Russell K. Hobbie
Bradley J. Roth

Intermediate Physics for Medicine and Biology

5th Edition

Springer

Russell K. Hobbie • Bradley J. Roth

Intermediate Physics for Medicine and Biology

Fifth Edition

Springer

Russell K. Hobbie
University of Minnesota
Minneapolis, Minnesota
USA

Bradley J. Roth
Oakland University
Rochester, Michigan
USA

ISBN 978-3-319-12681-4 ISBN 978-3-319-12682-1 (eBook)
DOI 10.1007/978-3-319-12682-1

Library of Congress Control Number: 2014954086

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

From the Preface to the Third Edition, by Russell K. Hobbie:

> Between 1971 and 1973 I audited all the courses medical students take in their first 2 years at the University of Minnesota. I was amazed at the amount of physics I found in these courses and how little of it is discussed in the general physics course.
>
> I found a great discrepancy between the physics in some papers in the biological research literature and what I knew to be the level of understanding of most biology majors or premed students who have taken a year of physics. It was clear that an intermediate level physics course would help these students. It would provide the physics they need and would relate it directly to the biological problems where it is useful.
>
> This book is the result of my having taught such a course since 1973. It is intended to serve as a text for an intermediate course taught in a physics department and taken by a variety of majors. Since its primary content is physics, I hope that physics faculty who might shy away from teaching a conventional biophysics course will consider teaching it. I also hope that research workers in biology and medicine will find it a useful reference to brush up on the physics they need or to find a few pointers to the current literature in a number of areas of biophysics. (The bibliography in each chapter is by no means exhaustive; however, the references should lead you quickly into a field.) The course offered at the University of Minnesota is taken by undergraduates in a number of majors who want to see more physics with biological applications and by graduate students in physics, biophysical sciences, biomedical engineering, physiology, and cell biology.
>
> Because the book is intended primarily for students who have taken only one year of physics, I have tried to adhere to the following principles in writing it:
>
> 1. Calculus is used without apology. When an important idea in calculus is used for the first time, it is reviewed in detail. These reviews are found in the appendices.
> 2. The reader is assumed to have taken physics and know the basic vocabulary. However, I have tried to present a logical development from first principles, but shorter than what would be found in an introductory course. An exception is found in Chaps. 14–18, where some results from quantum mechanics are used without deriving them from first principles. (My students have often expressed surprise at this change of pace.)
> 3. I have not intentionally left out steps in most derivations. Some readers may feel that the pace could be faster, particularly after a few chapters. My students have objected strongly when I have suggested stepping up the pace in class.
> 4. Each subject is approached in as simple a fashion as possible. I feel that sophisticated mathematics, such as vector analysis or complex exponential notation, often hides physical reality from the student. I have seen electrical engineering students who could not tell me what is happening in an RC circuit but could solve the equations with Laplace transforms.

The Fourth Edition followed the tradition of earlier editions. The book added a second author: Bradley J. Roth of Oakland University. Both of us have enjoyed this collaboration immensely. We added a chapter on sound and ultrasound, deleting or shortening topics elsewhere, in order to keep the book only slightly longer than the Third Edition.

The Fifth Edition does not add any new chapters, but almost every page has been improved and updated. Again, we fought the temptation to expand the book and deleted material when possible. Some of the deleted material is available at the book's website: http://www.oakland.edu/~roth/hobbie.htm. The Fifth Edition has 12 % more end-of-chapter problems than the Fourth Edition; most highlight biological applications of the physical principles. Many of the problems extend the material in the text. A solutions manual is available to those teaching the course. Instructors can use it as a reference or provide selected solutions to their students. The solutions manual makes it much easier for an instructor to guide an independent-study student. Information about the solutions manual is available at the book's website.

Chapter 1 reviews mechanics. Translational and rotational equilibrium are introduced, with the forces in the heel and hip joint as clinical examples. Stress and strain, hydrostatics, incompressible viscous flow, and the Poiseuille–Bernoulli equation are discussed, with examples from the circulatory system. The chapter concludes with a discussion of Reynolds number.

Chapter 2 is essential to nearly every other chapter in the book. It discusses exponential growth and decay and gives examples from pharmacology and physiology (including clearance). The logistic equation is discussed. Students are also shown how to use semilog and log–log plots and to determine power-law coefficients using a spreadsheet. The chapter concludes with a brief discussion of scaling.

Chapter 3 is a condensed treatment of statistical physics: average quantities, probability, thermal equilibrium, entropy, and the first and second laws of thermodynamics. Topics treated include the following: the Boltzmann factor and its corollary, the Nernst equation; the principle of equipartition of energy; the chemical potential; the general thermodynamic relationship; the Gibbs free energy; and the chemical potential of a solution. You can plow through this chapter if you are a slave to thoroughness, touch on the highlights, or use it as a reference as the topics are needed in later chapters.

Chapter 4 treats diffusion and transport of solute in an infinite medium. Fick's first and second laws of diffusion are developed. Steady-state solutions in one, two, and three dimensions are described. An important model is a spherical cell with pores providing transport through the cell membrane. It is shown that only a small number of pores are required to keep up with the rate of diffusion toward or away from the cell, so there is plenty of room on the cell surface for many different kinds of pores and receptor sites. The combination of diffusion and drift (or solvent drag) is also discussed. Finally, a simple random-walk model of diffusion is introduced.

Chapter 5 discusses transport of fluid and neutral solutes through a membrane. This might be a cell membrane, the basement membrane in the glomerulus of the kidney, or a capillary wall. The phenomenological transport equations including osmotic pressure are introduced as the first (linear) approximation to describe these flows. Countercurrent transport is described. Finally, a hydrodynamic model is developed for right-cylindrical pores. This model provides expressions for the phenomenological coefficients in terms of the pore radius and length. It is also used to calculate the net force on the membrane when there is flow.

After reviewing the electric field, electric potential, and circuits, *Chap. 6* describes the electrochemical changes that cause an impulse to travel along a nerve axon or along a muscle fiber before contraction. Two models are considered: electrotonus (when the membrane obeys Ohm's law) and the Hodgkin–Huxley model (when the membrane is nonlinear). Saltatory conduction in myelinated fibers is described. The dielectric properties of the membrane are modeled in terms of its molecular structure. Some simple changes to the membrane conductivity give rise to a periodically repeating action potential. Finally, a general relationship is developed between diffusive transport, resistance, and capacitance for a given geometry.

Chapter 7 shows how an electric potential is generated in the medium surrounding a nerve or muscle cell. This leads to the current dipole model for the electrocardiogram. The model is refined to account for the anisotropy of the electrical conductivity of the heart. We then discuss electrical stimulation, which is important for pacemakers, stimulating nerve and muscle cells, and defibrillation. Finally, the model is extended to the electroencephalogram.

Chapter 8 shows how the currents in a conducting nerve or muscle cell generate a magnetic field, leading to the magnetocardiogram and the magnetoencephalogram. Some bacteria and higher organisms contain magnetic particles used for determining spatial orientation in the earth's magnetic field. The mechanism by which these bacteria are oriented is described. The detection of weak magnetic fields and the use of changing magnetic fields to stimulate nerve or muscle cells are also discussed.

Chapter 9 covers a number of topics at the cellular and membrane level. It begins with Donnan equilibrium, where the presence of an impermeant ion on only one side of a membrane leads to the buildup of a potential difference across the membrane, and the Gouy–Chapman model for how ions redistribute near the membrane to generate this potential difference. The Debye–Hückel model is a simple description of the neutralization of ions by surrounding counterions. The Nernst–Planck equation provides the basic model for describing combined diffusion and drift in an applied electric field. It also forms the basis for the Goldman–Hodgkin–Katz model for zero total current in a membrane with a constant electric field. Gated membrane channels are then discussed. Noise is inescapable in all signalling situations. After developing the basic properties of shot noise and Johnson noise, we show how a properly adapted shark can detect very weak electric fields with a reasonable signal-to-noise ratio. The chapter concludes with a discussion of the basic physical principles that must be kept in mind when assessing the possibility of biological effects of weak electric and magnetic fields.

Chapter 10 describes feedback systems in the body. It starts with the regulation of breathing rate to stabilize the carbon dioxide level in the blood, moves to linear feedback systems with one and two time constants, and then to nonlinear models. We show how nonlinear systems described by simple difference equations can exhibit chaotic behavior, and how chaotic behavior can arise in continuous systems as well. Examples of feedback systems include Cheyne–Stokes respiration, heat stroke, pupil size, oscillating white-blood-cell counts, waves in excitable media, and period doubling and chaos in the heart.

Chapter 11 shows how the method of least squares underlies several important techniques for analyzing data. These range from simple curve fitting to discrete and continuous Fourier series, power spectra, correlation functions, and the Fourier transform. We then describe the frequency response of a linear system and the frequency spectrum of noise. We conclude with a brief discussion of testing data for chaotic behavior and the important concept of stochastic resonance.

Armed with the tools of the previous chapter, we turn to images in *Chap. 12*. Images are analyzed from the standpoint of linear systems and convolution. This leads to the use of Fourier analysis to describe the spatial frequencies in an image and the reconstruction of an image from its projections. Both Fourier techniques and filtered back projection are discussed.

Chapter 13 analyzes sound, hearing, and medical ultrasound. The wave equation is derived, and the wave speed and acoustic impedance are related to the tissue properties. The structure and function of the ear is described. Finally, methods for ultrasonic imaging are discussed, including pulse echo techniques and Doppler imaging.

Chapter 14 discusses the visible, infrared, and ultraviolet regions of the electromagnetic spectrum. The scattering and absorption cross sections are introduced and are used here and in the next three chapters. We then describe the diffusion model for photon transport in turbid media. Thermal radiation emitted by the body can be detected; the emission of thermal radiation by the sun includes ultraviolet light, which injures skin. Protection from ultraviolet light is both possible and prudent. The definitions of various radiometric quantities have varied from one field of research to another. We present a coherent description of radiometric, photometric, and actinometric definitions. We then turn to the eye, showing how spectacle lenses are used to correct errors of refraction. The chapter closes with a description of the quantum limitations to dark-adapted vision.

Chapter 15, like Chap. 3, has few biological examples but sets the stage for later work. It describes how photons and ionizing charged particles such as electrons lose energy in traversing matter. These interaction mechanisms, both in the body and in the detector, are fundamental to the formation of a radiographic image and to the use of radiation to treat cancer.

Chapter 16 describes the use of x rays for medical diagnosis and treatment. It moves from production to detection, to the diagnostic radiograph. We discuss image quality and noise, followed by angiography, mammography, fluoroscopy, and computed tomography. After briefly reviewing radiobiology, we discuss therapy and dose measurement. The chapter closes with a section on the risks from radiation.

Chapter 17 introduces nuclear physics and nuclear medicine. The different kinds of radioactive decay are described. Dose calculations are made using the fractional absorbed dose method recommended by the Medical Internal Radiation Dose Committee of the Society of Nuclear Medicine and Molecular Imaging. Auger electrons can magnify the dose delivered to a cell or to DNA. This can potentially provide new methods of treatment. Diagnostic imaging includes single photon emission tomography and positron emission tomography. Therapies include brachytherapy and internal radiotherapy. A section on the nuclear physics of radon closes the chapter.

Chapter 18 develops the physics of magnetic resonance imaging (MRI). We show how the basic pulse sequences are formed and used for slice selection, readout, image reconstruction, and to manipulate image contrast. We close with chemical shift imaging, flow effects, functional MRI, and diffusion and diffusion tensor MRI.

Biophysics is a very broad subject. Nearly every branch of physics has something to contribute, and the boundaries between physics and engineering are blurred. Each chapter could be much longer; we have attempted to provide the essential physical tools. Molecular biophysics has been almost completely ignored: excellent texts already exist, and this is not our area of expertise. This book has become long enough.

We would appreciate receiving any corrections or suggestions for improving the book.

Finally, thanks to our long-suffering families. We never understood what these common words really mean, nor the depth of our indebtedness, until we wrote the book.

Russell K. Hobbie
Professor of Physics Emeritus, University of Minnesota
(hobbie@umn.edu)
Bradley J. Roth
Professor of Physics, Oakland University
(roth@oakland.edu)

Contents

1 Mechanics 1
1.1 Distances and Sizes 1
1.2 Models 3
1.3 Forces and Translational Equilibrium 3
1.4 Rotational Equilibrium 4
1.5 Vector Product 6
1.6 Force in the Achilles Tendon 6
1.7 Forces on the Hip 7
1.8 The Use of a Cane 10
1.9 Work 11
1.10 Stress and Strain 12
1.11 Shear 13
1.12 Hydrostatics 13
1.13 Buoyancy 15
1.14 Compressibility 15
1.15 Diving 15
1.16 Viscosity 15
1.17 Viscous Flow in a Tube 16
1.18 Pressure–Volume Work 19
1.19 The Human Circulatory System 20
1.20 Turbulent Flow and the Reynolds Number 22
Symbols Used 24
Problems 25
References 30

2 Exponential Growth and Decay 33
2.1 Exponential Growth 33
2.2 Exponential Decay 35
2.3 Semilog Paper 36
2.4 Variable Rates 38
2.5 Clearance 40
2.6 The Chemostat 40
2.7 Multiple Decay Paths 41
2.8 Decay Plus Input at a Constant Rate 41
2.9 Decay With Multiple Half-Lives and Fitting Exponentials 41
2.10 The Logistic Equation 42

2.11 Log–log Plots, Power Laws, and Scaling . . . 43
2.11.1 Log–log Plots and Power Laws . . . 43
2.11.2 Food Consumption, Basal Metabolic Rate, and Scaling . . . 44
Symbols Used . . . 45
Problems . . . 45
References . . . 51

3 Systems of Many Particles . . . 53
3.1 Gas Molecules in a Box . . . 54
3.2 Microstates and Macrostates . . . 56
3.3 The Energy of a System: The First Law of Thermodynamics . . . 57
3.4 Ensembles and the Basic Postulates . . . 59
3.5 Thermal Equilibrium . . . 60
3.6 Entropy . . . 62
3.7 The Boltzmann Factor . . . 62
3.8 The Nernst Equation . . . 63
3.9 The Pressure Variation in the Atmosphere . . . 64
3.10 Equipartition of Energy and Brownian Motion . . . 64
3.11 Heat Capacity . . . 65
3.12 Equilibrium When Particles Can Be Exchanged: the Chemical Potential . . . 66
3.13 Concentration Dependence of the Chemical Potential . . . 67
3.14 Systems That Can Exchange Volume . . . 68
3.15 Extensive Variables and Generalized Forces . . . 68
3.16 The General Thermodynamic Relationship . . . 69
3.17 The Gibbs Free Energy . . . 70
3.17.1 Gibbs Free Energy . . . 70
3.17.2 An Example: Chemical Reactions . . . 71
3.18 The Chemical Potential of a Solution . . . 72
3.19 Transformation of Randomness to Order . . . 74
Symbols Used . . . 75
Problems . . . 76
References . . . 83

4 Transport in an Infinite Medium . . . 85
4.1 Flux, Fluence, and Continuity . . . 85
4.1.1 The Continuity Equation in One Dimension . . . 86
4.1.2 The Continuity Equation in Three Dimensions . . . 86
4.1.3 The Integral Form of the Continuity Equation . . . 87
4.1.4 The Differential Form of the Continuity Equation . . . 88
4.1.5 The Continuity Equation with a Chemical Reaction . . . 89
4.2 Drift or Solvent Drag . . . 89
4.3 Brownian Motion . . . 89
4.4 Motion in a Gas: Mean Free Path and Collision Time . . . 90
4.5 Motion in a Liquid . . . 91
4.6 Diffusion: Fick's First Law . . . 92
4.7 The Einstein Relationship Between Diffusion and Viscosity . . . 93
4.8 Fick's Second Law of Diffusion . . . 95
4.9 Time-Independent Solutions . . . 97
4.10 Example: Steady-State Diffusion to a Spherical Cell and End Effects . . . 98
4.10.1 Diffusion Through a Collection of Pores, Corrected . . . 100
4.10.2 Diffusion from a Sphere, Corrected . . . 100
4.10.3 How Many Pores Are Needed? . . . 100
4.10.4 Other Applications of the Model . . . 101

4.11 Example: A Spherical Cell Producing a Substance . . . 101
4.12 Drift and Diffusion in One Dimension . . . 102
4.13 A General Solution for the Particle Concentration as a Function of Time . . . 104
4.14 Diffusion as a Random Walk . . . 105
Symbols Used . . . 107
Problems . . . 108
References . . . 114

5 Transport Through Neutral Membranes . . . 117
5.1 Membranes . . . 117
5.2 Osmotic Pressure in an Ideal Gas . . . 118
5.3 Osmotic Pressure in a Liquid . . . 120
5.4 Some Clinical Examples . . . 121
5.4.1 Edema Due to Heart Failure . . . 122
5.4.2 Nephrotic Syndrome, Liver Disease, and Ascites . . . 122
5.4.3 Edema of Inflammatory Reaction . . . 122
5.4.4 Headaches in Renal Dialysis . . . 123
5.4.5 Osmotic Diuresis . . . 123
5.4.6 Osmotic Fragility of Red Cells . . . 123
5.5 Volume Transport Through a Membrane . . . 123
5.6 Solute Transport Through a Membrane . . . 125
5.7 Example: The Artificial Kidney . . . 126
5.8 Countercurrent Transport . . . 127
5.9 A Continuum Model for Volume and Solute Transport in a Pore . . . 128
5.9.1 Volume Transport . . . 128
5.9.2 Solute Transport . . . 130
5.9.3 Summary . . . 133
5.9.4 Reflection Coefficient . . . 133
5.9.5 The Effect of Pore Walls on Diffusion . . . 134
5.9.6 Net Force on the Membrane . . . 134
Symbols Used . . . 135
Problems . . . 135
References . . . 139

6 Impulses in Nerve and Muscle Cells . . . 141
6.1 Physiology of Nerve and Muscle Cells . . . 141
6.2 Coulomb's Law, Superposition, and the Electric Field . . . 143
6.3 Gauss's Law . . . 144
6.4 Potential Difference . . . 147
6.5 Conductors . . . 148
6.6 Capacitance . . . 149
6.7 Dielectrics . . . 149
6.8 Current and Ohm's Law . . . 151
6.9 The Application of Ohm's Law to Simple Circuits . . . 153
6.10 Charge Distribution in the Resting Nerve Cell . . . 154
6.11 The Cable Model for an Axon . . . 155
6.12 Electrotonus or Passive Spread . . . 159
6.13 The Hodgkin–Huxley Model for Membrane Current . . . 160
6.13.1 Voltage Clamp Experiments . . . 161
6.13.2 Potassium Conductance . . . 163
6.13.3 Sodium Conductance . . . 164
6.13.4 Leakage Current . . . 165

6.14 Voltage Changes in a Space-Clamped Axon . . . 165
6.15 Propagating Nerve Impulse . . . 166
6.16 Myelinated Fibers and Saltatory Conduction . . . 167
6.17 Membrane Capacitance . . . 169
6.18 Rhythmic Electrical Activity . . . 172
6.19 The Relationship Between Capacitance, Resistance, and Diffusion . . . 172
6.19.1 Capacitance and Resistance . . . 172
6.19.2 Capacitance and Diffusion . . . 173
Symbols Used . . . 174
Problems . . . 175
References . . . 183

7 The Exterior Potential and the Electrocardiogram . . . 185
7.1 The Potential Outside a Long Cylindrical Axon . . . 185
7.2 The Exterior Potential is Small . . . 187
7.3 The Potential Far from the Axon . . . 188
7.4 The Exterior Potential for an Arbitrary Pulse . . . 189
7.5 Electrical Properties of the Heart . . . 193
7.6 The Current-Dipole Vector of the Heart as a Function of Time . . . 195
7.7 The Electrocardiographic Leads . . . 195
7.8 Some Electrocardiograms . . . 198
7.9 Refinements to the Model . . . 199
7.9.1 The Fiber Has a Finite Radius . . . 200
7.9.2 Nonuniform Exterior Conductivity . . . 200
7.9.3 Anisotropic Conductivity: The Bidomain Model . . . 200
7.10 Electrical Stimulation . . . 201
7.11 The Electroencephalogram . . . 205
Symbols Used . . . 206
Problems . . . 206
References . . . 211

8 Biomagnetism . . . 213
8.1 The Magnetic Force on a Moving Charge . . . 213
8.1.1 The Lorentz Force . . . 213
8.1.2 The Cyclotron . . . 215
8.2 The Magnetic Field of a Moving Charge or a Current . . . 215
8.2.1 The Divergence of the Magnetic Field is Zero . . . 215
8.2.2 Ampere's Circuital Law . . . 216
8.2.3 The Biot–Savart Law . . . 216
8.2.4 The Displacement Current . . . 217
8.3 The Magnetic Field Around an Axon . . . 218
8.4 The Magnetocardiogram . . . 219
8.5 The Magnetoencephalogram . . . 223
8.6 Electromagnetic Induction . . . 224
8.7 Magnetic Stimulation . . . 225
8.8 Magnetic Materials and Biological Systems . . . 226
8.8.1 Magnetic Materials . . . 226
8.8.2 Measuring Magnetic Properties in People . . . 228
8.8.3 Magnetic Orientation . . . 228
8.8.4 Magnetic Nanoparticles . . . 229
8.9 Detection of Weak Magnetic Fields . . . 229
Symbols Used . . . 231
Problems . . . 231
References . . . 236

9 Electricity and Magnetism at the Cellular Level . . . 239
9.1 Donnan Equilibrium . . . 239
9.2 Potential Change at an Interface: The Gouy–Chapman Model . . . 241
9.3 Ions in Solution: The Debye–Hückel Model . . . 244
9.4 Saturation of the Dielectric . . . 245
9.5 Ion Movement in Solution: The Nernst–Planck Equation . . . 247
9.6 Zero Total Current in a Constant-Field Membrane: The Goldman Equations . . . 249
9.7 Membrane Channels . . . 250
9.8 Noise . . . 254
9.8.1 Shot Noise . . . 254
9.8.2 Johnson Noise . . . 255
9.9 Sensory Transducers . . . 256
9.10 Possible Effects of Weak External Electric and Magnetic Fields . . . 256
9.10.1 Strong Fields . . . 257
9.10.2 Power Frequency (50–60 Hz) Fields . . . 257
9.10.2.1 Fields in Homes are Weak . . . 257
9.10.2.2 Epidemiological Studies . . . 257
9.10.2.3 Laboratory Studies . . . 258
9.10.2.4 Reviews and Panel Reports . . . 258
9.10.2.5 Electric Fields in the Body . . . 258
9.10.2.6 Electric Fields in a Spherical Cell . . . 259
9.10.3 Electrical Interactions and Noise . . . 259
9.10.4 Magnetic Interactions and Noise . . . 260
9.10.5 Microwaves, Mobile Phones, and Wi-Fi . . . 261
Symbols Used . . . 262
Problems . . . 263
References . . . 266

10 Feedback and Control . . . 269
10.1 Steady-State Relationships Among Variables . . . 270
10.2 Determining the Operating Point . . . 271
10.3 Regulation of a Variable and Open-Loop Gain . . . 271
10.4 Approach to Equilibrium without Feedback . . . 273
10.5 Approach to Equilibrium in a Feedback Loop with One Time Constant . . . 273
10.6 A Feedback Loop with Two Time Constants . . . 276
10.7 Proportional, Derivative, and Integral Control . . . 278
10.8 Models Using Nonlinear Differential Equations . . . 279
10.8.1 Describing a Nonlinear System . . . 280
10.8.2 An Example of Phase Resetting: The Radial Isochron Clock . . . 281
10.8.3 Stopping an Oscillator . . . 283
10.9 Difference Equations and Chaotic Behavior . . . 284
10.9.1 The Logistic Map: Period Doubling and Deterministic Chaos . . . 284
10.9.2 The Bifurcation Diagram . . . 285
10.9.3 Quasiperiodicity . . . 286
10.10 A Feedback Loop with a Time Constant and a Fixed Delay . . . 287
10.11 Negative Feedback Loops: A Summary . . . 288
10.12 Additional Examples . . . 289
10.12.1 Cheyne–Stokes Respiration . . . 289
10.12.2 Hot Tubs and Heat Stroke . . . 289
10.12.3 Pupil Size . . . 289
10.12.4 Oscillating White-Blood-Cell Counts . . . 290

10.12.5 Waves in Excitable Media . . . 290
10.12.6 Period Doubling and Chaos in Heart Cells . . . 291
Symbols Used . . . 292
Problems . . . 292
References . . . 300

11 The Method of Least Squares and Signal Analysis . . . 303
11.1 The Method of Least Squares and Polynomial Regression . . . 303
11.1.1 The Simplest Example . . . 303
11.1.2 A Linear Fit . . . 304
11.1.3 A Polynomial Fit . . . 305
11.1.4 Variable Weighting . . . 306
11.2 Nonlinear Least Squares . . . 306
11.3 The Presence of Many Frequencies in a Periodic Function . . . 308
11.4 Fourier Series for Discrete Data . . . 308
11.4.1 Determining the Parameters . . . 309
11.4.2 Equally Spaced Data Points Simplify the Equations . . . 310
11.4.3 The Standard Form for the Discrete Fourier Transform . . . 310
11.4.4 Complex Exponential Notation . . . 311
11.4.5 Example: The Square Wave . . . 311
11.4.6 Example: When the Sampling Time is not a Multiple of the Period of the Signal . . . 312
11.4.7 Example: Spontaneous Births . . . 313
11.4.8 Example: Photosynthesis in Plants . . . 314
11.4.9 Pitfalls of Discrete Sampling: Aliasing . . . 314
11.4.10 Fast Fourier Transform . . . 315
11.5 Fourier Series for a Periodic Function . . . 315
11.6 The Power Spectrum . . . 317
11.7 Correlation Functions . . . 317
11.7.1 Cross-Correlation of a Pulse . . . 318
11.7.2 Cross-Correlation of a Nonpulse Signal . . . 318
11.7.3 Cross-Correlation Example . . . 318
11.7.4 Autocorrelation . . . 318
11.7.5 Autocorrelation Examples . . . 319
11.8 The Autocorrelation Function and the Power Spectrum . . . 320
11.9 Nonperiodic Signals and Fourier Integrals . . . 320
11.9.1 Introduce Negative Frequencies and Make the Coefficients Half as Large . . . 321
11.9.2 Make the Period Infinite . . . 322
11.9.3 Complex Notation . . . 322
11.9.4 Example: The Exponential Pulse . . . 322
11.10 The Delta Function . . . 323
11.11 The Energy Spectrum of a Pulse and Parseval's Theorem . . . 324
11.11.1 Parseval's Theorem . . . 324
11.11.2 Example: The Exponential Pulse . . . 325
11.12 The Autocorrelation of a Pulse and its Relation to the Energy Spectrum . . . 325
11.13 Noise . . . 326
11.14 Correlation Functions and Noisy Signals . . . 327
11.14.1 Detecting Signals in Noise . . . 327
11.14.2 Signal Averaging . . . 328
11.14.3 Power Spectral Density . . . 328
11.14.4 Units . . . 329
11.15 Frequency Response of a Linear System . . . 330
11.15.1 Example of Calculating the Frequency Response . . . 330
11.15.2 The Decibel . . . 331
11.15.3 Example: Impulse Response . . . 331

11.16 The Frequency Spectrum of Noise . . . 332
11.16.1 Johnson Noise . . . 332
11.16.2 Shot Noise . . . 335
11.16.3 $1/f$ Noise . . . 335
11.17 Testing Data for Chaotic Behavior . . . 335
11.17.1 Embedding . . . 335
11.17.2 Surrogate Data . . . 336
11.18 Stochastic Resonance . . . 337
11.18.1 Threshold Detection . . . 337
11.18.2 Feynman's Ratchet . . . 337
Symbols Used . . . 338
Problems . . . 339
References . . . 343

12 Images . . . 345
12.1 The Convolution Integral and Its Fourier Transform . . . 345
12.1.1 One Dimension . . . 345
12.1.2 Two Dimensions . . . 346
12.2 The Relationship Between the Object and the Image . . . 347
12.2.1 Point Spread Function . . . 347
12.2.2 Optical, Modulation, and Phase Transfer Functions . . . 348
12.2.3 Line and Edge Spread Functions . . . 349
12.3 Spatial Frequencies in an Image . . . 349
12.3.1 Summary . . . 351
12.4 Two-Dimensional Image Reconstruction from Projections by Fourier Transform . . . 351
12.5 Reconstruction from Projections by Filtered Back Projection . . . 352
12.6 An Example of Filtered Back Projection . . . 355
Symbols Used . . . 358
Problems . . . 358
References . . . 362

13 Sound and Ultrasound . . . 363
13.1 The Wave Equation . . . 363
13.1.1 Plane Waves in an Elastic Rod . . . 363
13.1.2 Plane Waves in a Fluid . . . 364
13.1.3 Shear Waves . . . 365
13.2 Properties of the Wave Equation . . . 365
13.3 Acoustic Impedance . . . 366
13.3.1 Relationships Between Pressure, Displacement and Velocity in a Plane Wave . . . 366
13.3.2 Reflection and Transmission of Sound at a Boundary . . . 367
13.4 Comparing Intensities: Decibels . . . 368
13.4.1 The Decibel . . . 368
13.4.2 Measuring Hearing Response . . . 368
13.5 The Ear and Hearing . . . 369
13.6 Attenuation . . . 370
13.7 Diagnostic Uses of Ultrasound . . . 371
13.7.1 Ultrasound Transducers . . . 371
13.7.2 Pulse Echo Imaging . . . 373
13.7.3 The Doppler Effect . . . 374
13.7.4 Elastography . . . 375
13.7.5 Safety . . . 375
13.8 Therapeutic Uses of Ultrasound . . . 375
Symbols Used . . . 376
Problems . . . 376
References . . . 379

14 Atoms and Light . . . 381
14.1 The Nature of Light: Waves and Photons . . . 381
14.2 Electron Waves and Particles: The Electron Microscope . . . 383
14.3 Atomic Energy Levels and Atomic Spectra . . . 383
14.4 Molecular Energy Levels . . . 384
14.5 Scattering and Absorption of Radiation; Cross Section . . . 387
14.6 The Diffusion Approximation to Photon Transport . . . 389
14.6.1 Diffusion Approximation . . . 389
14.6.2 Continuous Measurements . . . 390
14.6.3 Pulsed Measurements . . . 391
14.6.4 Refinements to the Model . . . 391
14.7 Biological Applications of Infrared Scattering . . . 392
14.7.1 Near Infrared (NIR) . . . 392
14.7.2 Optical Coherence Tomography (OCT) . . . 392
14.7.3 Raman Spectroscopy . . . 393
14.7.4 Far Infrared or Terahertz Radiation . . . 394
14.8 Thermal Radiation . . . 394
14.9 Infrared Radiation from the Body . . . 398
14.9.1 Atherosclerotic Coronary Heart Disease . . . 399
14.9.2 Photodynamic Therapy . . . 399
14.10 Blue and Ultraviolet Radiation . . . 400
14.10.1 Treatment of Neonatal Jaundice . . . 400
14.10.2 The Ultraviolet Spectrum . . . 400
14.10.3 Response of the Skin to Ultraviolet Light . . . 401
14.10.4 Ultraviolet Light Causes Skin Cancer . . . 402
14.10.5 Protection From Ultraviolet Light . . . 403
14.10.6 Ultraviolet Light Damages the Eye . . . 403
14.10.7 Ultraviolet Light Therapy . . . 403
14.11 Heating Tissue with Light . . . 404
14.12 Radiometry and Photometry . . . 405
14.12.1 Radiometric Definitions . . . 407
14.12.1.1 Radiant Energy and Power . . . 407
14.12.1.2 Point Source: Radiant Intensity . . . 407
14.12.1.3 Extended Source: Radiance . . . 407
14.12.1.4 Energy Striking a Surface: Irradiance . . . 408
14.12.1.5 Plane-Wave Relationships . . . 409
14.12.1.6 Isotropic Radiation: Lambert's Law . . . 409
14.12.1.7 The Spectrum . . . 409
14.12.2 Photometric Definitions . . . 409
14.12.3 Actinometric Definitions . . . 410
14.13 The Eye . . . 410
14.14 Quantum Effects in Dark-Adapted Vision . . . 413
14.15 Color Vision . . . 415
Symbols Used . . . 415
Problems . . . 417
References . . . 421

15 Interaction of Photons and Charged Particles with Matter . . . 425
15.1 Atomic Energy Levels and X-ray Absorption . . . 425
15.2 Photon Interactions . . . 426
15.2.1 Photoelectric Effect . . . 426
15.2.2 Compton and Incoherent Scattering . . . 427
15.2.3 Coherent Scattering . . . 427

15.2.4 Inelastic Scattering . . . 427
15.2.5 Pair Production . . . 428
15.2.6 Energy Dependence . . . 428
15.3 The Photoelectric Effect . . . 428
15.4 Compton Scattering . . . 428
15.4.1 Kinematics . . . 428
15.4.2 Cross Section: Klein–Nishina Formula . . . 430
15.4.3 Incoherent Scattering . . . 431
15.4.4 Energy Transferred to the Electron . . . 431
15.5 Coherent Scattering . . . 431
15.6 Pair Production . . . 432
15.7 The Photon Attenuation Coefficient . . . 432
15.8 Compounds and Mixtures . . . 433
15.9 Deexcitation of Atoms . . . 434
15.10 Energy Transfer from Photons to Electrons . . . 436
15.11 Charged-Particle Stopping Power . . . 438
15.11.1 Interaction with Target Electrons . . . 442
15.11.2 Scattering from the Nucleus . . . 445
15.11.3 Stopping of Electrons . . . 446
15.11.4 Compounds . . . 446
15.12 Linear Energy Transfer and Restricted Collision Stopping Power . . . 447
15.13 Range, Straggling, and Radiation Yield . . . 447
15.14 Track Structure . . . 448
15.15 Energy Transferred and Energy Imparted; Kerma and Absorbed Dose . . . 450
15.15.1 An Example . . . 450
15.15.2 Energy Transferred and Kerma . . . 451
15.15.3 Energy Imparted and Absorbed Dose . . . 452
15.15.4 Net Energy Transferred, Collision Kerma and Radiative Kerma . . . 452
15.16 Charged-Particle Equilibrium . . . 452
15.16.1 Radiation Equilibrium . . . 452
15.16.2 Charged-Particle Equilibrium . . . 453
15.17 Buildup . . . 454
Symbols Used . . . 455
Problems . . . 456
References . . . 459

16 Medical Uses of X-Rays . . . 461
16.1 Production of X-Rays . . . 461
16.1.1 Characteristic X-Rays . . . 461
16.1.2 Bremsstrahlung . . . 462
16.2 Quantities to Describe Radiation Interactions . . . 463
16.2.1 Radiation Chemical Yield . . . 463
16.2.2 Mean Energy per Ion Pair . . . 463
16.2.3 Exposure . . . 464
16.3 Detectors . . . 464
16.3.1 Film and Screens . . . 465
16.3.2 Scintillation Detectors . . . 466
16.3.3 Gas Detectors . . . 468
16.3.4 Semiconductor Detectors . . . 469
16.3.5 Thermoluminescent Dosimeters . . . 469
16.3.6 Chemical Dosimetry . . . 469
16.3.7 Digital Detectors . . . 470

16.4 The Diagnostic Radiograph . . . 470
16.4.1 X-Ray Tube and Filter . . . 470
16.4.2 Collimation . . . 471
16.4.3 Attenuation in the Patient: Contrast Material . . . 471
16.4.4 Antiscatter Grid . . . 473
16.4.5 Detector . . . 474
16.5 Image Quality . . . 474
16.6 Angiography and Digital Subtraction Angiography . . . 476
16.7 Mammography . . . 477
16.8 Computed Tomography . . . 477
16.9 Biological Effects of Radiation . . . 480
16.9.1 Cell-Culture Experiments . . . 480
16.9.2 Chromosome Damage . . . 481
16.9.3 The Linear-Quadratic Model . . . 482
16.9.4 The Bystander Effect . . . 483
16.9.5 Tissue Irradiation . . . 483
16.9.6 A Model for Tumor Eradication . . . 485
16.10 Radiation Therapy . . . 485
16.10.1 Classical Radiation Therapy . . . 486
16.10.2 Modern X-Ray Therapy . . . 487
16.10.3 Charged Particles and Neutrons . . . 488
16.11 Dose Measurement . . . 489
16.12 The Risk of Radiation . . . 490
16.12.1 Equivalent and Effective Dose . . . 490
16.12.1.1 Equivalent Dose . . . 490
16.12.1.2 Detriment and Effective Dose . . . 491
16.12.2 Comparison With Natural Background . . . 491
16.12.3 Calculating Risk . . . 493
16.12.3.1 The Linear No-Threshold Model and Collective Dose . . . 493
16.12.3.2 Other Models . . . 494
16.12.4 Radon . . . 495
Symbols Used . . . 496
Problems . . . 496
References . . . 500

17 Nuclear Physics and Nuclear Medicine . . . 503
17.1 Nuclear Systematics . . . 503
17.2 Nuclear Decay: Decay Rate and Half-Life . . . 506
17.3 Gamma Decay and Internal Conversion . . . 507
17.4 Atomic Deexcitation . . . 507
17.5 Beta Decay and Electron Capture . . . 507
17.6 Calculating the Absorbed Dose from Radioactive Nuclei within the Body: the MIRD Method . . . 510
17.6.1 Activity and Cumulated Activity . . . 511
17.6.1.1 The General Distribution Problem: Residence Time . . . 512
17.6.1.2 Immediate Uptake with No Biological Excretion . . . 512
17.6.1.3 Immediate Uptake with Exponential Biological Excretion . . . 512
17.6.1.4 Immediate Uptake Moving through Two Compartments . . . 513
17.6.1.5 More Complicated Situations . . . 514
17.6.1.6 Activity per Unit Mass . . . 514
17.6.2 Mean Energy Emitted Per Unit Cumulated Activity . . . 514
17.6.3 Calculation of the Absorbed Fraction . . . 514
17.6.3.1 Nonpenetrating Radiation . . . 514
17.6.3.2 Infinite Source in an Infinite Medium . . . 514
17.6.3.3 Point Source of Monoenergetic Photons in Empty Space . . . 514

17.6.3.4 Point Source of Monoenergetic Photons in an Infinite Isotropic Absorber . . . 515
17.6.3.5 More Complicated Cases—the MIRD Tables . . . 515
17.6.4 Sample Dose Calculation . . . 517
17.7 Radiopharmaceuticals and Tracers . . . 517
17.7.1 Physical Properties . . . 517
17.7.2 Biological Properties . . . 519
17.8 Detectors; The Gamma Camera . . . 520
17.9 Single-Photon Emission Computed Tomography . . . 521
17.10 Positron Emission Tomography . . . 523
17.11 Brachytherapy and Internal Radiotherapy . . . 523
17.12 Radon . . . 524
Symbols Used . . . 526
Problems . . . 527
References . . . 532

18 Magnetic Resonance Imaging . . . 535
18.1 Magnetic Moments in an External Magnetic Field . . . 535
18.2 The Source of the Magnetic Moment . . . 536
18.3 The Magnetization . . . 537
18.4 Behavior of the Magnetization Vector . . . 538
18.5 A Rotating Coordinate System . . . 539
18.5.1 Transforming to the Rotating Coordinate System . . . 539
18.5.2 An Additional Oscillating Field . . . 540
18.5.3 Nutation . . . 541
18.5.4 π and $\pi/2$ Pulses . . . 541
18.6 Relaxation Times . . . 542
18.7 Detecting the Magnetic Resonance Signal . . . 544
18.8 Some Useful Pulse Sequences . . . 545
18.8.1 Free-Induction-Decay (FID) Sequence . . . 546
18.8.2 Inversion-Recovery (IR) Sequence . . . 546
18.8.3 Spin-Echo (SE) Sequence . . . 546
18.8.4 Carr–Purcell (CP) Sequence . . . 547
18.8.5 Carr–Purcell–Meiboom–Gill (CPMG) Sequence . . . 547
18.9 Imaging . . . 548
18.9.1 Slice Selection . . . 548
18.9.2 Readout in the Direction . . . 550
18.9.3 Projection Reconstruction . . . 551
18.9.4 Phase Encoding . . . 551
18.9.5 Other Pulse Sequences . . . 553
18.9.6 Image Contrast and the Pulse Parameters . . . 554
18.9.7 Safety . . . 555
18.10 Chemical Shift . . . 555
18.11 Flow Effects . . . 555
18.12 Functional MRI . . . 557
18.13 Diffusion and Diffusion Tensor MRI . . . 557
18.14 Hyperpolarized MRI of the Lung . . . 558
Symbols Used . . . 559
Problems . . . 559
References . . . 564

A Appendix A
Plane and Solid Angles . . . 567
A.1 Plane Angles . . . 567
A.2 Solid Angles . . . 567

B Appendix B
Vectors; Displacement, Velocity, and Acceleration . . . 569
B.1 Vectors and Vector Addition . . . 569
B.2 Components of Vectors . . . 570
B.3 Position, Velocity, and Acceleration . . . 570

C Appendix C
Properties of Exponents and Logarithms . . . 573

D Appendix D
Taylor's Series . . . 575

E Appendix E
Some Integrals of Sines and Cosines . . . 579

F Appendix F
Linear Differential Equations with Constant Coefficients . . . 581
F.1 First-Order Equation . . . 582
F.2 Second-Order Equation . . . 582

G Appendix G
The Mean and Standard Deviation . . . 585

H Appendix H
The Binomial Probability Distribution . . . 587

I Appendix I
The Gaussian Probability Distribution . . . 591

J Appendix J
The Poisson Distribution . . . 595

K Appendix K
Integrals Involving e^{-ax^2} . . . 599

L Appendix L
Spherical and Cylindrical Coordinates . . . 601

M Appendix M
Joint Probability Distributions . . . 605
M.1 Discrete Variables . . . 605
M.2 Continuous Variables . . . 605

N Appendix N
Partial Derivatives . . . 607

O Appendix O
Some Fundamental Constants and Conversion Factors . . . 609

Index . . . 611

Mechanics

1

This chapter introduces some concepts from mechanics that are of biological or medical interest. We begin with a discussion of sizes important in biology. Then we turn to the forces on an object that is in equilibrium and calculate the forces experienced by various bones and muscles. In Sect. 1.9, we introduce the concept of mechanical work, which will recur throughout the book. The next two sections describe how materials deform when forces act on them. Sections 1.12 through 1.16 discuss the forces in stationary and moving fluids. These concepts are then applied to laminar viscous flow in a pipe, which is a model for the flow of blood and the flow of fluid through pores in cell membranes. The chapter ends with a discussion of the circulatory system.

1.1 Distances and Sizes

In biology and medicine, we study objects that span a wide range of sizes: from giant redwood trees to individual molecules. Therefore, we begin with a brief discussion of length scales. The basic unit of length in the metric system[1] is the meter (m): about the height of a 3-year-old child. For objects much larger or smaller than a meter, we add a prefix as shown in Table 1.1. For example, a kilometer is formed by adding the prefix "kilo," which means times one thousand (10^3 m = 1 km). Living organisms rarely, if ever, reach a size of 1 km; the tallest trees are about 0.1 km (100 m) high. A few animals (whales, dinosaurs) reach the size of tens of meters, but most organisms are a few meters or less in size.

The diversity of life becomes more obvious as we move down to smaller length scales. One one-hundredth of a meter is called a centimeter (1 cm = 10^{-2} m). The centimeter is still common in the medical literature, although it is going out of style among metric purists who prefer to use only prefixes that are factors of 1000.[2] One one-thousandth of a meter is a millimeter (1 mm = 10^{-3} m), about the thickness of a dime. We can still see objects of this size, but we cannot study their detailed structure with the unaided eye.

Table 1.1 Common prefixes used in the metric system

Prefix	Abbreviation	Multiply by
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

The microscope enables us to study objects many times smaller than 1 mm. The natural unit for measuring such objects is 10^{-6} m or 10^{-3} mm, called a micrometer (1 μm = 10^{-6} m). The nickname for the micrometer is the "micron." Figure 1.1 shows the relative sizes of objects in the range of 1 mm–1 μm and encompasses the length scale of cell biology. Many small structures of our body are this size. For instance, our lungs consist of a branching network of tubes through which air flows. These tubes end in small, nearly spherical air sacs called *alveoli* (Fig. 1.1b). Each alveolus has a diameter of about 250 μm, and this size is set by the diffusion properties of air (Chap. 4). *Protozoans* are a type of small one-celled animal. A paramecium is a protozoan about 250 μm long (Fig. 1.1a). The cells in multicellular animals tend to be somewhat smaller than protozoans. For instance, the mammalian *cardiac cell* (a muscle cell found in the heart, Chap. 7) shown in Fig. 1.1c is about

[1] The metric system is officially called the SI system (systeme internationale). It used to be called the MKS (meter kilogram second) system.

[2] We find that restricting ourselves to prefixes that are a multiple of 1000 makes it easier to remember relative sizes.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_1, © Springer International Publishing Switzerland 2015

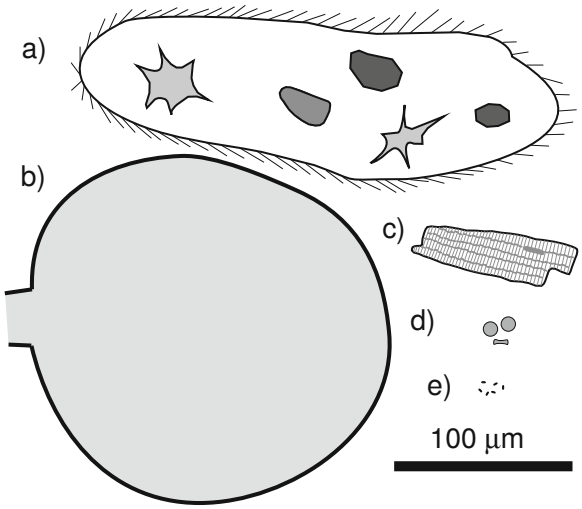


Fig. 1.1 Objects ranging in size from 1 mm down to 1 μm. **a** A paramecium. **b** An alveolus (air sac in the lung). **c** A cardiac cell. **d** Red blood cells. **e** *Escherichia coli* bacteria

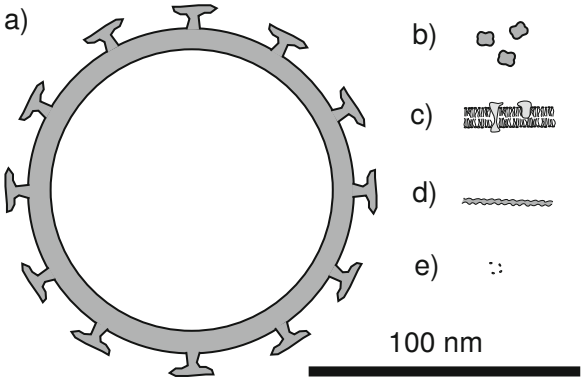


Fig. 1.2 Objects ranging in size from 1 μm down to 1 nm. **a** The human immunodeficiency virus (HIV). **b** Hemoglobin molecules. c A cell membrane. **d** A DNA molecule. **e** Glucose molecules

100-μm long and 20 μm in diameter. Nerve cells have a long fiber-like extension called an *axon*. Axons come in a variety of sizes, from 1-μm diameter up to tens of microns. The squid contains a giant axon nearly 1 mm in diameter. This axon plays an important role in our understanding of how nerves work (Chap. 6). Our red blood cells (*erythrocytes*) carry oxygen to all parts of our body. (Actually, red blood cells are not true cells at all, but rather "corpuscles"). Red blood cells are disk-shaped, with a diameter of about 8 μm and a thickness of 2 μm (Fig. 1.1d). Blood flows through a branching network of vessels (Sect. 1.19), the smallest of which are *capillaries*. Each capillary has a diameter of about 8 μm, meaning that the red blood cells can barely pass through it single-file.

One valuable skill in physics is the ability to make order-of-magnitude estimates, meaning to calculate something approximately right. For instance, suppose we want to calculate the number of cells in the body. This is a difficult calculation, because cells come in all sizes and shapes. But for some purposes we only need an approximate answer (say, within a factor of ten). For example: cells are roughly 10 μm in size, so their volume is about $(10\ \mu\text{m})^3$, or $(10\times10^{-6})^3 = 10^{-15}\ \text{m}^3$. An adult is roughly 2 m tall and about 0.3 m wide, so our volume is about 2 m × 0.3 m × 0.3 m, or 0.18 m^3. We are made up almost entirely of cells, so the number of cells in our body is about $(0.18\ \text{m}^3)/(10^{-15}\ \text{m}^3)$, or roughly 2×10^{14}. Some problems at the end of the chapter ask you to make similar order-of-magnitude calculations.

Most cells are larger than a few microns. But many cells (called *eukaryotes*) are complex structures that contain *organelles* about this size. *Mitochondria*, organelles where many of the chemical processes providing cells with energy take place, are typically about 2 μm long. *Protoplasts*, organelles found in plant cells where photosynthesis changes light energy to chemical energy, are also about 2 μm long.

The simplest cells are called *prokaryotes* and contain no subcellular structures. *Bacteria* are the most common prokaryotic cells. The bacterium *E. coli* is about 2 μm long (Fig. 1.1e), and has been studied extensively.

To examine structures smaller than bacteria, we must measure lengths that are smaller than a micron. One-thousandth of a micron is called a nanometer (1 nm = 10^{-9} m). Figure 1.2 shows objects having lengths from 1 nm to 1 μm. *E. coli*, which seemed so tiny compared to cells in Fig. 1.1, are giants on the nanometer length scale, being 20 times longer than the 100-nm scale bar in Fig. 1.2. *Viruses* are tiny packets of genetic material encased in protein. On their own they are incapable of metabolism or reproduction, so some scientists do not even consider them as living organisms. Yet, they can infect a cell and take control of its metabolic and reproductive functions. The length scale of viruses is one-tenth of a micron, or 100 nm. For instance, *The human immunodeficiency virus (HIV)*, the virus that causes AIDS, is roughly spherical with a diameter of about 120 nm (Fig. 1.2a). Some viruses, called *bacteriophages*, infect and destroy bacteria. Most viruses are too small to see in a light microscope. The resolution of a microscope is limited by the wavelength of light, which is about 500 nm (Chap. 14). Thus, with a microscope we can study cells in detail, we can see bacteria without much resolution, and we can barely see viruses, if we can see them at all.

Below 100 nm, we enter the world of individual molecules. *Proteins* are large, complex macromolecules that are vitally important for life. For example, *hemoglobin* is the protein in red blood cells that binds to and carries oxygen. Hemoglobin is roughly spherical, about 6 nm in diameter (Fig. 1.2b). Many biological functions occur in the cell *membrane* (see Chap. 5). Membranes are made up of layers of *lipid* (fat), often with proteins and other molecules embedded in them (Fig. 1.2c). A typical cell membrane is about 10 nm

Table 1.2 Approximate sizes of biological objects

Object	Size
Protozoa	100 μm
Cells	10 μm
Bacteria	1 μm
Viruses	100 nm
Macromolecules	10 nm
Molecules	1 nm
Atoms	100 pm

thick. The molecule *adenosine triphosphate* (ATP), crucial for energy production and distribution in cells, is about 2 nm long (Chap. 3). Chemical energy is stored in molecules called *carbohydrates*. A common (and relatively small) carbohydrate is *glucose* ($C_6H_{12}O_6$), which is about 1 nm long (Fig. 1.2e). Genetic information is stored in long, helical strands of *deoxyribonucleic acid* (DNA). DNA is about 2.5 nm wide, and the helix completes a turn every 3.4 nm along its length (Fig. 1.2d).

At the 1-nm scale and below, we reach the world of small molecules and individual atoms. Water is the most common molecule in our body. It consists of two atoms of hydrogen and one of oxygen. The distance between adjacent atoms in water is about 0.1 nm. The distance 0.1 nm (100 pm) is used so much at atomic length scales that it has earned a nickname: the angstrom (A). Like the centimeter, this unit is going out of fashion as the use of nanometer becomes more common. Individual atoms have diameters of 100 or 200 pm.

Below the level of 100 pm, we leave the realm of biology and enter the world of subatomic physics. The nuclei of atoms (Chap. 17) are very small, and their sizes are measured in femtometers (1 fm $= 10^{-15}$ m).

One cannot possibly memorize the size of all biological objects: there are simply too many. The best one can do is remember a few mileposts along the way. Table 1.2 contains a rough guide to how large a few important biological objects are. Think of these as rules of thumb. Given the diversity of life, one can certainly find exceptions to these rules, but if you memorize Table 1.2 you will have a rough framework to organize your thinking about size. To examine the relative sizes of objects in more detail, see Morrison et al. (1994) or Goodsell (2009).

1.2 Models

Throughout this book we construct *mathematical models* of physical and biological systems. We start with general principles such as Newton's laws and apply them to a simplified model such as the leg in Fig. 1.3. The forces acting on the leg are much more complicated, but we model them with just three forces.

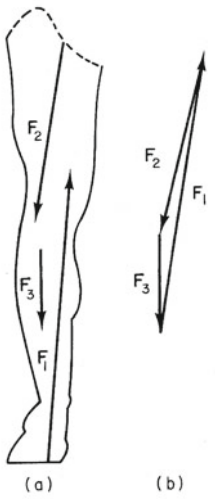


Fig. 1.3 Forces on the leg in equilibrium. Each force is exerted by some other object. **a** The points of application are widely separated. **b** The sum of the forces is zero

Biologists and physicists tend to make models differently (Blagoev et al. 2013). Biologists are used to dealing with complexity and diversity in biological systems. Physicists seek to explain as many phenomena with as few overarching principles as possible. Modeling a process is second nature to physicists. They willingly ignore some features of the biological system while seeking these principles. It takes experience and practice to decide what can be simplified and what can not.

A model incorporates some biological information, such as the ion currents in the Hodgkin–Huxley model (Chap. 6). The HH model can be extended by incorporating more ions and ion permeabilities measured in different nerves and muscles and other species. In other cases, the model has no specific details about the physiologic process, but captures an important feature in simplified form, that may have widespread applicability. We call this a *toy model*. The *radial isochron clock* (page 281) is a good example.

In many cases, simple models are developed in the homework problems at the end of each chapter. Working these problems will provide practice in the art of modeling.

1.3 Forces and Translational Equilibrium

There are several ways that we could introduce the idea of force, depending on the problem at hand and our philosophical bent. For our present purposes, it will suffice to say that a force is a *push* or a *pull*, that forces have both a magnitude and a direction, and that they give rise to accelerations through Newton's second law, $\mathbf{F} = m\mathbf{a}$. Experiments show that forces add like *displacements*, so they can be represented

by *vectors*. (Some of the properties of vectors are reviewed in Appendix B; others are introduced as needed.) Vectors will be denoted by boldfaced characters. The force is measured in newtons (N). A newton is a kg m s^{-2}.

One finds experimentally that an object is in *translational equilibrium* if the vector sum of all the forces acting on the body is zero. *Equilibrium* means that the object either remains at rest or continues to move with a constant velocity. That is, it is not accelerated. *Translational* means that only changes of position are being considered; changes of orientation of the object with respect to the axes are ignored.

We must consider *all* the forces that act *on* the object. If the object is a person standing on both feet, the forces are the upward force of the floor on each foot and the downward force of gravity on the person (more accurately, the vector sum of the gravitational force on every cell in the person). We do *not* consider the downward force that the person's feet exert *on the floor*. It is also possible to replace the sum of the gravitational force on each cell of the body with a single downward gravitational force acting at one point, the *center of gravity* of the body.

The forces that add to zero to give translational equilibrium need not all act at one point on the object. If the object is a person's leg and the leg is at rest, there are three forces exerted on the leg by other objects (Fig. 1.3). Force $\mathbf{F}_1$ is the push of the floor up on the bottom of the foot. The various pushes and pulls of the rest of the body on the leg through the hip joint and surrounding muscles have been added together to give $\mathbf{F}_2$. The gravitational pull of the earth downward on the leg is $\mathbf{F}_3$. Force $\mathbf{F}_1$ acts on the bottom of the leg, $\mathbf{F}_2$ acts on the top, and $\mathbf{F}_3$ acts somewhere in between. If the leg is in equilibrium the sum of these forces is zero, as shown in Fig. 1.3b. Although the points of application of the forces can be ignored in considering translational equilibrium, they are important in determining whether or not the object is in rotational equilibrium. This is discussed shortly.

The Greek letter Σ (capital sigma) is usually used to mean a sum of things. With this notation, the condition for translational equilibrium can be written as

$$\sum_i \mathbf{F}_i = 0. \tag{1.1}$$

The subscript i is used to label the different forces acting on the body. A notation this compact has a lot hidden in it. This is a vector equation, standing for three equations:

$$\sum_i F_{ix} = 0,$$

$$\sum_i F_{iy} = 0, \tag{1.2}$$

$$\sum_i F_{iz} = 0.$$

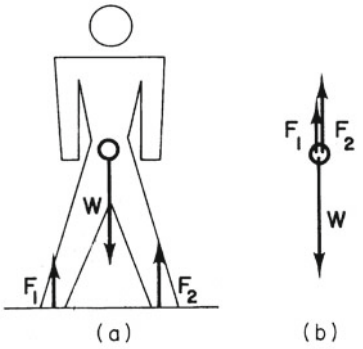


Fig. 1.4 A person standing. **a** The forces on the person. **b** A free-body or force diagram

Often the subscript i is omitted and the equations are written as $\sum F_x = 0$, $\sum F_y = 0$, and $\sum F_z = 0$. In this notation, a component is positive if it points along the positive axis and negative if it points the other way.

Sometimes, as in the next example, we draw forces in particular directions and assume that these directions are positive. If the subsequent algebra happens to give a solution that is negative, the force points opposite the direction assumed.

As an example, consider a person standing on both feet as in Fig. 1.4. The earth pulls down with force $\mathbf{W}$. The floor pushes up on the right foot with force $\mathbf{F}_1$ and on the left foot with force $\mathbf{F}_2$. To determine what the condition for translational equilibrium tells us about the forces, draw the force diagram or *free-body diagram* of Fig. 1.4b. This diagram is an abstraction that ignores the points at which the forces are applied to the body. We can get away with this abstraction because we are considering only translation. When we consider rotational equilibrium, we will have to redraw the diagram showing the points at which the various forces act on the person. If all the forces are vertical, then there is only one component of each force to worry about, and the equilibrium condition gives $F_1 + F_2 - W = 0$ or $F_1 + F_2 = W$. The total force of the floor pushing up on both feet is equal to the pull of the earth down.

If there is a sideways force on each foot, translational equilibrium provides two conditions: $F_{1x} + F_{2x} = 0$ and $F_{1y} + F_{2y} - W = 0$.

This is all that can be learned from the condition for translational equilibrium. If the person stands on one foot, then $F_1 = 0$ and $F_2 = W$. If the person stands with equal force on each foot, then $F_1 = F_2 = W/2$.

1.4 Rotational Equilibrium

If the object is in *rotational equilibrium*, then another condition must be placed upon the forces. Rotational equilibrium

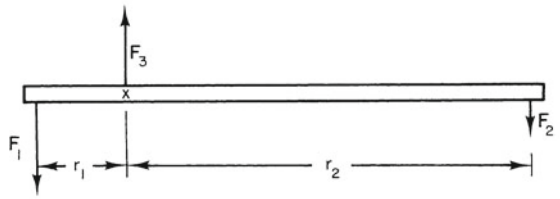


Fig. 1.5 A rigid rod free to rotate about a pivot at point X

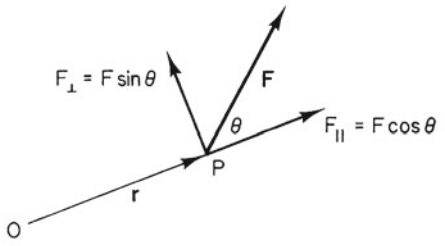


Fig. 1.6 A force **F** is applied to an object at point P. The object can rotate about point O. Vectors **r** and **F** determine the plane of the paper

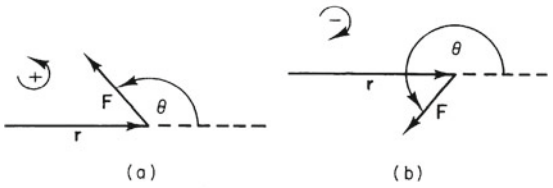


Fig. 1.7 **a** When θ is between 0 and 180°, both $\sin\theta$ and the torque are positive. **b** When θ is between 180 and 360°, both $\sin\theta$ and the torque are negative

means that the object either does not rotate or continues to rotate at a constant rate (with a constant number of rotations per second). Consider the object of Fig. 1.5, which is a rigid rod pivoted at point X so that it can rotate in the plane of the paper. Forces $\mathbf{F}_1$ and $\mathbf{F}_2$ are applied to the rod in the plane of the paper at distances r_1 and r_2 from the pivot and perpendicular to the rod. The pivot exerts the force $\mathbf{F}_3$ on the rod needed to maintain translational equilibrium. If both $\mathbf{F}_1$ and $\mathbf{F}_2$ are perpendicular to the rod, they are parallel to each other. They must also be parallel to $\mathbf{F}_3$, and translational equilibrium requires that $\mathbf{F}_3 = \mathbf{F}_1 + \mathbf{F}_2$.

Experiment shows that there is no rotation of the rod if $F_1 r_1 = F_2 r_2$. The condition for rotational equilibrium can be stated in a form analogous to that for translational equilibrium if we define the *torque*, τ, to be

$$\tau_i = r_i F_i . \tag{1.3}$$

With this definition goes an algebraic sign convention: the torque is positive if it tends to produce a counterclockwise rotation. The rod is in rotational equilibrium if the algebraic sum of all the torques is zero:

$$\sum_i \tau_i = \sum_i r_i F_i = 0. \tag{1.4}$$

Note that $\mathbf{F}_3$ contributes nothing to the torque because r_3 is zero.

The torque is defined about a certain point, X. It depends on the distance from the point of application of each force to X.[3] As long as the object is in translational equilibrium, the torque can be evaluated around any point. This theorem, which we will not prove, often allows calculations to be simplified, because taking torques about certain points can cause some forces not to contribute to the torque equation.

The torque can also be calculated if the force is not at right angles to the rod. Imagine an object free to rotate about point O in Fig. 1.6. Force **F** lies in the plane of the paper but is applied in some arbitrary direction at point P. The vectors **r** and **F** determine the plane of the paper if they are not parallel. Force **F** can be resolved into two components: one parallel to **r**, $F_\parallel = F\cos\theta$, and the other perpendicular to **r**, $F_\perp = F\sin\theta$. The component parallel to **r** will not cause any rotation about point O. (Pull on an open door parallel to the plane of the door; there is no rotation.) The torque is therefore

$$\tau = r F_\perp = r F \sin\theta. \tag{1.5}$$

The perpendicular distance from the line along which the force acts to point O is $r\sin\theta$. It is often called the *moment arm*, and the torque is the magnitude of the force multiplied by the moment arm.

The angle θ is the angle of rotation from the direction of **r** to the direction of **F**. It is called positive if the rotation is counterclockwise. For the angle shown in Fig. 1.6, $\sin\theta$ has a positive value, and the torque is positive. Figure 1.7a shows an angle between 90 and 180° for which the torque and $\sin\theta$ are still positive. Figure 1.7b shows an angle between 180 and 360°, for which both the torque and $\sin\theta$ are negative. In all cases, Eq. 1.5 gives the correct sign for the torque.

To summarize: the torque due to force **F** applied to a body at point P must be calculated about some point O. If **r** is the vector from O to P, the magnitude of the torque is equal to the magnitude of **r** times the magnitude of **F** times the sine of the angle between **r** and **F**. The angle is measured counterclockwise from **r** to **F**.

[3] The discussion associated with Fig. 1.5 suggests that torque is taken about an axis, rather than a point. In a three-dimensional problem the torque is taken about a point.

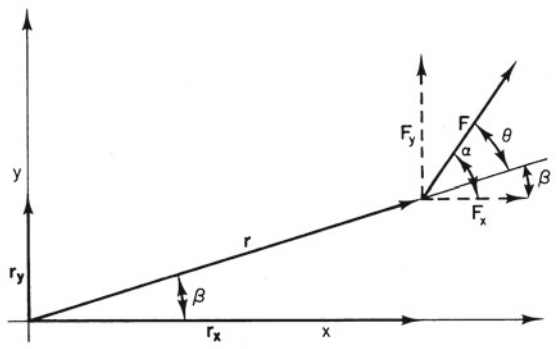


Fig. 1.8 The cross product $\mathbf{r} \times \mathbf{F}$ is calculated by resolving $\mathbf{r}$ and $\mathbf{F}$ into components

1.5 Vector Product

Torque can be thought of as a vector, $\boldsymbol{\tau}$. Its magnitude is $Fr \sin\theta$. The only direction uniquely defined by vectors $\mathbf{r}$ and $\mathbf{F}$ is perpendicular to the plane in which they lie. This is also the direction of an axis about which the torque would cause a rotation. However, there is ambiguity about which direction along this line to assign to the torque. The convention is to say that a positive torque points in the direction of the thumb of the right hand when the fingers curl in the direction of positive rotation from $\mathbf{r}$ to $\mathbf{F}$.[4] When $\mathbf{r}$ and $\mathbf{F}$ point in the same direction, so that no plane is defined, the magnitude of the torque is zero.

The product of two vectors according to the foregoing rules is called the *cross product* or *vector product* of the two vectors. One can use a shorthand notation,

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}. \tag{1.6}$$

There is another way to write the cross product. If both $\mathbf{r}$ and $\mathbf{F}$ are resolved into components, as shown in Fig. 1.8, then the cross product can be calculated by applying the rules above to the components. Since $\mathbf{F}_y$ is perpendicular to $\mathbf{r}_x$ and parallel to $\mathbf{r}_y$, its only contribution is a counterclockwise torque $r_x F_y$. The only contribution from $\mathbf{F}_x$ is a clockwise torque, $-r_y F_x$. The magnitude of the cross product is therefore

$$\tau = r_x F_y - r_y F_x. \tag{1.7}$$

Note that this is the (signed) sum of each component of the force multiplied by its moment arm.

[4] This arbitrariness in assigning the sense of $\boldsymbol{\tau}$ means that it does not have quite all the properties that vectors usually have. It is called an axial vector or a pseudovector. It will not be necessary in this book to worry about the difference between a real vector and an axial vector.

The equivalence of this result to Eq. 1.5 can be verified by writing Eq. 1.7 as

$$\tau = (r\cos\beta)(F\sin\alpha) - (r\sin\beta)(F\cos\alpha),$$

$$\tau = rF\,(\sin\alpha\cos\beta - \cos\alpha\sin\beta)\,.$$

There is a trigonometric identity that

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta.$$

Since $\theta = \alpha - \beta$ (from Fig. 1.8), this is equivalent to $\tau = rF\sin\theta$.

When vectors $\mathbf{r}$ and $\mathbf{F}$ lie in the xy plane, $\boldsymbol{\tau}$ points along the z axis. If $\mathbf{r}$ and $\mathbf{F}$ point in arbitrary directions, Eq. 1.7 gives the z component of $\boldsymbol{\tau}$. One can apply the same reasoning for other components and show that

$$\begin{aligned} \tau_x &= r_y F_z - r_z F_y, \\ \tau_y &= r_z F_x - r_x F_z, \\ \tau_z &= r_x F_y - r_y F_x. \end{aligned} \tag{1.8}$$

If you are familiar with the rules for evaluating determinants, you will see that this is equivalent to the notation,

$$\boldsymbol{\tau} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ r_x & r_y & r_z \\ F_x & F_y & F_z \end{vmatrix}. \tag{1.9}$$

1.6 Force in the Achilles Tendon

The equilibrium conditions can be used to understand many problems in clinical orthopedics. Two are discussed in this book: forces that sometimes cause the *Achilles tendon* at the back of the heel to break, and forces in the hip joint.

The Achilles tendon connects the calf muscles (the gastrocnemius and the soleus) to the calcaneus at the back of the heel (Fig. 1.9). To calculate the force exerted by this tendon on the calcaneus when a person is standing on the ball of one foot, assume that the entire foot can be regarded as a rigid body. This is our first example of creating a *model* of the actual situation. We try to simplify the real situation to make the calculation possible while keeping the features that are important to what is happening. In this model, the internal forces within the foot are being ignored.

Figure 1.10 shows the force exerted by the tendon on the foot ($\mathbf{F}_T$), the force of the leg bones (tibia and fibula) on the foot ($\mathbf{F}_B$), and the force of the floor upward, which is equal to the weight of the body ($\mathbf{W}$). The weight of the foot is small compared to these forces and will be neglected. Measurements on a few people suggest that the angle the Achilles tendon makes with the vertical is about 7 °.

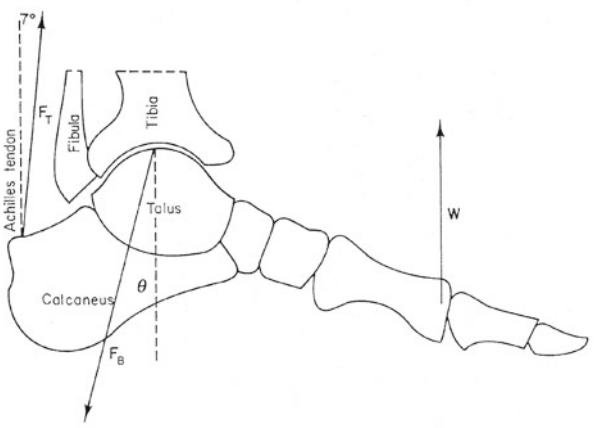


Fig. 1.9 Simplified anatomy of the foot

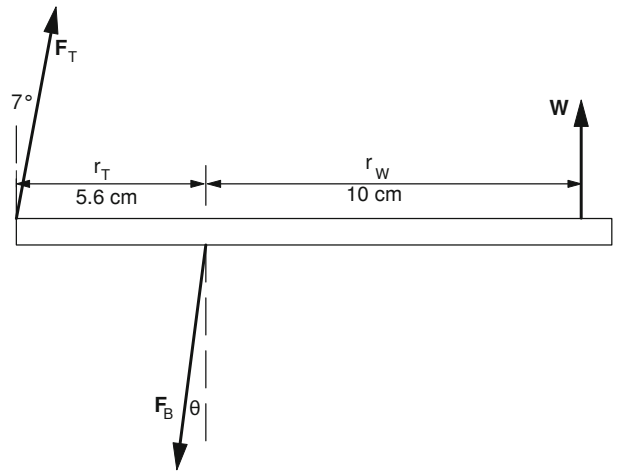


Fig. 1.10 Forces on the foot, neglecting its own weight

Translational equilibrium requires that

$$F_T \cos(7^\circ) + W - F_B \cos\theta = 0, \qquad (1.10)$$
$$F_T \sin(7^\circ) - F_B \sin\theta = 0.$$

To write the condition for rotational equilibrium, we need to know the lengths of the appropriate vectors $\mathbf{r}_T$ and $\mathbf{r}_W$, assuming that the torques are taken about the point where $\mathbf{F}_B$ is applied to the foot. In our simple model, we ignore the contributions of the horizontal components of any forces to the torque equation. This is not essential (if we are willing to make more detailed measurements), but it simplifies the equations and thereby makes the process clearer. The horizontal distances measured by one of the authors are $r_T = 5.6$ cm and $r_W = 10$ cm, as shown in Fig. 1.10. The torque equation is

$$10W - 5.6F_T \cos 7^\circ = 0. \qquad (1.11)$$

This equation can be solved for the tension in the tendon:

$$F_T = \frac{10W}{5.6\cos 7^\circ} = 1.8W. \qquad (1.12)$$

This result can now be used in Eq. 1.10 to find $F_{By} = F_B \cos\theta$:

$$(1.8)(W)(0.993) + W = F_B \cos\theta,$$
$$2.8W = F_B \cos\theta. \qquad (1.13)$$

From Eqs. 1.10 and 1.12, we get

$$(1.8)(W)(0.122) = F_B \sin\theta,$$
$$0.22W = F_B \sin\theta. \qquad (1.14)$$

Equations 1.13 and 1.14 are squared and summed and the square root taken to give $F_B = 2.8W$, while they can be divided to give

$$\tan\theta = \frac{0.22}{2.8} = 0.079,$$
$$\theta = 4.5^\circ.$$

The tension in the Achilles tendon is nearly twice the person's weight, while the force exerted on the leg by the talus is nearly three times the body weight. One can understand why the tendon might rupture.

1.7 Forces on the Hip

The forces in the hip joint can be several times a person's weight, and the use of a cane can be very effective in reducing them.

As a person walks, there are moments when only one foot is on the ground. There are then two forces acting on the body as a whole: the downward pull of the earth W and the upward push of the ground on the foot N. The pull of the earth may be regarded as acting at the center of gravity of the body (Serway and Jewett 2013, p. 219). The center of gravity is located on the midline (if the limbs are placed symmetrically), usually in the lower abdomen (Williams and Lissner 1962, Chap. 5.) If torques are taken about the foot, then the center of gravity must be directly over the foot so that there will be no torque from either force. This situation is shown in Fig. 1.11. The condition for translational equilibrium requires that $N = W$.

The anatomy of the pelvis, hip, and leg is shown schematically in Fig. 1.12. Fourteen muscles and several ligaments connect the pelvis to the femur. Extensive measurements of the forces exerted by the abductor[5] muscles in the hip have

[5] *To abduct* means to move away from the midline of the body.

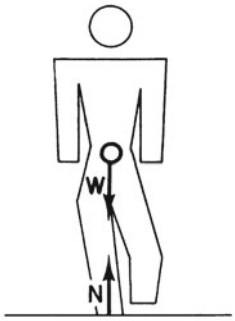


Fig. 1.11 A person standing on one foot must place the foot under the center of gravity, which is on or near the midline

been made by Inman (1947). If the leg is considered an isolated system as in Fig. 1.12, the following forces act:

F: The net force of the abductor muscles, acting on the greater trochanter. These muscles are primarily the gluteus medius and gluteus minimus, shown as a single band of muscle in Fig. 1.12.

R: The force of the acetabulum (the socket of the pelvis) on the head of the femur.

N: The upward force of the floor on the bottom of the foot (in this case, equal to W).

$\mathbf{W}_L$: The weight of the leg, acting vertically downward at the center of gravity of the leg. $W_L \approx W/7$ (Williams and Lissner 1962, Chap. 5).

Inman found that **F** acts at about a 70° angle to the horizontal. In a typical adult, the distance from the greater trochanter to the midline is about 18 cm, the horizontal distance from the greater trochanter to the center of gravity of the leg is about 10 cm, and the distance from the greater trochanter to the middle of the head of the femur is about 7 cm.

A free body diagram is shown in Fig. 1.13. The middle of the head of the femur will turn out to be very close to the intersection of the line along which **R** acts and a horizontal line drawn from the point where **F** acts. This means that if torques are taken about this intersection point (point O), there will be no contributions from **R** or from the horizontal component of **F**. The intersection is about 7 cm toward the midline from the point of application of **F**. Since $N = W$ and $W_L \approx W/7$, the equilibrium equations are

$$\sum F_y = F\sin(70^\circ) - R_y - W/7 + W = 0, \qquad (1.15)$$

$$\sum F_x = F\cos(70^\circ) - R_x = 0, \qquad (1.16)$$

$$\sum \tau = -F\sin(70^\circ)(7) - (W/7)(10-7) + W(18-7) = 0.$$

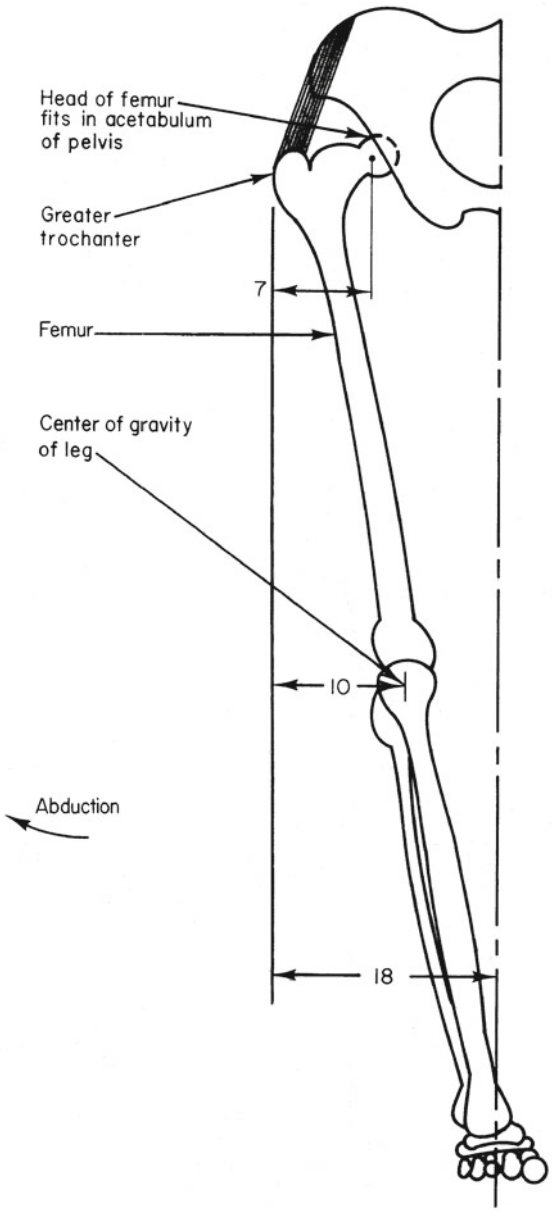


Fig. 1.12 Pertinent features of the anatomy of the leg

The last of these equations can be written as $11W - \frac{3}{7}W - 6.6F = 0$, from which $F = 1.6W$. The magnitude of the force in the abductor muscles is about 1.6 times the body weight.

Equations 1.15 and 1.16 can now be used to find R_x and R_y:

$$R_x = F\cos(70^\circ) = (1.6W)(0.342) = 0.55W,$$

$$R_y = F\sin(70^\circ) + \frac{6}{7}W = (1.6W)(0.94) + 0.86W = 2.36W.$$

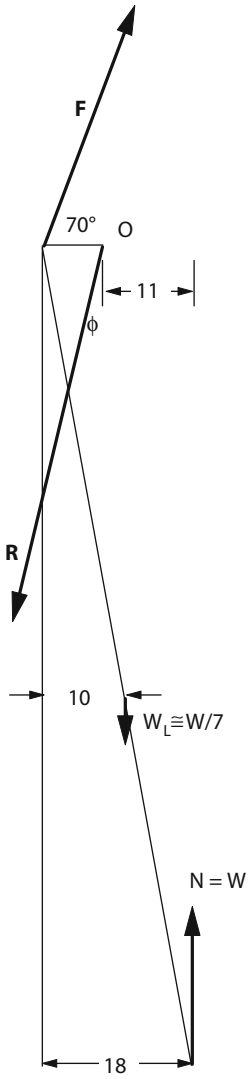


Fig. 1.13 A free-body diagram of the forces acting on the leg. Torques are taken about point O, which is the intersection of a line along which **R** acts and a horizontal line through the point at which **F** is applied. This point is 7 cm toward the midline (medially) from the greater trochanter

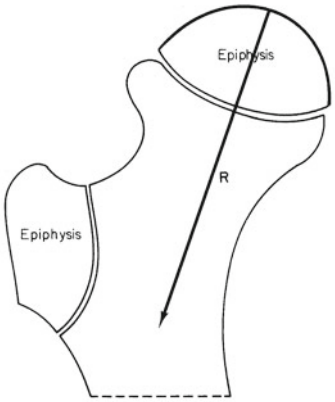


Fig. 1.14 The femoral epiphysis and the direction of **R**



Fig. 1.15 X-ray of a slipped femoral epiphysis in an adolescent male. (Courtesy of the Department of Diagnostic Radiology, University of Minnesota)

The angle that **R** makes with the vertical is given by

$$\tan\phi = \frac{R_x}{R_y} = 0.23,$$

$$\phi = 13^\circ.$$

The magnitude of **R** is $R = (R_x^2 + R_y^2)^{1/2} = 2.4W$.

If the patient did not have to put the foot under the center of gravity of the body, the moment arm of the only positive torque, $11W$, could have been much less, and this would have been balanced by a smaller value of F. This can be done by having the patient use a cane on the *opposite* side, so that the foot need not be right under the center of gravity. This will be explored in the next section. Conversely, if the patient were carrying a suitcase in the opposite hand, the center of mass would be moved away from the midline, the foot would still have to be placed under the center of mass, and the moment arm, and hence F, would be even larger (Problem 11).

One very interesting conclusion of Inman's study was that the force **R** always acts along the neck of the femur in such a direction that the femoral epiphysis has very little sideways force on it. The epiphysis is the growing portion of the bone (Fig. 1.14) and is not very well attached to the rest of the bone. If there were an appreciable sideways force, the epiphysis would slip sideways, and indeed it sometimes does (Fig. 1.15). This is a serious problem, since if the blood supply to the epiphysis is compromised, there will be no more bone growth.

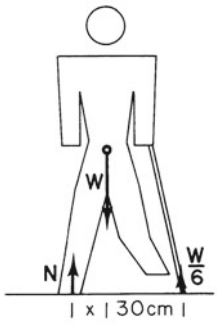


Fig. 1.16 A person using a cane on the left side (*front view*) to favor the right hip

Suppose that, for some reason, the gluteal muscles are severed. The patient can no longer apply force **F** to the greater trochanter; Eq. 1.16 shows that then R_x must be zero. This change in the direction of **R** causes a rotation of the epiphyseal plate and a gradual reshaping of the femur.

1.8 The Use of a Cane

A cane is beneficial if used on the side opposite to the affected hip (Fig. 1.16). We ignore the fact that the arm holding the cane has moved, thereby shifting slightly the center of mass, and we assume that the force of the ground on the cane is vertical. If we assume that the tip of the cane is about 30 cm (12 in.) from the midline and supports one-sixth of the body weight, then we can apply the equilibrium conditions to learn that $N + \frac{1}{6}W - W = 0$, so $N = \frac{5}{6}W$. Torques taken about the center of mass give $(30)(\frac{W}{6}) - x(\frac{5}{6})W = 0$, $x = 6$ cm. (Figure 1.16 is not to scale.)

Having the foot 6 cm from the midline reduces the force in the muscle and the joint. To find out how much, consider the force diagram in Fig. 1.17. The most difficult part of the problem is working out the various moment arms. Assume that the slight movement of the leg has not changed the point about which we take torques (point O). Again, **R** contributes no torque about this point. The horizontal distance of **F** from this point is still 7 cm. The force of the ground on the leg is now $5W/6$, and its moment arm is $18-6-7=5$ cm. The weight of the leg, $W/7$, acts at the center of mass of the leg, which is still $\frac{10}{18}$ of the distance from the greater trochanter to the foot. Its horizontal position is therefore $\frac{10}{18}$ of the horizontal distance from the greater trochanter to the foot: $(10)(12)/18 = 6.67$ cm. The moment arm is $7-6.67$ cm $= 0.33$ cm. The torque equation is

$$-F\sin(70^\circ)(7) + \left(\frac{W}{7}\right)(0.33) + \left(\frac{5W}{6}\right)(5) = 0.$$

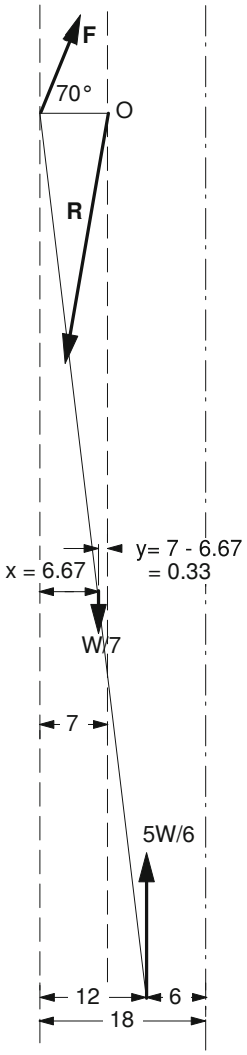


Fig. 1.17 A force diagram for the leg when a cane is being used and the leg is 6 cm from the midline

It is solved by writing it as

$$-6.58F + 0.047W + 4.17W = 0,$$
$$F = 0.64W.$$

Even though the cane supports only one-sixth of the body weight, F has been reduced from $1.6W$ to $0.64W$ by the change in the moment arm.

The force of the acetabulum on the head of the femur can be determined from the conditions for translational equilibrium:

$$F\cos(70^\circ) - R_x = 0,$$
$$R_x = 0.22W,$$
$$F\sin(70^\circ) - R_y - \frac{W}{7} + \frac{5}{6}W = 0,$$
$$R_y = 1.29W.$$

The resultant force **R** has magnitude $(R_x^2 + R_y^2)^{1/2} = 1.3W$. This compares to the value $2.4W$ without the cane. The force in the joint has been reduced by slightly more than the body weight. It is interesting to read what an orthopedic surgeon had to say about the use of a cane. The following is from the presidential address of W. P. Blount, M.D., to the Annual Meeting of the American Academy of Orthopedic Surgeons, January 30, 1956:

> The patient with a wise orthopedic surgeon walks with crutches for six months after a fracture of the neck of the femur. He uses a stick for a longer time—the wiser the doctor, the longer the time. If his medical adviser, his physical therapist, his friends, and his pride finally drive him to abandon the cane while he still needs one, he limps. He limps in a subconscious effort to reduce the strain on the weakened hip. If there is restricted motion, he cannot shift his body weight, but he hurries to remove the weight from the painful hip joint when his pride makes him reduce the limp to a minimum. The excessive force pressing on the aging hip takes its toll in producing degenerative changes. He should not have thrown away the stick.[6]

1.9 Work

So far this chapter has considered only those situations in which a mass m is in equilibrium. If the total force on the object is not zero, the object experiences an acceleration **a** given by Newton's second law:

$$\mathbf{F} = m\mathbf{a}.$$

The study of how forces produce accelerations is called *dynamics*. It is an extensive field that will be discussed only briefly here.

Suppose the object moves along the x axis with velocity v_x. If it is subject to a force in the x direction F_x, it will be accelerated, and the velocity will change according to $F_x = ma_x = m\,(dv_x/dt)$. If F_x is known as a function of time, then this equation can be written as $dv_x = (1/m)\,F_x(t)dt$, and it can be integrated, at least numerically.

In this context it is useful to define the *kinetic energy*

$$E_k = \frac{1}{2}mv_x^2. \tag{1.17}$$

As long as F_x acts, the object is accelerated and the kinetic energy changes. We can gain some understanding of how it changes by noting that

$$\frac{d}{dt}\left(\frac{1}{2}mv_x^2\right) = mv_x\frac{dv_x}{dt} = F_xv_x. \tag{1.18}$$

Therefore F_xv_x is the rate at which the kinetic energy is changing with time. It is called the *power* due to force F_x. The units of kinetic energy are $\mathrm{kg\,m^2\,s^{-2}}$ or joules (J); the units of power are $\mathrm{J\,s^{-1}}$ or watts (W).

If v_x and F_x are both positive, the acceleration increases the object's velocity, the kinetic energy increases, and the power is positive. If v_x and F_x are both negative, v_x decreases—becomes more negative—but the magnitude of the velocity increases. The kinetic energy increases with time, and the power is positive. If v_x and F_x point in opposite directions, then the effect of the acceleration is to reduce the magnitude of v_x, the kinetic energy decreases, and the power is negative.

Equation 1.18 can be written as

$$\frac{d}{dt}\left(\frac{1}{2}mv_x^2\right) = F_x\frac{dx}{dt}.$$

Both sides of this equation can be integrated with respect to t:

$$\int_{t_1}^{t_2}\frac{d}{dt}\left(\frac{1}{2}mv_x^2\right)dt = \int_{t_1}^{t_2}F_x\,(t)\,\frac{dx}{dt}\,dt.$$

The indefinite integral corresponding to the left-hand side is the integral with respect to time of the derivative of $\frac{1}{2}mv_x^2$ and is therefore $\frac{1}{2}mv_x^2$. If F_x is known not as a function of t but as a function of x, it is convenient to write the right-hand side as

$$\int_{x_1}^{x_2}F_x(x)\,dx = W.$$

This quantity is called the *work* done by force F_x on the object as it moves from x_1 to x_2. The complete equation is therefore

$$\left[\frac{1}{2}mv_x^2\right]_2 - \left[\frac{1}{2}mv_x^2\right]_1 = \int_{x_1}^{x_2}F_x(x)\,dx = W. \tag{1.19}$$

The increase in kinetic energy of the body as it moves from position 1 (at time 1) to position 2 (at time 2) is equal to the work done *on* the body *by* the force F_x. The work done on the body by force F_x is the area under the curve of F_x versus x, between points x_1 and x_2. This is shown in Fig. 1.18.

If several forces act on the body, then the acceleration is given by Newton's second law, where **F** is the *total* force on the body. The change in kinetic energy is therefore the work done by the total force or the sum of the work done by each individual force.

When the force and displacement vectors point in any direction, the kinetic energy is defined to be

$$E_k = \frac{1}{2}mv^2 = \frac{1}{2}m(v_x^2 + v_y^2 + v_z^2). \tag{1.20}$$

Differentiating this expression with respect to time shows that the power is given by an extension of Eq. 1.18:

$$\frac{dE_k}{dt} = F_xv_x + F_yv_y + F_zv_z.$$

[6] Quoted with permission from Blount (1956). Copyright © 1956 *J Bone Joint Surg.* This article was first quoted to the physics community by Benedek and Villars (1973)

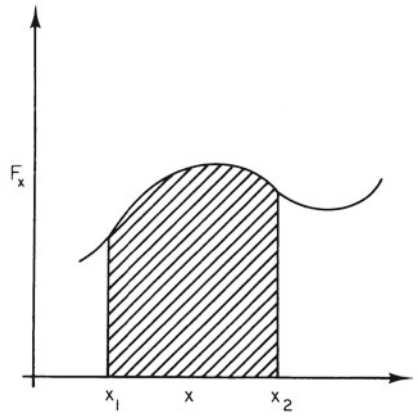


Fig. 1.18 The work done by F_x is the shaded area under the curve between x_1 and x_2

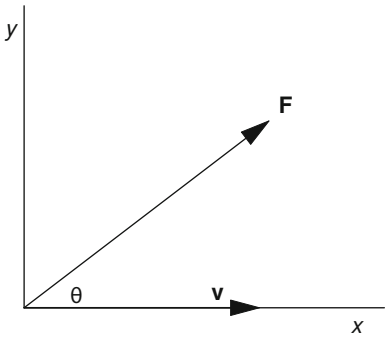


Fig. 1.19 Aligning the axes so that **v** is along the x axis and **F** in the xy plane shows that an alternative expression for $\mathbf{F}\cdot\mathbf{v}$ is $Fv\cos\theta$

This particular combination of vectors **F** and **v** is called the *scalar product* or *dot product*. It is written as $\mathbf{F}\cdot\mathbf{v}$.

There is another way to write the scalar product. If **F** and **v** are not parallel, they define a plane. Align the x axis with **v** so that v_y and v_z are zero, and choose the direction of y so that **F** is in the xy plane (Fig. 1.19). Then it is easy to see that $\mathbf{F}\cdot\mathbf{v} = F_x v_x = Fv\cos\theta$, where θ is the angle between **F** and **v**.

To summarize, the power is

$$P = \frac{dE_k}{dt} = \mathbf{F}\cdot\mathbf{v} = Fv\cos\theta = F_x v_x + F_y v_y + F_z v_z. \tag{1.21}$$

Equation 1.21 can be integrated in the same manner as above to obtain

$$\Delta E_k = \int F_x\,dx + \int F_y\,dy + \int F_z\,dz = \int \mathbf{F}\cdot d\mathbf{s}. \tag{1.22}$$

This is the general expression for the work done by force **F** on a point mass that undergoes displacement **s**.

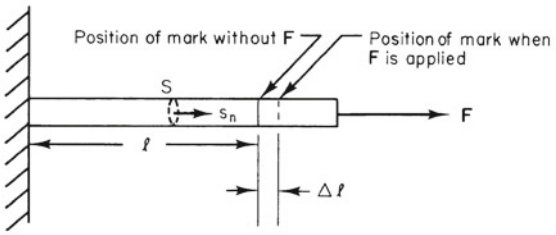


Fig. 1.20 A rod subject to a force **F** along it

1.10 Stress and Strain

Whenever a force acts on an object, it undergoes a change of shape or *deformation*. Often these deformations can be ignored, as they were in the previous sections. In other cases, such as the contraction of a muscle, the expansion of the lungs, or the propagation of a sound wave, the deformation is central to the problem and must be considered. This book will not develop the properties of deformable bodies extensively; nevertheless, deformable body mechanics is important in many areas of biology (Fung 1993). We will develop the subject only enough to be able to consider viscous forces in fluids.

Consider a rod of cross-sectional area S. One end is anchored, and a force F is exerted on the other end parallel to the rod (Fig. 1.20). Effects of weight will be ignored. A *surface force* is transmitted across any surface defined by an imaginary cut perpendicular to the axis of the rod. A surface force is exerted by the substance to the right of the cut on the substance to the left (and vice versa, in accordance with Newton's third law: when object A exerts a force on object B, object B exerts an equal and opposite force on object A). The surface force per unit area is called the *stress*. In this case, when the surface is perpendicular to the axis of the rod and the force is along the axis of the rod, it is called a *normal stress*:

$$s_n = \frac{F}{S}. \tag{1.23}$$

In the general case there can also be a component of stress parallel to the surface.

The *strain* ϵ_n is the fractional change in the length of the rod:

$$\epsilon_n = \frac{\Delta l}{l}. \tag{1.24}$$

If increasing stress is applied to a typical substance, the strain increases linearly with the stress for small stresses. Then it increases even more rapidly. At higher strains it may be necessary to reduce the stress to maintain the same strain. If the stress is not reduced, the rod elongates further and breaks. Finally, at a high enough strain, the sample

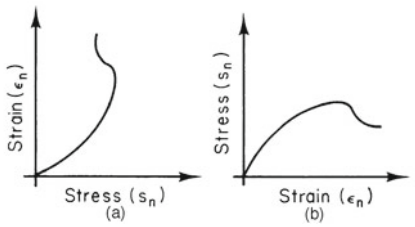


Fig. 1.21 A typical stress–strain relationship. On the left, stress is the independent variable. On the right, strain is the independent variable. Strain is usually used as the independent variable because it is often a double-valued function of the stress

Table 1.3 Young's modulus, tensile strength, and compressive strength of various materials in pascal

Material	E	Tensile strength	Compressive strength
Steel (approx.)[a]	20×10^{10}	50×10^7	–
Femur (wet)[b]	1.4×10^{10}	8.3×10^7	1.8×10^7
Walnut[c]	0.8×10^{10}	4.1×10^7	5.2×10^7

[a] *American Institute of Physics Handbook* (1957). New York, McGraw-Hill, p. 2–70
[b] cf. Kummer (1972)
[c] cf. U.S. Department of Agriculture (1955)

breaks. This is plotted in Fig. 1.21. Because of the double-valuedness of the strain as a function of stress, the strain is usually plotted as the independent variable, as on the right in Fig. 1.21.

In the linear region, the relationship between stress and strain is written as

$$s_n = E\epsilon_n. \tag{1.25}$$

The proportionality constant E is called *Young's modulus*. Since the strain is dimensionless, E has the dimensions of stress. Various units are $\mathrm{N\,m^{-2}}$ or pascal (Pa), $\mathrm{dyn\,cm^{-2}}$, psi (pound per square inch), and bar ($1\ \mathrm{bar} = 14.5\ \mathrm{psi} = 10^5\ \mathrm{Pa} = 10^6\ \mathrm{dyn\,cm^{-2}}$).

If the stress is increased enough, the bar breaks. The value of the stress when the bar breaks under tension is called the *tensile strength*. The material will also rupture under compressive stress; the rupture value is called the *compressive strength*. Table 1.3 gives values of Young's modulus, the tensile strength, and the compressive strength for steel, long bone (femur), and wood (walnut).

In some materials, the stress depends not only on the strain, but on the rate at which the strain is produced. It may take more stress to stretch the material rapidly than to stretch it slowly, and more stress to stretch it than to maintain a fixed strain. Such materials are called *viscoelastic*. They are often important biologically but will not be discussed here (Fung 1993).

Still other materials exhibit *hysteresis*. The stress–strain relationship is different when the material is being stretched than when it is allowed to return to its unstretched state. This difference is observed even if the strain is changed so slowly that viscoelastic effects are unimportant.

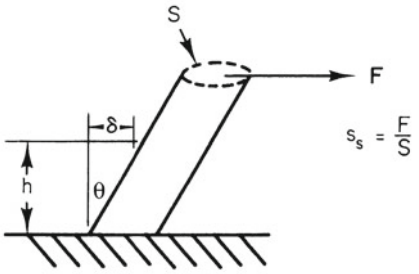


Fig. 1.22 Shear stress and strain

1.11 Shear

In a *shear stress*, the force is parallel to the surface across which it is transmitted.[7] In a *shear strain*, the deformation increases as one moves in a direction perpendicular to the deformation. An example of shear stress and strain is shown in Fig. 1.22. The shear stress is

$$s_s = \frac{F}{S}, \tag{1.26}$$

and the shear strain is

$$\epsilon_s = \frac{\delta}{h}. \tag{1.27}$$

It is possible to define a *shear modulus* G analogous to Young's modulus when the shear strain is small:

$$s_s = G\epsilon_s. \tag{1.28}$$

1.12 Hydrostatics

We now turn to some topics in the mechanics of fluids that will be useful for understanding several phenomena, including the circulation and fluid movement through membranes in Chap. 5. *Hydrostatics* is the description of fluids at rest. A fluid is a substance that will not support a shear when it is at rest. When the fluid is in motion, there can be a shear force arising from *viscosity*.

[7] This discussion of stress and strain has been made simpler than is often the case. In general, the force **F** across any surface is a vector. It can be resolved into a component perpendicular to the surface and two components parallel to the surface. One can speak of nine components of stress: $s_{xx}, s_{xy}, s_{xz}, s_{yx}, s_{yy}, s_{yz}, s_{zx}, s_{zy}, s_{zz}$. The first subscript denotes the direction of the force and the second denotes the normal to the surface across which the force acts. Components s_{xx}, s_{yy}, and s_{zz} are normal stresses; the others are shear stresses. It can be shown that $s_{xy} = s_{yx}$, and so forth.

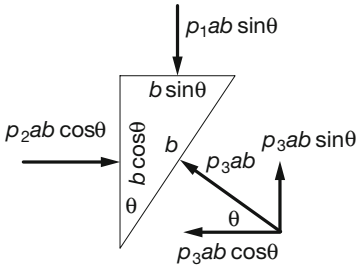


Fig. 1.23 A volume element of fluid used to show that the pressure in a fluid at rest is the same in all directions

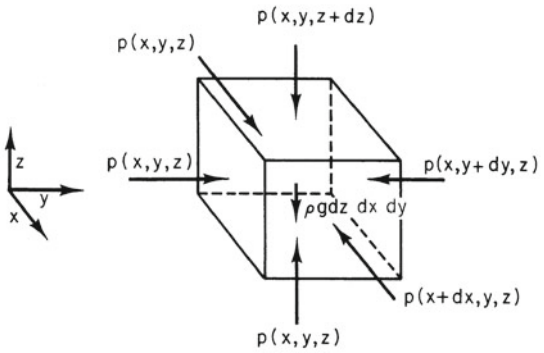


Fig. 1.24 The fluid in volume $dxdydz$ is in equilibrium

An immediate consequence of the definition of a fluid is that when the fluid is at rest, all the stress is normal. The normal stress is called the *pressure*. The pressure at any point in the fluid is the same in all directions. This can be demonstrated experimentally, and it can be derived from the conditions for equilibrium. Consider the small volume of fluid shown in Fig. 1.23. It has a length a perpendicular to the page. This volume is in equilibrium. Since the fluid at rest cannot support a shear, the pressure is perpendicular to each face, and there is no other force across each face. To prove this, assume that the pressures perpendicular to the three faces can be different, and call them p_1, p_2, and p_3. The force exerted across face 1 is $p_1ab\sin\theta$, acting downward. The force across face 2 is $p_2ab\cos\theta$, acting to the right. Across face 3 it is p_3ab, with vertical component $p_3ab\sin\theta$ and horizontal component $p_3ab\cos\theta$. The vertical components sum to zero only if $p_1 = p_3$, while the horizontal components sum to zero only if $p_2 = p_3$. Since this result is independent of the value of θ, the pressure must be the same in every direction.

Next, consider how the pressure changes with position. Suppose that p depends on the coordinates $p = p(x, y, z)$ and that the *density* of the fluid is ρ kg m^{-3}. The only external force acting is gravity in the direction of the $-z$ axis. The fluid in the volume $dxdydz$ of Fig. 1.24 is in equilibrium. In the y direction, there is a force to the right across the left-hand face equal to $p(x, y, z)dxdz$ and to the left across the right-hand face equal to $-p(x, y + dy, z)dxdz$. These are the only forces in the y direction, and their magnitudes must be the same. Therefore, p does not change in the y direction. A similar argument shows that p does not change in the x direction. In the z direction there are three terms: the upward force across the bottom face, the downward force across the top face, and the pull of gravity. The weight of the fluid is its mass ($\rho\, dxdydz$) times the gravitational acceleration g ($g = 9.8$ m s^{-2}). The three forces must add to zero:

$$p(x, y, z)\, dxdy - p(x, y, z + dz)\, dxdy - \rho g\, dxdydz = 0.$$

For small changes in height, dz, it is possible to approximate[8] $p(x, y, z + dz)$ by $p(x, y, z) + (dp/dz)\, dz$. With this approximation, the equilibrium equation is

$$dxdydz\left(-\frac{dp}{dz} - \rho g\right) = 0.$$

This equation can be satisfied only if

$$\frac{dp}{dz} = -\rho g. \tag{1.29}$$

This is a *differential equation* for $p(z)$. It is a particularly simple one, since the right-hand side is constant if ρ and g are constant: $dp = -\rho g dz$. Integrating this gives

$$\int dp = -\rho g \int dz,$$

$$p = -\rho g z + c.$$

The constant of integration is determined by knowing the value of p for some value of z. If $p = p_0$ when $z = 0$, then $p_0 = c$ and

$$p = p_0 - \rho g z. \tag{1.30}$$

With a constant gravitational force per unit volume acting on the fluid, the pressure decreases linearly with increasing height. The SI unit of pressure is N m^{-2} or pascal (Pa). The density is expressed in kg m^{-3}, so that ρg has units of N m^{-3} and $\rho g z$ is in N m^{-2}. Pressures are often given as equivalent values of z in some substance, for example, in millimeters of mercury (torr) or centimeters of water. In such cases, the value of z must be converted to an equivalent value of $\rho g z$ before calculations involving anything besides pressure are done. The density of water is 1 g cm^{-3} or 10^3 kg m^{-3}. The density of mercury is 13.6×10^3 kg m^{-3}, so 1 torr = 133 Pa. Another common unit for pressure is the atmosphere (atm), equal to 1.01×10^5 Pa. One atmosphere is approximately the atmospheric pressure at sea level.

[8] See Appendix D on Taylor series for a more complete discussion of this approximation.

1.13 Buoyancy

Buoyancy is important when an object is immersed in a fluid. We are all familiar with buoyant effects when swimming; they are also important in instruments such as the centrifuge. Consider an object of density ρ immersed in a fluid of density ρ_{fluid}. The net force on such an object is the sum of the gravitational force and a force arising from the pressure gradient in the fluid. To visualize this, consider a small object with sides dx, dy, and dz. We have just seen that the pressure on the bottom face is greater than the pressure on the top face. Therefore, there is an upward force on the cube. The total force on the object is then

$$F = \left(-\frac{dp}{dz} - \rho g\right) dx\, dy\, dz.$$

Since the pressure gradient in the fluid is $-\rho_{\text{fluid}} g$, the total force is

$$F = (\rho_{\text{fluid}} - \rho)\, gV, \tag{1.31}$$

where V is the volume of the object. The second term is the object's weight, directed downward. The first term is called the buoyant force and is directed upward. The buoyant force reduces the "effective weight" of the object and depends on the difference of densities of the object and the surrounding fluid.

Animals are made up primarily of water, so their density is approximately 10^3 kg m^{-3}. The buoyant force depends on the animal's environment. Terrestrial animals live in air, which has a density of 1.2 kg m^{-3}. The buoyant force on terrestrial animals is very small compared to their weight. Aquatic animals live in water, and their density is almost the same as the surrounding fluid. The buoyant force almost cancels the weight, so the animal is essentially "weightless." Gravity plays a major role in the life of terrestrial animals, but only a minor role for aquatic animals. Denny (1993) explores the differences between terrestrial and aquatic animals in more detail.

1.14 Compressibility

Increasing the pressure on a fluid causes a deformation and a decrease in volume. The *compressibility* κ is defined as

$$\frac{\Delta V}{V} = -\kappa \Delta p. \tag{1.32}$$

Since $\Delta V/V$ is dimensionless, κ has the units of inverse pressure, N^{-1} m^2 or Pa^{-1}. In many liquids, the compressibility is quite small (e.g., 5×10^{-10} Pa^{-1} for water), and for many purposes, such as flow through pipes, compressibility can be ignored. Other effects, such as the transmission of sound through a fluid, depend on deformation, and compressibility cannot be ignored. The bulk modulus is the reciprocal of the compressibility.

1.15 Diving

Air is easily compressible, so swimming at large depths can be dangerous as the volume of the air in the lungs decreases. One can swim safely for depths of tens of meters (several atmospheres of pressure) using a self-contained underwater breathing apparatus (SCUBA). Compressed air tanks are used to supply air to the lungs, and the pressure of the air is adjusted to match the pressure of the surrounding water.

One physiological effect of breathing high-pressure air is that nitrogen dissolves into the blood, which can lead to a mental impairment known as nitrogen narcosis. Moreover, if the swimmer returns rapidly to the surface after a long deep dive, the lowered pressure allows the dissolved nitrogen to form bubbles in the blood that block blood flow and cause decompression sickness, often called "the bends" (Benedek and Villars 2000). To avoid the bends, swimmers must return to the surface slowly, or replace nitrogen by other gasses, such as helium, that are less soluble in blood.

1.16 Viscosity

A fluid at rest does not support a shear. If the fluid is moving, a shear force can exist. At large velocities the flow of the fluid is turbulent and may be difficult or impossible to calculate. We will consider only those cases in which the velocity is low enough so that the flow is smooth. This means that particles of dye that are introduced into the fluid to monitor its motion flow along smooth lines called *streamlines*. A streamline is tangent to the velocity vector of the fluid at every point along its path. There is no mixing of fluid across streamlines; the flow is *laminar* (in layers). Laminar flow is often used in rooms where dirt or bacterial contamination is to be avoided, such as operating rooms or manufacturing clean rooms. Clean air enters and passes through the room without mixing. Any contaminants picked up are carried out in the air.

A fluid can support a viscous shear stress if the shear strain is changing. One way to create such a situation is to immerse two parallel plates, each of area S, in the fluid, and to move one parallel to the other as in Fig. 1.25. If the fluid in contact with each plate sticks to the plate[9],

[9] This is called the "no-slip" boundary condition. There are exceptions.

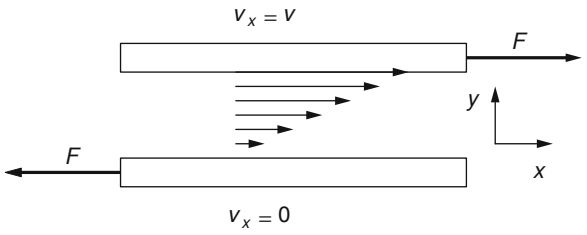


Fig. 1.25 Forces **F** and $-\mathbf{F}$ are needed to make the top plate move in a viscous fluid while the bottom plate remains stationary. The velocity profile is also shown

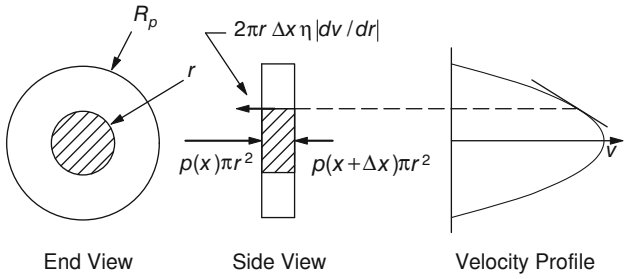


Fig. 1.26 Longitudinal and transverse cross sections of the tube. Newton's first law is applied to the shaded volume

the fluid in contact with the lower plate is at rest and that in contact with the upper plate moves with the same velocity as the plate. Between the plates the fluid flows parallel to the plates, with a speed that depends on position as shown in Fig. 1.25. The variation of velocity between the plates gives rise to a velocity gradient dv_x/dy. Note that this is the rate of change of the shear strain.

In order to keep the top plate moving and the bottom one stationary, it is necessary to exert a force of magnitude F on each plate: to the right on the upper plate and to the left on the lower plate. The resulting shear stress or force per unit area is in many cases proportional to the velocity gradient:

$$\frac{F}{S} = \eta \frac{dv_x}{dy}. \tag{1.33}$$

Often this equation is written with a minus sign, in which case **F** is the force of the fluid on the plate rather than the plate on the fluid. The constant η is called the *coefficient of viscosity*. The units of η are $\mathrm{N\,s\,m^{-2}}$ or $\mathrm{kg\,m^{-1}\ s^{-1}}$ or Pa s. Older units are the $\mathrm{dyn\,cm^{-2}}$ or poise, the centipoise, and the micropoise. 1 poise = 0.1 Pa s. Water has a viscosity of about 10^{-3} Pa s at room temperature. Equation 1.33 gives the force exerted by fluid above the plane at height y on the fluid below the plane. In the case of the parallel plates, the force from above on fluid in the slab between y and $y+dy$ is the same in magnitude as (and opposite in direction to) the force exerted by the fluid below the slab. Therefore, there is no net force on the fluid in the slab, and the fluid moves with constant velocity. Fluids that are described by Eq. 1.33 are called *Newtonian fluids*. Many fluids are not Newtonian.

Since dv_x/dy is the rate of change of the shear strain, Eqs. 1.27 and 1.33 can be written as

$$s_s = \frac{F}{S} = \eta \frac{d\epsilon_s}{dt}.$$

The rate of change of the shear strain is also called the *shear rate*.

1.17 Viscous Flow in a Tube

Biological fluid dynamics is a well-developed area of study (Lighthill 1975; Mazumdar 1992; Vogel 1994). External biological fluid dynamics is concerned with locomotion—from single-celled organisms to swimming fish and flying birds. Internal biological fluid dynamics deals with mass transport within the organism. Two obvious examples are flow in the airways and the flow of blood.

Consider laminar viscous flow of fluid through a pipe of constant radius R_p and length Δx. Ignore for now the gravitational force. The pressure at the left end of a segment of pipe is $p(x)$; at the right end it is $p(x + \Delta x)$. For now consider the special case in which none of the fluid is accelerated, so the total force on any volume element of the fluid is zero. The velocity profile must be as shown in Fig. 1.26: zero at the walls and a maximum at the center. Our problem is to determine $v(r)$.

Let us determine the forces acting on the shaded cylinder of fluid of radius r shown in Fig. 1.26. Since gravity is ignored, there are only three forces acting on the volume. The fluid on the left exerts a force $\pi r^2 p(x)$ acting to the right in the direction of the positive x axis. The fluid on the right exerts a force $-\pi r^2 p(x+\Delta x)$ (the minus sign because it points to the left). The slower moving fluid outside the shaded region exerts a viscous drag force across the cylindrical surface at radius r. The area of the surface is $2\pi r \Delta x$. The force points to the left. Its magnitude is $2\pi r \Delta x\, \eta\, |dv/dr|$. Since dv/dr is negative, we obtain the correct sign by writing it as $2\pi r\, \Delta x\, \eta\, (dv/dr)$. Since the fluid is not accelerating, the forces sum to zero:

$$\pi r^2 [p(x) - p(x + \Delta x)] + 2\pi r\, \Delta x\, \eta\, (dv/dr) = 0, \tag{1.34}$$

which can be rearranged to give

$$\frac{dv}{dr} = \frac{r}{2\eta}\left(\frac{p(x + \Delta x) - p(x)}{\Delta x}\right) = \frac{dp}{dx}\frac{r}{2\eta}. \tag{1.35}$$

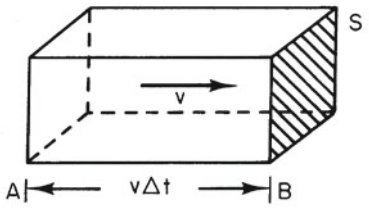


Fig. 1.27 Flow of fluid across the plane at B

This can be integrated:

$$\int dv = \frac{1}{2\eta}\left(\frac{dp}{dx}\right)\int r\,dr,$$
$$v(r) = \frac{1}{4\eta}\left(\frac{dp}{dx}\right)r^2 + A. \tag{1.36}$$

For flow to the right, dp/dx is negative. Therefore it is convenient to write Δp as the pressure drop from x to $x + dx$: $\Delta p = p(x) - p(x + \Delta x)$. Then the first term in Eq. 1.36 is $-(1/4\eta)(\Delta p/\Delta x)r^2$. The constant of integration can be determined assuming the "no-slip" boundary condition: the velocity of the fluid immediately adjacent to a solid is the same as the velocity of the solid itself. Because the wall is at rest, the velocity of the fluid is zero at the wall ($r = R_p$). The final result is

$$v(r) = \frac{1}{4\eta}\frac{\Delta p}{\Delta x}(R_p^2 - r^2). \tag{1.37}$$

The total flow rate or *volume flux* or *volume current* i is the volume of fluid per second moving through a cross section of the tube. Its units are $\mathrm{m^3\,s^{-1}}$. The *volume fluence rate* or *volume flux density*[10] or *current density* j_v is the volume per unit area per unit time across some small area in the tube. The units of j_v are $\mathrm{m^3\,s^{-1}\,m^{-2}}$ or $\mathrm{m\,s^{-1}}$.

In fact, j_v is just the velocity of the fluid at that point. To see this, consider the flow of an incompressible fluid during time Δt. In Fig. 1.27, the fluid moves to the right with velocity v. At $t = 0$, the fluid just to the left of plane B crosses the plane; at $t = \Delta t$, the fluid that was at A at $t = 0$ crosses plane B. All the fluid between plane A and plane B crosses plane B during the time interval Δt. The volume fluence rate is

$$j_v = \frac{(\text{volume transported})}{(\text{area})(\text{time})} = \frac{Sv\Delta t}{S\Delta t} = v. \tag{1.38}$$

It may seem unnecessarily confusing to call the fluence rate or flux density j_v instead of v; however, this notation corresponds to a more general notation in which j means the fluence rate or flux density of anything per unit area per unit time, and the subscript v, s, or q tells us whether it is the fluence rate of volume, solute particles, or electric charge.

To find the volume current i, j_v must be integrated over the cross-sectional area of the pipe. The volume of fluid crossing the washer-shaped area $2\pi r dr$ is $j_v 2\pi r dr = v 2\pi r dr$. The total flux through the tube is

$$i = \int_0^{R_p} j_v(r) 2\pi r\,dr,$$
$$i = \frac{2\pi}{4\eta}\frac{\Delta p}{\Delta x}\int_0^{R_p}\left(R_p^2 - r^2\right) r\,dr. \tag{1.39}$$

To integrate this, let $u = R_p^2 - r^2$. Then $du = -2r dr$ and the integral is $R_p^4/4$. Therefore

$$i = \frac{\pi R_p^4}{8\eta}\frac{\Delta p}{\Delta x} \tag{1.40}$$

is the flux of a viscous fluid through a pipe of radius R_p due to a pressure gradient $(\Delta p/\Delta x)$ along the pipe. The dependence of i on R_p^4 means that small changes in diameter cause large changes in flow.

This relationship was determined experimentally in painstaking detail by a French physician, Jean Leonard Marie Poiseuille, in 1835. He wanted to understand the flow of blood through capillaries. His work and knowledge of blood circulation at that time have been described by Herrick (1942).

As an example of the use of Eq. 1.40, consider a pore of the following size, which might be found in the basement membrane of the glomerulus of the kidney:

$$\begin{gathered} R_p = 5\,\mathrm{nm}, \\ \Delta p = 15.4\,\mathrm{torr}, \\ \eta = 1.4\times 10^{-3}\,\mathrm{kg\,m^{-1}\,s^{-1}}, \\ \Delta x = 50\,\mathrm{nm}. \end{gathered} \tag{1.41}$$

It is first necessary to convert 15.4 torr to Pa using Eq. 1.30 and the value of ρ for mercury, $13.55\times 10^3\,\mathrm{kg\,m^{-3}}$:

$$\Delta p = \rho g \Delta z = (13.55\times 10^3)(9.8)(15.4\times 10^{-3})$$
$$= 2.04\times 10^3\,\mathrm{Pa}.$$

Then Eq. 1.40 can be used:

$$i = \frac{(3.14)(5\times 10^{-9})^4(2.04\times 10^3)}{(8)(1.4\times 10^{-3})(50\times 10^{-9})} = 7.2\times 10^{-21}\,\mathrm{m^3\,s^{-1}}.$$

Now consider the general case in which we have not only viscosity, but the fluid may be accelerated and gravity is important. We continue to write Δp as the pressure drop and consider four contributions, each of which will be discussed:

$$\Delta p = p_1 - p_2 = -\int_{x_1}^{x_2}(dp/dx)\,dx$$
$$= \Delta p_{\mathrm{visc}} + \Delta p_{\mathrm{grav}} + \Delta p_{\mathrm{accel1}} + \Delta p_{\mathrm{accel2}}. \tag{1.42}$$

[10] Some authors call j_v the flux. The nomenclature used here is consistent throughout the book.

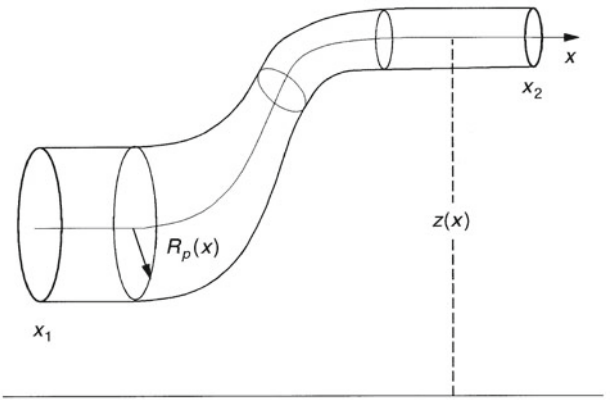


Fig. 1.28 A pipe of circular cross section with radius and height varying along the pipe

For simplicity, we restrict the derivation to an incompressible fluid and a pipe of circular cross section where the radius can change. Distance along the pipe is x, the radius of the pipe is $R_p(x)$, and the cross-sectional area is $S(x) = \pi R_p^2(x)$. Gravitational force acts on the fluid, and the height of the axis of the pipe above some reference plane is $z(x)$, as shown in Fig. 1.28.

Because the fluid is incompressible, the total current i is independent of x. We define the average velocity by

$$i = \overline{v}(x)S(x), \tag{1.43}$$

so

$$\overline{v}(x) = \frac{i}{S(x)} = \frac{i}{\pi R_p^2(x)}. \tag{1.44}$$

If the area decreases, the average velocity increases, and vice versa. This is a special case of the continuity equation, which is discussed further in Chap. 4.

Assume that changes in pipe radius occur slowly enough so that the velocity profile remains parabolic at every point in the pipe and that we can treat x as though it were distance along the axis of the cylinder. We can use Eq. 1.37 to rewrite the velocity profile as

$$v(r,x) = 2\overline{v}\left[1 - \frac{r^2}{R_p^2(x)}\right] = \frac{2i}{\pi R_p^2(x)}\left[1 - \frac{r^2}{R_p^2(x)}\right]. \tag{1.45}$$

The first term in Eq. 1.42 is the drop in pressure because of viscous drag. We can rewrite Eq. 1.35 as

$$\frac{dp_{\text{visc}}}{dx} = \frac{2\eta}{r}\frac{dv}{dr}.$$

Using Eq. 1.45, we can write

$$\frac{dp_{\text{visc}}}{dx} = -\frac{8\eta\, i}{\pi R_p^4(x)}. \tag{1.46}$$

We saw this earlier, solved for i in a pipe of constant radius, as Eq. 1.40. The pressure drop is obtained by integration:

$$\begin{aligned}\Delta p_{\text{visc}} &= -\int_{x_1}^{x_2} dp_{\text{visc}} = -\int_{x_1}^{x_2}\left(\frac{dp_{\text{visc}}}{dx}\right)dx \\ &= +\frac{8\eta i}{\pi}\int_{x_1}^{x_2}\frac{dx}{R_p^4(x)}.\end{aligned} \tag{1.47}$$

To go further requires knowing $R_p(x)$.

The next term p_{grav} is the hydrostatic pressure change that we saw in Eq. 1.30:

$$\Delta p_{\text{grav}} = -\int_{x_1}^{x_2} dp_{\text{grav}} = -\int \frac{dp_{\text{grav}}}{dz}\,dz = \rho g(z_2 - z_1). \tag{1.48}$$

The last two terms of Eq. 1.42 are pressure differences required to accelerate the fluid. When the flow is steady—that is, the velocity depends only on position, and the velocity at a fixed position does not change with time—there can still be an acceleration if the cross section of the pipe changes. The third term, Δp_{accel1}, is the pressure drop required to cause this acceleration. It can be derived as follows: Imagine a streamline in the fluid. No fluid crosses the streamline. Consider a small length of streamline ds and a small area dA perpendicular to it. Note that ds is a small displacement along a streamline, while dx is along the axis of the pipe. The edge of dA defines another set of streamlines that form a tube of flow, and $dAds$ defines a small volume of fluid. Make ds and dA small enough so that v is nearly the same at all points within the volume. The mass of fluid in the volume is $dm = \rho dAds$. We ignore viscosity and gravity, so the only pressure difference is due to acceleration. The net force on the volume is

$$dF = -\frac{dp}{ds}\,ds\,dA. \tag{1.49}$$

This is equal to the mass times the acceleration dv/dt. The acceleration of the fluid in the element is then

$$\frac{dv}{dt} = \frac{dF}{dm} = \frac{-\left(\frac{dp}{ds}\right)dsdA}{\rho dsdA} = -\frac{1}{\rho}\left(\frac{dp}{ds}\right). \tag{1.50}$$

We are considering only velocity changes that occur because the fluid moves along a streamline to a different position. We use the chain rule to write

$$\frac{dv}{dt} = \left(\frac{dv}{ds}\right)\left(\frac{ds}{dt}\right) = v\left(\frac{dv}{ds}\right).$$

Combining these gives

$$\frac{dp_{\text{accel1}}}{ds} = -\rho v\left(\frac{dv}{ds}\right). \tag{1.51}$$

This can be integrated along the streamline to give

$$\Delta p_{\text{accel1}} = -\int_{s_1}^{s_2} \left(\frac{dp_{\text{accel1}}}{ds}\right) ds = +\rho \int_{x_1}^{x_2} v \left(\frac{dv}{ds}\right) ds$$
$$= \frac{\rho v_2^2}{2} - \frac{\rho v_1^2}{2}. \tag{1.52}$$

This is sometimes called the dynamic pressure.

The final term Δp_{accel2} is the pressure drop required to accelerate the fluid between points 1 and 2 if the velocity of the fluid at a fixed position is changing with time (unsteady flow). This happens, for example, to blood that is accelerated as it is ejected from the heart during systole. To derive this term, again imagine a small length of streamline ds and a small area dA perpendicular to it. In addition to ignoring gravity and viscosity, we ignore changes in velocity because of changes in cross section. There is acceleration only if the velocity at a fixed location is changing. The acceleration is $\partial v/\partial t$. The derivative is written with ∂s to signify the fact that we are considering only changes in the velocity with time that occur at a fixed position. The net force required to accelerate this mass is provided by the pressure difference Eq. 1.49:

$$dF = -dA\, dp_{\text{accel2}} = dm \left(\frac{\partial v}{\partial t}\right) = \rho \left(\frac{\partial v}{\partial t}\right) dA\, ds,$$

$$dp_{\text{accel2}} = -\rho \left(\frac{\partial v}{\partial t}\right) ds,$$

$$\Delta p_{\text{accel2}} = -\int_{s_1}^{s_2} dp_{\text{accel2}} = \rho \int_{s_1}^{s_2} \left(\frac{\partial v}{\partial t}\right) ds. \tag{1.53}$$

All of these effects can be summarized in the *generalized Bernoulli equation*:

$$p_1 - p_2 = \Delta p = \underbrace{\rho \int_{s_1}^{s_2} \frac{\partial v}{\partial t}\, ds}_{\Delta p_{\text{accel2}}} + \underbrace{\int_{s_1}^{s_2} \left(-\frac{dp_{\text{visc}}}{ds}\right) ds}_{\Delta p_{\text{visc}}}$$
$$\underbrace{+\frac{\rho v_2^2}{2} - \frac{\rho v_1^2}{2}}_{\Delta p_{\text{accel1}}} + \underbrace{\rho g\,(z_2 - z_1)}_{\Delta p_{\text{grav}}}. \tag{1.54}$$

Equation 1.54 is valid for nonuniform viscous flow that may be laminar or turbulent if the integral is taken along a streamline (see, for example, Synolakis and Badeer 1989).

1.18 Pressure–Volume Work

An important example of work is that done in a biological system when the volume of a container (such as the lungs or the heart or a blood vessel) changes while the fluid within the container is exerting a force on the walls.

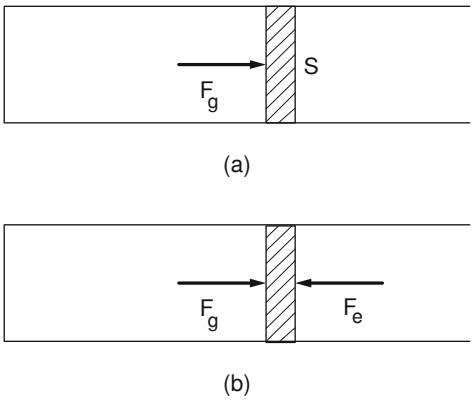


Fig. 1.29 **a** A cylinder containing gas has a piston of area S at one end. b The force exerted on the piston by the gas is balanced by an external force if the piston is at rest

To deduce an expression for pressure–volume work, consider a cylinder of gas fitted with a piston, Fig. 1.29a. If the piston has area S, the gas exerts a force $F_g = pS$ on the piston. If no other force is exerted on the piston to restrain it, it will be accelerated to the right and gain kinetic energy as the gas does work on it:

$$(\text{work done by gas}) = F_g dx = pSdx = pdV. \tag{1.55}$$

If the piston is prevented from accelerating by an external force $\mathbf{F}_e$, equal and opposite to that exerted by the gas (Fig. 1.29b), then the external force does work on the piston:

$$(\text{work done by external force}) = -F_e dx \tag{1.56}$$
$$= -pSdx = -pdV,$$

which is the negative of the work done on the piston by the expanding gas. The result is that the kinetic energy of the piston does not change. The gas does work on the surroundings as it expands, increasing the energy of the surroundings; the surroundings, through the external force, do *negative* work on the gas; that is, they decrease the energy of the gas. (The meaning of "energy of the gas" and "energy of the surroundings" is discussed in Chap. 3.) If the gas is compressed, the situation is reversed: the surroundings do positive work on the gas and the gas does negative work on the surroundings.

For a large change in volume from V_1 to V_2, the pressure may change as the volume changes. In that case the work done by the gas on the surroundings is

$$W_{\text{by gas}} = \int_{V_1}^{V_2} p\, dV. \tag{1.57}$$

This work is the shaded area in Fig. 1.30. If the gas is compressed, the change in volume is negative and the work done *by* the gas is negative.

Let us apply this model to the heart. Suppose that the left ventricle of the heart contracts at constant pressure, so that

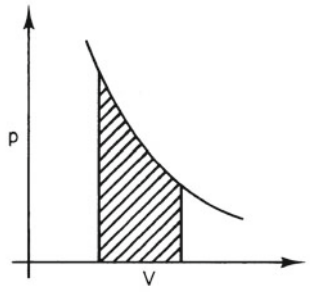


Fig. 1.30 A plot of p versus V, showing the work done by the gas as it expands

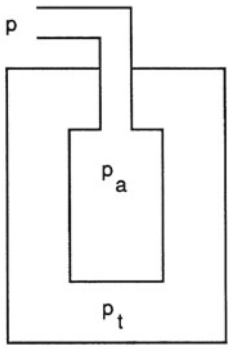


Fig. 1.31 A model of the thorax, lungs, and airways that can be used to understand some features of breathing

it changes volume by $\Delta V = V_2 - V_1$. (Since $V_2 < V_1$, the quantity ΔV is negative. A volume of blood $-\Delta V$ is ejected into the aorta.) The work done by the heart wall on the blood is $-p\Delta V$ and is positive, since ΔV is negative.

As another example of pressure–volume work, we can develop a model to estimate the work necessary to breathe. Consider the model of the lungs and airways shown in Fig. 1.31. The pressure at the nose is the atmospheric pressure p. In the alveoli (air sacs), the pressure is p_a. If there is no flow taking place, $p_a = p$. For air to flow in, p_a must be less than p; for it to flow out, p_a must be greater than atmospheric. The work done by the walls of the alveoli on the gas in them is $-\int p_a\, dV$. The net value of this integral for a respiratory cycle is positive. Perhaps the easiest way to see this is to imagine an inspiration, in which the alveolar pressure is $p_a = p - \Delta p$ and the volume change is ΔV. The work done on the gas is $-(p - \Delta p)\Delta V$. This is followed by an expiration at pressure $p_a = p + \delta p$, for which the work is $-(p + \delta p)(\Delta V)$. The net work done on the gas is $(\Delta p + \delta p)\Delta V$. The energy imparted to the gas shows up as a mixture of heating because of frictional losses and kinetic energy of the exhaled air.

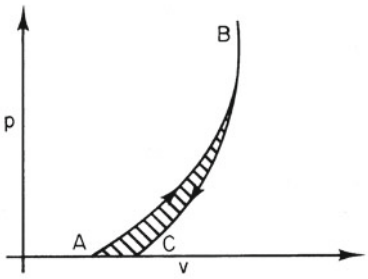


Fig. 1.32 A hypothetical plot of the pressure–volume relationship for inhalation and exhalation

There is another mechanism by which work is done in breathing. Refer again to Fig. 1.31. The pressure in the chest cavity (thorax) is p_t. (The pressure measured in mid-esophagus is a good estimate of p_t.) Because of contractile forces in the lung tissue, $p_a > p_t$. The gas in the alveoli and the fluid in the thorax both do work on the lung tissue. The latter has opposite sign, since a positive displacement dx of a portion of the alveolar wall is in the direction of the force exerted by the alveolar gas but is opposite to the direction of the force exerted by the thoracic fluid. The elastic recoil pressure, multiplied by dV, gives the net work done by both forces on the wall of the lung.

Figure 1.32 shows the elastic recoil pressure versus lung volume. The elastic recoil pressure is the difference between the pressure in the alveoli (air sacs) of the lung and the pressure in the thorax just outside the lung. During inspiration (curve AB), the elastic recoil pressure $p_a - p_t$ is greater than that during expiration (curve BC). The net work done on the lung wall during the respiratory cycle goes into frictional heating of the lung tissue.

1.19 The Human Circulatory System

The human circulatory system is responsible for pumping blood and its life-sustaining nutrients to all parts of the body (Vogel 1992). The circulatory system has two parts: the *systemic circulation* and the *pulmonary circulation*, as shown in Fig. 1.33. The left heart pumps blood into the systemic circulation: organs, muscles, etc. The right heart pumps blood through the lungs. As the heart beats, the pressure in the blood leaving the heart rises and falls. The maximum pressure during the cardiac cycle is the *systolic pressure*. The minimum is the *diastolic pressure*. (A blood pressure reading is in the form systolic/diastolic, measured in torr. A typical blood pressure might be 110/70.)

A *sphygmomanometer* is used to measure blood pressure. Air is pumped into a cuff placed around the forearm. The applied pressure is measured using either a column of mercury or a mechanical pressure transducer. The cuff is inflated

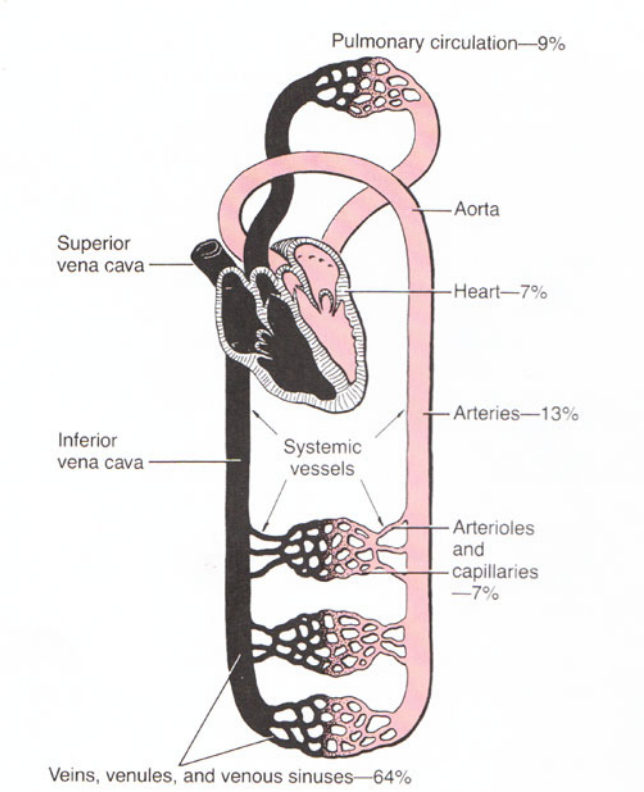


Fig. 1.33 The human circulatory system. The subject is facing you, so the left chambers of the heart are on the right in the picture. The left heart pumps oxygenated blood (red), and the right heart pumps deoxygenated blood (black). (Reprinted from Guyton 1991 © Elsevier Inc. Used with permission of Elsevier)

until flow in the brachial artery ceases. The cuff pressure is then slowly reduced until flow returns during *systole*. The flow can be detected by listening with a stethoscope for the sounds associated with the starting and stopping of flow (*Korotkoff sounds*), or with a *pulse oximeter* (see page 392). The cuff pressure is then further reduced until flow occurs continuously throughout the cardiac cycle including *diastole*.

The blood flows from the aorta to several large arteries, to medium-sized arteries, to small arteries, to arterioles, and finally to the capillaries, where exchange with the tissues of oxygen, carbon dioxide, and nutrients takes place. The blood emerging from the capillaries is collected by venules, flows into increasingly larger veins, and finally returns to the heart through the vena cava.

At any given time, blood is flowing in only a fraction of the capillaries. The state of flow in the capillaries is continually changing to provide the amount of oxygen required by each organ. In skeletal muscle, terminal arterioles constrict and dilate to control distribution of blood to groups of capillaries. In smooth muscle and skin, a precapillary sphincter muscle controls the flow to each capillary (Patton et al. 1989, p. 860). Since the blood is incompressible and is conserved,[11] the total volume flow i remains the same at all generations of branching in the vascular tree. Table 1.4 shows average values for the pressure and vessel sizes at different generations of branching. Most of the pressure drop occurs in the arterioles.

We define the *vascular resistance* R in a pipe or a segment of the circulatory system as the ratio of pressure difference across the pipe or segment to the flow through it:

$$R = \frac{\Delta p}{i}. \tag{1.58}$$

The units are $\mathrm{Pa\,m^{-3}\,s}$. Physiologists use the *peripheral resistance unit (PRU)*, which is $\mathrm{torr\,ml^{-1}\,min}$. For Poiseuille flow, the resistance can be calculated from Eq. 1.40:

$$R = \frac{8\eta\Delta x}{\pi R_p^4}. \tag{1.59}$$

The resistance decreases rapidly as the radius of the vessel increases.

If vessels of different diameters are connected in series so that the flow i is the same through each one and the total pressure drop is the sum of the drops across each vessel, then the total resistance is the sum of the resistances of each vessel:

$$R_{\mathrm{tot}} = R_1 + R_2 + R_3 + \cdots. \tag{1.60}$$

If there is branching so that several vessels are in parallel with the same pressure drop across each one, the total flow through all the branches equals the flow in the vessel feeding them. The total resistance is then given by

$$\frac{1}{R_{\mathrm{tot}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots. \tag{1.61}$$

For the most part, the capillaries are arranged in parallel. Even though the resistance of an individual capillary is large because of its small radius (Eq. 1.59), the resistance of the capillaries as a whole is relatively small because there are so many of them (see Problem 42).

The pressure in the left ventricle changes during the cardiac cycle. It can be plotted versus time. It can also be plotted versus ventricular volume, as in Fig. 1.34. The p–V relationship moves counterclockwise around the curve during the cycle. Filling occurs at nearly zero pressure until the ventricle begins to distend when the volume exceeds 60 ml[12]. There is then a period of contraction at nearly constant volume

[11] This is not strictly true. Some fluid leaves the capillaries and returns to the heart through the lymphatic system instead of the venous system. See Chap. 5.

[12] $1\,\mathrm{ml} = 10^{-3}\,\mathrm{liter(l)} = 10^{-6}\,\mathrm{m^3}$.

Table 1.4 Typical values for the average pressure at the entrance to each generation of the major branches of the cardiovascular tree, the average blood volume in certain branches, and typical dimensions of the vessels

Location	Average pressure (torr)	Blood volume[a] (ml)	Diameter[b] (mm)	Length[b] (mm)	Wall thickness[b] (mm)	Avg. velocity[b] ($m\,s^{-1}$)	Reynolds number at maximum flow[c]
Systemic circulation							
Left atrium	5						
Left ventricle	100						
Aorta	100	156	20	500	2	0.48	9 400
Arteries	95	608	4	500	1	0.45	1 300
Arterioles	86	94	0.05	10	0.2	0.05	
Capillaries	30	260	0.008	1	0.001	0.001	
Venules	10	470	0.02	2	0.002	0.002	
Veins	4	2682	5	25	0.5	0.01	
Vena cava	3	125	30	500	1.5	0.38	3 000
Right atrium	3						
Pulmonary circulation							
Right atrium	3						
Right ventricle	25						
Pulmonary artery	25	52					
Arteries	20	91					7 800
Arterioles	15	6					
Capillaries	10	104					
Veins	5	215					2 200
Left atrium	5						

[a]From Plonsey (1995)
[b]From Mazumdar (1992)
[c]From Milnor (1989)

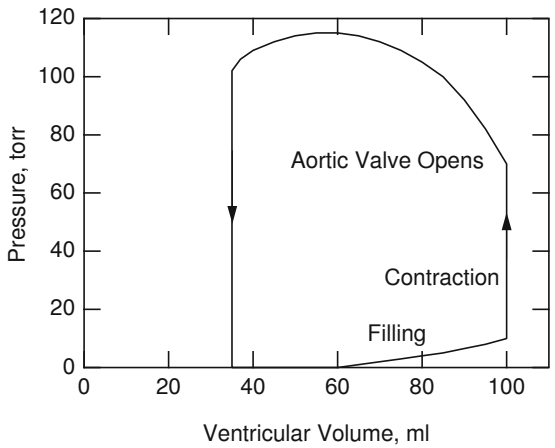


Fig. 1.34 Pressure–volume relationship in the left ventricle. The curve is traversed counterclockwise with increasing time. The stroke volume is 100−35 = 65 ml. Systolic pressure is 118 torr, and diastolic pressure is 70 torr. The ventricular pressure drops below diastolic while the pressure in the arteries remains about 70 torr because the aortic valve is closed and prevents back flow

that causes the ventricular pressure to rise until it exceeds the (diastolic) pressure in the aorta, and the aortic valve opens. The contraction continues, and the pressure rises further, but the ventricular volume decreases as blood flows into the aorta. The ventricle then relaxes. The aortic valve closes when the ventricular pressure drops below that in the aorta. The work done in one cycle is the area enclosed by the curve. For the curve shown, it is 6600 torr ml = 0.88 J. At 80 beats per minute, the power is 1.2 W. In this drawing the stroke volume is 100 − 35 = 65 ml, and the cardiac output is

$$i = (65\,\mathrm{ml\,beat^{-1}})(80\,\mathrm{beats/60\,s}) = 87 \times 10^{-6}\mathrm{m^3\,s^{-1}}.$$

1.20 Turbulent Flow and the Reynolds Number

Many features of the circulation can be modeled by Poiseuille flow. However, at least four effects—in addition to those in Eq. 1.42—cause departures from Poiseuille flow: (1) there may be *turbulence*; (2) there are departures from a parabolic velocity profile; (3) the vessel walls are elastic; and (4) the apparent viscosity depends on both the fraction of the blood volume occupied by red cells and the size of the vessel.

The importance of turbulence (nonlaminar flow) is determined by a dimensionless number characteristic of the system called the *Reynolds number* N_R. It is defined by

$$N_R = \frac{LV\rho}{\eta}, \tag{1.62}$$

where L is a length characteristic of the problem, V a velocity characteristic of the problem, ρ the density, and η the viscosity of the fluid. When N_R is greater than a few thousand, turbulence usually occurs.

The Reynolds number arises in the following way: If we were to write Newton's second law for a fluid (which we have not done) in terms of dimensionless primed variables such as $\mathbf{r}' = \mathbf{r}/L$, $\mathbf{v}' = \mathbf{v}/V$, and $t' = t/(L/V)$, we would find that the equations depended on the properties of the fluid only through the combination N_R (Mazumdar 1992, p. 14). With appropriate scaling of dimensions and times, flows with the same Reynolds number are identical.

There is ambiguity in defining the characteristic length and the characteristic velocity. Should one use the radius or the diameter of a tube? The maximum velocity or the average velocity? If one is solving the equations of motion, one knows what values of L and V were used to transform the equations. They are used to transform the solution back to "real world" coordinates. However, if one is making a statement such as "turbulence usually occurs for values of N_R greater than a few thousand," there is ambiguity. On the other hand, the statement is not very precise. Sometimes an additional subscript is used to specify how N_R was determined.

When N_R is large, inertial effects are important. External forces accelerate the fluid. This happens when the density is large and the viscosity is small. As the viscosity increases (for fixed L, V, and ρ), the Reynolds number decreases. When the Reynolds number is small, viscous effects are important. The fluid is not accelerated, and external forces that cause the flow are balanced by viscous forces. Since viscosity is a form of internal friction in the fluid, work done on the system by the external forces is transformed into thermal energy. The low-Reynolds-number regime is so different from our everyday experience that the effects often seem counterintuitive. They are nicely described by Purcell (1977).

Here is an example of an estimate expressed in terms of the Reynolds number. A pressure difference Δp acts on a segment of fluid of length Δx undergoing Poiseuille flow. The difference between the force exerted on the segment of fluid by the fluid "upstream" and that exerted by the fluid "downstream" is $\pi R_p^2 \Delta p$. If the average speed of the fluid is $\overline{v}$, then the net work done on the segment by the fluid upstream and downstream in time Δt is $W_{\text{visc}} = \pi R_p^2 \Delta p \overline{v} \Delta t$. Since the fluid is not accelerated, this work is converted into thermal energy. We can solve Eq. 1.40 for Δp and use Eq. 1.44 to write

$$W_{\text{visc}} = \pi R_p^2 \, \Delta p \, \overline{v} \Delta t = 8\eta\pi\overline{v}^2 \, \Delta x \, \Delta t.$$

The kinetic energy of the moving fluid in a cylinder of length $\overline{v}\Delta t$ is

$$E_k = \frac{mv^2}{2} = \frac{\rho \, \pi R_p^2 \, (\overline{v}\Delta t)\overline{v}^2}{2} = \frac{\rho \, \pi R_p^2 \, \overline{v}^3 \, \Delta t}{2},$$

and the ratio of the kinetic energy to the work done is

$$\frac{E_k}{W_{\text{visc}}} = \frac{\rho \, \overline{v} \, R_p^2}{16\eta \, \Delta x} = \frac{1}{16\xi} \frac{\rho \, \overline{v} \, R_p}{\eta} = \frac{1}{16\xi} N_R$$

where we write Δx as ξR_p. This result shows that the ratio of kinetic energy to viscous work is proportional to the Reynolds number. Another example is given in the problems.

The behavior of a sphere moving through a fluid illustrates how flow behavior depends on Reynolds number. At low Reynolds number, the viscous forces tend to make the fluid stick to the sphere, creating a large amount of viscous drag. This flow can be analyzed analytically (Schlichting and Gersten 2000). The drag force is $6\eta R v$, where R is the sphere radius, v is the speed of the sphere, and η is the viscosity, a result known as *Stokes' law*. At high Reynolds number, Bernoulli's equation (see Problem 36) tells us that high pressure is associated with low fluid speeds, and low pressure is associated with fast speeds. There is a region of high pressure in front of and in back of the sphere (where speeds are slow), and low pressure to either the left or right side (where speeds are fast). At very high Reynolds number, viscosity is small but still plays a role because of the no-slip boundary condition at the sphere surface. A thin layer of fluid, called the *boundary layer*, sticks to the solid surface, causing a large velocity gradient and therefore significant viscous drag (Schlichting and Gersten 2000). At extremely high Reynolds number, the flow undergoes *separation*, where eddies and turbulent flow occur downstream from the sphere, lowering the pressure in the sphere's wake, but they do not influence the high pressure in front of the sphere. Thus, pressure drag contributes to the total drag force, in addition to viscous drag. Similarly, if we consider a nonsymmetrical object instead of a sphere, we can make the flow speed and pressure differ on the left and right sides of the object, resulting in lift: a force perpendicular to the direction of the main fluid flow. Vogel (1994) discusses the biological implications of high Reynolds number flow, which is particularly important for flying animals and large swimmers. However, many of the biological fluid dynamics applications we will consider occur at low Reynolds number, where turbulence, separation, pressure drag, and boundary layers are not important, and Stokes' law dominates.

A large range of values of N_R occurs in the circulatory system. Typical values corresponding to the peak flow are given in Table 1.4. Blood flow is laminar except in the ascending aorta and main pulmonary artery, where turbulence may occur during peak flow. The Reynolds number in the capillaries is about 10^{-2}.

There are two main causes of departures from the parabolic velocity profile. First, a red cell is about the same diameter as a capillary. Red cells in capillaries line up single file, each nearly blocking the capillary. The plasma flows in

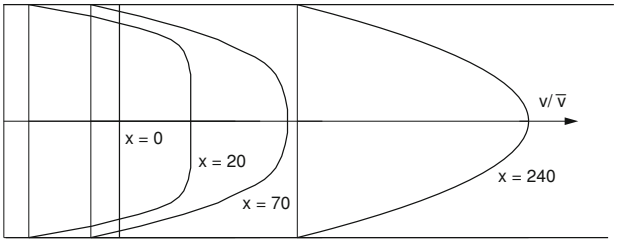


Fig. 1.35 Velocity profiles in steady laminar flow at the entrance to a tube, showing the development of the parabolic velocity profile. The velocity is given as $v/\overline{v}$. At the entrance, $v/\overline{v} = 1$. When the Poiseuille flow is fully developed, $v/\overline{v}$ is 2 at the center of the tube. These curves are calculated from a graph by Cebeci and Bradshaw (1977) for laminar flow in a tube of radius 2 mm and a pressure gradient of 20 torr m^{-1}, carrying a fluid with a viscosity of 3×10^{-3} N s m^{-2} and a density of 10^3 kg m^{-3}. The scales are different along the axis and radius of the tube; the tube radius is 2 mm and the entrance region is 240-mm long

small volumes between each red cell, with a velocity profile that is nearly independent of radius. Second, the *entry region* causes deviations from Poiseuille flow in larger vessels. Suppose that blood flowing with a nearly flat velocity profile enters a vessel, as might happen when blood flowing in a large vessel enters the vessel of interest, which has a smaller radius. At the wall of the smaller vessel, the flow is zero. Since the blood is incompressible, the *average* velocity is the same at all values of x, the distance along the vessel. (We assume the vessel has constant cross-sectional area.) However, the velocity profile $v(r)$ changes with distance x along the vessel. At the entrance to the vessel ($x = 0$), there is a very abrupt velocity change near the walls. As x increases, a parabolic velocity profile is attained. The transition or entry region is shown in Fig. 1.35. In the entry region, the pressure gradient is different from the value for Poiseuille flow. The velocity profile cannot be calculated analytically in the entry region. Various numerical calculations have been made, and the results can be expressed in terms of scaled variables (see, for example, Cebeci and Bradshaw 1977). The Reynolds number used in these calculations was based on the diameter of the pipe, $D = 2R_p$, and the average velocity. The length of the entry region is

$$L = 0.05DN_{R,D} = 0.1R_pN_{R,D} = 0.2R_pN_{R,R_p}. \quad (1.63)$$

Blood pressure is, of course, pulsatile. This means that the average velocity and $v(r)$ are changing with time and also departing from the parabolic profile. Also, at the peak pressure during systole, the aorta and arteries expand, storing some of the blood and releasing it gradually during the rest of the cardiac cycle. Pulsatile flow and the elasticity of vessel walls are discussed extensively by Caro et al. (1978) and Milnor (1989).

Blood is not a Newtonian fluid. The viscosity depends strongly on the fraction of volume occupied by red cells (the *hematocrit*). In blood vessels of less than 100-μm radius, the apparent viscosity decreases with tube radius. Since a red cell barely fits in a capillary, the velocity profile in capillaries is not parabolic. Flow in arterioles and arteries is often modeled as individual particles surrounded by plasma and transported by laminar flow, each red cell staying at its own distance from the central axis. However, high-speed motion pictures show that the red cells often collide with other red cells and with the wall. (See the articles by Trowbridge (1982, 1983) and Trowbridge and Meadowcroft (1983), and also the Caro et al. and Milnor articles.)

Symbols Used in Chapter 1

Symbol	Use	Units	First used page
a, a	Acceleration	m s^{-2}	3
a, b	Small distances	m	14
c	Constant of integration		14
g	Acceleration due to gravity	m s^{-2}	14
h	Small distance	m	13
i	Total volume flux or flow rate or current	m^3 s^{-1}	17
j_v	Volume fluence rate or flux density (flow of volume per unit area per second)	m s^{-1}	17
l	Length of rod	m	12
m	Mass	kg	3
p	Pressure	Pa	14
p_t	Pressure in thorax	Pa	20
p_a	Pressure in alveoli	Pa	20
r	Position	m	5
r	Distance from origin (radius) in polar coordinates	m	5
s	Displacement	m	12
s_n	Normal stress	Pa	12
s_s	Shear stress	Pa	13
s	Distance along a streamline	m	18
t	Time	s	11
v, **v**	Velocity	m s^{-1}	11
x, y, z	Coordinates	m	4
$\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$	Unit vectors along the x, y, and z axes		6
A	Constant of integration		17
dA	Small area perpendicular to a streamline	m^2	19
D	Pipe diameter	m	24
E	Young's modulus	Pa	13
E_k	Kinetic energy	J	11
F, **F**	Force	N	3
G	Shear modulus	Pa	13
L	Characteristic length	m	24
N, **N**	Force	N	8
N_R	Reynolds number		22
$N_{R,D}$	Reynolds number based on diameter		24
N_{R,R_p}	Reynolds number based on pipe radius		24
P	Power	W	12
R, **R**	Force	N	8
R_p	Radius of pipe	m	17
R	Vascular resistance	Pa m^{-3} s	21
S	Cross-sectional area	m^2	12
V	Volume	m^3	15

V	Velocity	$\mathrm{m\,s^{-1}}$	23
W, $\mathbf{W}$	Weight	N	4
W	Work	J	19
δ	A small distance	m	13
ϵ_n	Normal strain		12
ϵ_s	Shear strain		13
η	Viscosity	Pa s	16
$\alpha, \beta, \theta, \phi$	Angle		5
κ	Compressibility	$\mathrm{Pa^{-1}}$	15
ρ	Mass density	$\mathrm{kg\,m^{-3}}$	14
τ,$\boldsymbol{\tau}$	Torque	N m	5
ξ	Dimensionless ratio		23

Problems

Section 1.1

Problem 1. Estimate the number of hemoglobin molecules in a red blood cell. Red blood cells are little more than bags of hemoglobin, so it is reasonable to assume that the hemoglobin takes up all the volume of the cell.

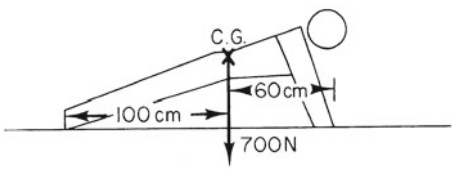
Problem 2. Our genetic information or genome is stored in the parts of the DNA molecule called base pairs. Our genome contains about 3 billion $\left(3 \times 10^9\right)$ base pairs, and there are two copies in each cell. Along the DNA molecule, there is one base pair every one-third of a nanometer. How long would the DNA helix from one cell be if it were stretched out in a line? If the entire DNA molecule were wrapped up into a sphere, what would be the diameter of that sphere?

Problem 3. Estimate the size of a box containing one air molecule. (Hint: What is the volume of one mole of gas at standard temperature and pressure?) Compare the size of the box to the size of an air molecule (about 0.1 nm).

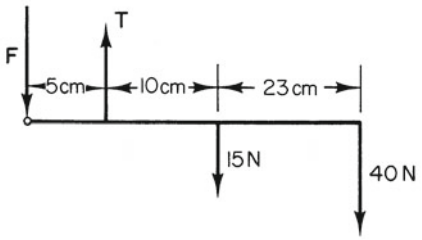
Problem 4. Estimate the density of water (H_2O) in $\mathrm{kg\,m^{-3}}$. Useful information: an oxygen atom contains eight protons and eight neutrons. A hydrogen atom contains one proton and no neutrons. The mass of the electron is negligible.

Section 1.4

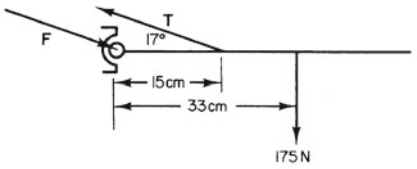
Problem 5. A person with mass $m = 70$ kg has a weight (mg) of about 700 N. If the person is doing push-ups as shown, what are the vertical components of the forces exerted by the floor on the hands and feet?



Problem 6. A person with upper arm vertical and forearm horizontal holds a mass of 4 kg. The mass of the forearm is 1.5 kg. Consider four forces acting on the forearm: **F** by the bones and ligaments of the upper arm at the elbow, **T** by the biceps, 40 N by the mass, and 15 N as the weight of the arm. The points of application are shown in the drawing. Calculate the vertical components of **F** and **T**.

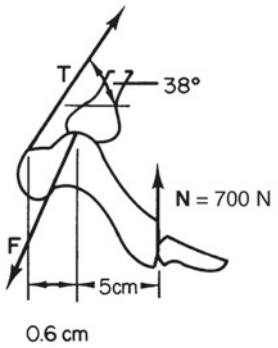


Problem 7. When the arm is stretched out horizontally, it is held by the deltoid muscle. The situation is shown schematically. Determine **T** and **F**.



Section 1.6

Problem 8. When a person crouches, the geometry of the heel is as shown. Determine **T** and **F**. Assume all the forces act in the plane of the drawing.

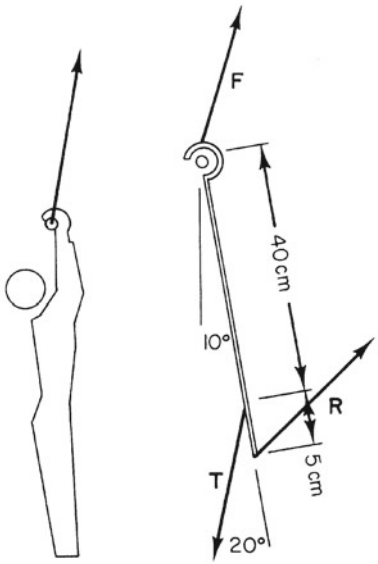


Problem 9. A person of weight W is suspended by both hands from a high bar as shown. The center of mass is directly below the bar.

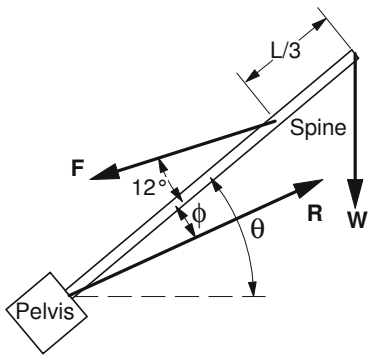
(a) Find the horizontal and vertical components F_x and F_y, where **F** is the force exerted by the bar on each of the two hands.

(b) Given the additional information about the arm shown in the second drawing, calculate the components of **R**, the force exerted by the humerus on the forearm through the elbow, and the tension **T** in the biceps tendon. Neglect

the weight of the arm, and assume that **T** and **R** are the only forces exerted on the forearm by the upper arm.



Problem 10. Consider the forces on the spine when lifting. Approximate the spinal column as a stiff bar of length L that has three forces acting on it. **W** is the downward force acting at the top of the spinal column (via the arms and shoulders), and equals the weight of the object being lifted. **F** is the force applied by the erector spinae muscle, which attaches to the spine about 1/3 of the way from the top of the column. Assume this muscle acts at an angle of 12 ° to the spinal column. **R** is the force the pelvis exerts on the spinal column. The weight of the trunk is neglected. Assume the spinal column makes an angle θ with the horizontal.



(a) Determine **R** and **F** in terms of **W** and θ.
(b) The spinal column may be injured if **R** is too large. Compare **R** when θ is 0 and 90 °. This problem explains why people say, "lift with your legs, not with your back."
(c) Compare the angle ϕ when θ is 0 and 90 °. If ϕ is not close to zero, there will be considerable transverse force at the discs in the lower back, which is not a good situation.

Section 1.8

Problem 11. Suppose that instead of using a cane, a person holds a suitcase of weight $W/4$ in one hand, 0.4 m from the midline. The person is standing on the opposite leg. Calculate the force exerted by the hip abductor muscles and by the acetabulum on that leg.

Section 1.10

Problem 12. Young's modulus for a spider's thread is about 0.2×10^{10} Pa, and the thread breaks when it undergoes a strain of about 50 % (Köhler and Vollrath 1995).
(a) Calculate the tensile strength of the thread and compare it to the tensile strength of steel.
(b) Calculate the strain that steel undergoes when it breaks. (Assume that a linear relationship between stress and strain holds until it breaks.) Compare the breaking strain to the spider's thread.

Problem 13. Assume an object undergoes a normal strain in all three directions: $\epsilon_x = \Delta x/l_x$, $\epsilon_y = \Delta y/l_y$, and $\epsilon_z = \Delta z/l_z$. Relate the three strains to the change in volume of the object. Assume the strains are small.

Section 1.11

Problem 14. Relate the shear strain to angle θ in Fig. 1.22. How does this relationship simplify if θ is small?

Section 1.12

Problem 15. The inspirational pressure difference p_{in} that the lung can generate is about 86 torr. What would be the absolute maximum depth at which a person could breathe through a snorkel device? (A safe depth is only about half this maximum, since the lung ventilation becomes very small at the maximum depth. Assume the lungs are 30 cm below the mouth.)

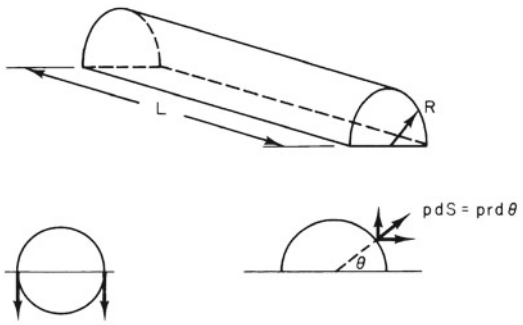
Problem 16. A person standing erect can in some cases be modeled by a column of water.
(a) Estimate the hydrostatic pressure difference between a person's head and foot in torr.
(b) Explain why blood pressure is measured in the arm at the same vertical height as the heart.
(c) Our body has adapted to having a larger hydrostatic pressure in our feet than in our head. Speculate on why you feel uncomfortable when you "stand on your head."

Problem 17. A medication dissolved in a saline solution is infused into a vein in the patient's arm (IV infusion). The

density of saline is the same as water. The pressure of the blood inside the vein is 5 torr above atmospheric pressure. How high above the insertion point must the container be hung so that there is sufficient hydrostatic pressure to force fluid into the vein?

Problem 18. The walls of a cylindrical pipe that has an excess pressure p inside are subject to a tension force per unit length T. (Consider only the force per unit length in the walls of the cylinder, not the force in any end caps of the pipe.) The force per unit length in the walls can be calculated by considering a different pipe made up of two parts as shown in the figure: a semicircular half-cylinder of radius R and length L attached to a flat plate of width $2R$ and length L. What is the force that the excess pressure exerts on the flat plate? Show that the tension force per unit length in the wall of the tube is $T = pR$. This is called the *Law of Laplace*. (Do not worry about any deformation.)

See if you can obtain the same answer by direct integration of the horizontal and vertical components of the force due to the excess pressure.



Problem 19. Find a relationship among the tension per unit length T across any element of the wall of a spherical soap bubble, the excess pressure inside the bubble, p, and the radius of the bubble, R. (Hint: Use the same technique as for the previous problem.)

Problem 20. The Law of Laplace, $T = pR$, relates the tension in an arterial wall, T, to the pressure p inside the artery and its radius R. Assume the wall obeys *Hooke's law*, $T = k(R - R_0)$, where R_0 is the radius of the artery when $p = 0$ and k is a measure of the wall stiffness.

(a) Derive an expression for R as a function of p. Sketch plots of r versus p and T versus p.

(b) Determine the critical pressure at which R goes to infinity. Physically, this is the pressure that guarantees a burst artery (an *aneurysm*).

(c) Arteries would be unstable if they were to balloon out and burst as the pressure approaches a critical value. They avoid this problem by becoming more stiff as the radius increases. Repeat part (a) using $k = cR$ for the stiffness. In this case is there a critical pressure at which the artery will burst?

The law of Laplace has many applications in biology and medicine (Basford 2002). For a discussion of how arteries become stiffer as R grows, see Vogel (1992).

Section 1.13

Problem 21. Suppose a fish has an average density of 1030 $\mathrm{kg\,m^{-3}}$, compared to the density of the surrounding water, 1000 $\mathrm{kg\,m^{-3}}$. One way the fish can keep from slowly sinking is by using an *air bladder* (the density of air is 1.2 $\mathrm{kg\,m^{-3}}$). What fraction of the fish's total volume must be air in order for the fish to be neutrally buoyant (the buoyant force is equal and opposite to the weight). Assume that the volume V of the fish's tissue is fixed, so in order to increase the volume U of the air bladder, the total volume of the fish $V + U$ must increase.

Problem 22. This problem explores the physics of a *centrifuge*. A cylinder of fluid of density ρ_{fluid} and length L is rotated at an angular velocity ω ($\mathrm{rad\,s^{-1}}$) in a horizontal plane about a vertical axis through one end of the tube. Neglect gravity. An object moving in a circle with constant angular velocity has a *centripetal acceleration* $a = -r\omega^2$ toward the center of the circle. Find the pressure in the fluid as a function of distance from the axis of rotation, assuming the pressure is p_0 at $r = 0$.

Problem 23. Buoyancy plays an important role in the centrifuge. Consider a small cubic particle of density ρ immersed in a fluid of density ρ_{fluid}.

(a) Write Newton's second law for the particle, considering only the centripetal acceleration and the pressure exerted by the fluid (Problem 22). Find an expression for the "effective weight" of the particle (analogous to Eq. 1.31) in terms of ρ, ρ_{fluid}, ω, r, and the particle volume V. Your result is more general than you might expect: it is true for a particle of any shape (Wick and Tooby 1977).

(b) Find the ratio of the "effective weight" derived in (a) to the "effective weight" due to gravity (Eq. 1.31).

(c) If the particle is 10 cm from the axis of a centrifuge spinning at 40,000 revolutions per minute, evaluate the ratio obtained in (b).

(d) The *density gradient* technique uses a sucrose solution of varying concentration to produce a fluid density that varies with r, $\rho_{\text{fluid}}(r)$. Explain how in this case the centrifuge can be used to separate particles of different densities.

Problem 24. For the centrifuge of Problem 23, assume there is one additional force: a viscous force proportional to the speed u of the particle relative to the fluid.

(a) Derive an expression for u, the *sedimentation velocity*, assuming the particle is not accelerating relative to the fluid.

(b) The sedimentation velocity per unit acceleration, S, is a parameter commonly used in centrifuge work. Divide the expression obtained in (a) by the centripetal acceleration to obtain an expression for S. The common unit for S is the svedberg ($1\,\text{Sv} = 10^{-13}\,\text{s}$).

(c) Consider two particles with $S = 50$ and 70 Sv. For the centrifuge of Problem 23(c), how long will it take for the particles to separate by 3 mm if they were initially at the same position? How long would this separation take if gravity were used instead of a centrifuge?

Section 1.14

Problem 25. What is the compressibility of a gas for which $pV = \text{const.}$? Compare the compressibility of water to that of air at atmospheric pressure. What are the implications of this for the volume of the lungs of a swimmer diving deep below the water surface?

Problem 26. Figure 1.20, showing a rod subject to a force along its length, is a simplification. Actually, the cross-sectional area of the rod shrinks as the rod lengthens. Let the axial strain and stress be along the z axis. They are related by Eq. 1.25, $s_z = E\epsilon_z$. The lateral strains ϵ_x and ϵ_y are related to s_z by $s_z = -(E/\nu)\epsilon_x = -(E/\nu)\epsilon_y$, where ν is called the *Poisson's ratio* of the material.

(a) Use the result of Problem 13 to relate E and ν to the fractional change in volume $\Delta V/V$.

(b) The change in volume caused by hydrostatic pressure is the sum of the volume changes caused by axial stresses in all three directions. Relate Poisson's ratio to the compressibility.

(c) What value of ν corresponds to an incompressible material?

(d) For an isotropic material, $-1 < \nu < 0.5$. How would a material with negative ν behave?

Elliott et al. (2002) measured Poisson's ratio for articular (joint) cartilage under tension and found $1 < \nu < 2$. This large value is possible because cartilage is *anisotropic*: its properties depend on direction.

Section 1.16

Problem 27. Consider the fluid flowing between two slabs as shown in Fig. 1.25. Since the work done by the external force on the system in time dt is $dW = Fvdt$, the rate of doing work is $P = dW/dt = Fv$, where v is the speed of the moving plate. Find the power dissipated per unit volume of the fluid in terms of the velocity gradient.

Problem 28. Consider a fluid that is flowing in the x direction, but with the velocity v_x changing in the y direction.

(a) Start with Newton's second law. Analyze the forces on a small cube of fluid and derive the equation

$$\rho\frac{\partial v_x}{\partial t} + \rho v_x\frac{\partial v_x}{\partial x} = -\frac{\partial p}{\partial x} + \eta\frac{\partial^2 v_x}{\partial y^2}.$$

This is a simplified version of the *Navier–Stokes equation* that governs fluid flow.

(b) Which term in the equation is nonlinear (that is, if p and v_x are doubled, which term does not double)? A nonlinear equation is needed to describe complicated flows such as turbulence.

Problem 29. Consider the simplified version of the Navier–Stokes equation in Problem 28. Assume the fluid speed is approximately V and all spatial changes occur over distances of order L. Take the ratio of the "inertial term" $\rho v_x(\partial v_x/\partial x)$ to the "viscous term" $\eta(\partial^2 v_x/\partial y^2)$ and show that you get the Reynolds number, Eq. 1.62.

Section 1.17

Problem 30. Consider laminar flow in a pipe of length Δx and radius R_p. Find the total viscous drag exerted by the pipe on the fluid.

Problem 31. The maximum flow rate from the heart is $500\,\text{ml}\,\text{s}^{-1}$. If the aorta has a diameter of 2.5 cm and the flow is Poiseuille, what are the average velocity, the maximum velocity at the center of the vessel, and the pressure gradient along the vessel? Plot the velocity versus distance from the center of the vessel. As an approximation to the viscosity of blood, use $\eta = 10^{-3}\,\text{kg}\,\text{m}^{-1}\,\text{s}^{-1}$.

Problem 32. The glomerular pore described in Eq. 1.41 has a flow $i = 7.2\times10^{-21}\,\text{m}^3\,\text{s}^{-1}$. How many molecules of water per second flow through it? What is their average speed?

Problem 33. Organisms may use shear stress to determine the appropriate size of vessels for fluid transport (LaBarbera 1990). Consider a parent vessel of radius R_p that branches into two daughter vessels of radii R_{d1} and R_{d2}.

(a) Find a relationship between the radii R_p, R_{d1}, and R_{d2} such that the shear stress on the vessel wall is the same in each vessel. (Hint: Use conservation of the volume flow.) This relationship is called *Murray's Law*.

(b) If a 100-μm parent vessel branches into two identical daughter vessels, what is the radius of each daughter vessel? What is the cross-sectional area of the parent vessel, and what is the sum of the cross-sectional areas of the daughter vessels?

(c) If the two daughter vessels branch into subsequent generations of even smaller vessels, all obeying Murray's law, and the daughter vessels of any generation are all the same size, then find a relationship between the number of vessels in the nth generation, the radius of the single parent vessel, and the radii of the nth generation's daughter vessels.

(d) We have one aorta of radius 10 mm. Use Murray's law to estimate how many capillaries we have, each of radius 5 μm. Calculate the cross-sectional area of the aorta and the sum of the cross-sectional area of all our capillaries. Warning: Murray's law is a good approximation, but may not be exact for our smallest vessels.

Problem 34. Sap flows up a tree at a speed of about 1 mm s^{-1} through its vascular system (*xylem*), which consists of cylindrical pores of 20-μm radius. Assume the viscosity of sap is the same as the viscosity of water. What pressure difference between the bottom and top of a 100-m tall tree is needed to generate this flow? How does it compare to the hydrostatic pressure difference caused by gravity?

Problem 35.

(a) Consider a small cube of incompressible fluid. Analyze the volume fluence rate for each face of the cube and show that the divergence of **v** is zero. (The divergence is defined in Chap. 4.)

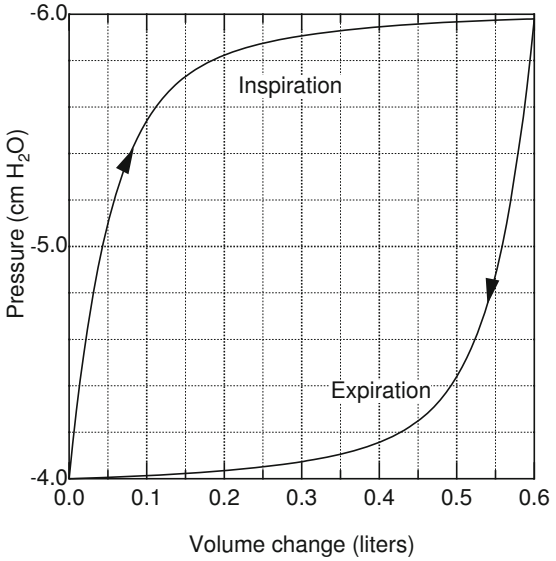
(b) Use the velocity distribution given in Problem 46 and the material in Appendix L to show that for this flow the fluid is incompressible.

Problem 36. Consider Eq. 1.54 when viscosity is negligible and the flow is steady ($\partial v/\partial t = 0$). Show that it reduces to the *Bernoulli equation*

$$p_1 + \rho\frac{v_1^2}{2} + \rho g z_1 = p_2 + \rho\frac{v_2^2}{2} + \rho g z_2.$$

Section 1.18

Problem 37. The accompanying figure shows the negative pressure (below atmospheric) that must be maintained in the thorax during the respiratory cycle by a patient with airway obstruction in order to breathe. Viscous effects are included. Estimate the work in joules done by the body during a breath.



Section 1.19

Problem 38. The volume of blood in a typical person is 5 l, and the volume current through the aorta is about 5 l min^{-1}.

(a) What is the total volume current through all the systemic capillaries?

(b) What is the total volume current through all the pulmonary capillaries?

(c) How long does the blood take to make one complete circuit through the circulatory system?

Problem 39. Find the conversion factor between PRU and Pa m^{-3} s. The total resistance of the systemic circulation was calculated in the text to be 1.66×10^8 Pa m^{-3} s. Express this in PRU.

Problem 40. Equation 1.59 relates the resistance of a vessel to its radius. In the circulatory system, the resistance of an arteriole increases when the smooth muscle surrounding the arteriole contracts, thereby decreasing its radius. By what factor does the resistance increase if the radius decreases by 10 %?

Problem 41. Derive the equations for resistance in a collection of vessels in series and in parallel. Remember that when several vessels are in series, the current is constant and the total pressure change is the sum of the pressure changes along the length of each vessel. When vessels are in parallel, each has the same pressure drop, but the current before the vessels branch is the sum of the currents in each branch.

Problem 42. The velocity of the blood in the aorta is about 0.5 m s^{-1}, and the velocity of the blood in a capillary is about 0.001 m s^{-1}. We have only one aorta, with a diameter of 20 mm, but many capillaries in parallel, each with a diameter of 8 μm. Estimate how many capillaries are typically open at any one time.

Problem 43. Suppose a student asked you, "How can blood be moving more slowly in a capillary than in the aorta? For an incompressible fluid, when the cross-sectional area along a pipe decreases, the velocity increases, so that the volume current *i* is the same. The capillary has a much smaller cross-sectional area than the aorta. Therefore, the blood should move faster in the capillary than in the aorta!" How would you respond to this student?

Problem 44. For Poiseuille flow, find an expression for the maximum shear rate in each vessel from Eq. 1.45. Where in the vessel does it occur? Typical maximum shear rates are 50 s^{-1} in the aorta, 150 s^{-1} in the femoral artery, and 400 s^{-1} in an arteriole.

Problem 45. A sphere of radius a moving through a fluid with speed v is subject to a viscous drag $F_{\text{drag}} = 6\pi\eta a v$. Make an argument similar to that in the text to show that the ratio of kinetic energy of a sphere of fluid of the same size moving at the same speed to the viscous work done to displace the sphere by its own diameter is $N_R/18$.

Problem 46. Consider a stationary sphere of radius a placed in a fluid of viscosity η moving uniformly with speed V. For low Reynolds number flow, the radial and tangential components of the fluid velocity and the pressure surrounding the sphere are

$$v_r = V\cos\theta\left(1-\frac{3a}{2r}+\frac{a^3}{2r^3}\right)$$
$$v_\theta = -V\sin\theta\left(1-\frac{3a}{4r}-\frac{a^3}{4r^3}\right)$$
$$p = -\eta V\cos\theta\frac{3a}{2r^2}.$$

(a) Show that the no-slip boundary condition is satisfied.
(b) Integrate the shear force and the pressure force over the sphere surface and find an expression for the net drag force on the sphere (Stoke's law). What fraction of this force arises from pressure drag, and what fraction from viscous drag?

Problem 47. Find an expression for the entry length in terms of the tube size, the pressure gradient, and the properties of the fluid. Estimate the length of the entry region in the aorta, in an artery, and in an arteriole of radius 20 μm. Use $\eta = 10^{-3}\,\mathrm{kg\,m^{-1}\,s^{-1}}$.

Problem 48. Estimate the tension per unit length and the stress in the walls of various blood vessels using the data in Table 1.4.

Problem 49. Compare the magnitude of the four terms in Eq. 1.42 in the following two cases. Ignore branching. Assume the vessels are vertical. Use $\rho = 10^3\,\mathrm{kg\,m^{-3}}$ and $\eta = 10^{-3}\,\mathrm{Pa\,s}$.

(a) The descending aorta. Assume the length is 35 cm, the radius is 1 cm (independent of distance along the aorta), the peak acceleration of the blood is 1800 $\mathrm{cm\,s^{-2}}$, and the peak velocity (during the cardiac cycle) is 70 $\mathrm{cm\,s^{-1}}$ at the entrance and 60 $\mathrm{cm\,s^{-1}}$ at the exit. (These velocities are different because some of the blood leaves the aorta in major arteries.)
(b) An arteriole of radius 50 μm, length 10 mm, and constant velocity of 5 $\mathrm{mm\,s^{-1}}$ at both entrance and exit.

Problem 50. The viscosity of water (and therefore of blood) is a rapidly decreasing function of temperature. Water at 5 °C is twice as viscous as water at 35 °C. Speculate on the implications of this extreme temperature dependence for the circulatory system of cold-blooded animals. (For a further discussion, see Vogel 1994, pp. 27–31.)

Section 1.20

Problem 51. Estimate the Reynolds number for the following flows. In each case, determine whether the Reynolds number is high ($\gg 1$) or low ($\ll 1$).

(a) *E. coli* (length 2 μ) swim in water at speeds of about 0.01 $\mathrm{mm\,s^{-1}}$.
(b) An Olympic swimmer (length 2 m) swims in water at speeds of up to 2 $\mathrm{m\,s^{-1}}$.
(c) A bald eagle (wingspan 2 m) flies in air (density = 1.2 $\mathrm{kg\,m^{-3}}$, viscosity = 1.8×10^{-5} Pa s) at speeds of 20 $\mathrm{km\,hr^{-1}}$.

Problem 52. Estimate the Reynolds number of blood flow in a capillary, using the data in Table 1.4. How does this compare to that in the aorta?

Problem 53. Consider a sphere of radius R moving at speed v through a fluid of density ρ and viscosity η.

(a) If the Reynolds number is low, then viscous effects dominate and the drag force F_{visc} depends on η and not ρ. Assume that F_{visc} depends only on R, η, and v and use dimensional analysis[13] to determine the form of Stokes' law (i.e., the power to which each variable is raised).
(b) At high Reynolds number, the force needed to accelerate the fluid out of the way is important, and the drag force F_{pres} depends on ρ and not η. Find the dependence of F_{pres} on the relevant variables.
(c) Find, to within a dimensionless factor, the critical speed at which $F_{\mathrm{visc}} = F_{\mathrm{pres}}$.

References

Basford J (2002) Law of Laplace and its relevance to contemporary medicine and rehabilitation. Arch Phys Med Rehabil 83: 1165–1170

Benedek GB, Villars FMH (1973) Physics with illustrative examples from medicine and biology, vol 1. Mechanics. Addison-Wesley, Reading, pp 3–8

Benedek GB, Villars FMH (2000) Physics with illustrative examples from medicine and biology: mechanics. vol 1. Springer-Verlag, New York

Blagoev KB, Shukla K, Levine H (2013) We need theoretical physics approaches to study living systems. Phys Biol 10:040201. doi:10.1088/1478-3975/10/4/040201

Blount WP (1956) Don't throw away the cane. J Bone Joint Surg 38A:695–708

Caro CG, Pedley TJ, Schroter RC, Seed WA (1978) The mechanics of the circulation. Oxford University Press, Oxford

Cebeci T, Bradshaw P (1977) Momentum transfer in boundary layers. Hemisphere, Washington

Denny MW (1993) Air and water: the biology and physics of life's media. Princeton University Press, Princeton

Elliott DM, Narmoneva DA, Setton LA (2002) Direct measurement of the Poisson's ratio of human patella cartilage in tension. J Biomech Eng 124:223–228

Fung YC (1993) Biomechanics: mechanical properties of living tissue, 2nd edn. Springer-Verlag, New York

Goodsell DS (2009). The machinery of life, 2nd edn. Springer-Verlag, New York

Guyton AC (1991) Textbook of medical physiology, 8th edn. Elsevier, Philadelphia p 151

[13] For a detailed discussion of dimensional analysis, see Jensen (2013).

Herrick JF (1942) Poiseuille's observations on blood flow lead to a new law in hydrodynamics. Am J Phys 10:33–39
Inman VT (1947) Functional aspects of the abductor muscles of the hip. J Bone Joint Surg 29:607–619
Jensen JH (2013) Introducing fluid dynamics using dimensional analysis. Am J Phys 81(9):688–694
Köhler T, Vollrath F (1995) Thread biomechanics in the two orb-weaving spiders, *Araneus diadematus* (Araneae, Araneidae) and *Uloborus walckenaerius* (Araneae, Uloboridae). J Exp Zool 271: 1–17
Kummer BKF (1972) Biomechanics of bone. In: Fung YC et al (eds) Biomechanics—Its foundations and objectives. Prentice-Hall, Englewood Cliffs, p 237
LaBarbera M (1990) Principles of design of fluid transport systems in zoology. Science 249:992–1000
Lighthill J (1975) Mathematical biofluiddynamics. Society for Industrial and Applied Mathematics, Philadelphia
Mazumdar JN (1992) Biofluid mechanics. World Scientific, Singapore
Milnor WR (1989) Hemodynamics, 2nd edn. Williams & Wilkins, Baltimore
Morrison P, Morrison P, The Office of Charles and Ray Eames (1994) Powers of ten. Scientific American Library, New York
Patton HD, Fuchs AF, Hille B, Scher AM, Steiner R (eds) (1989) Textbook of physiology, 21st edn. Saunders, Philadelphia
Plonsey R (1995) Physiologic Systems. In: Bronzino JD (ed) The biomedical engineering handbook. CRC Press, Boca Raton, pp 9–10
Purcell EM (1977) Life at low Reynolds number. Am J Phys 45:3–11
Serway RA, Jewett JW (2013) Principles of physics, 5th edn. Brooks/Cole, Boston
Schlichting H, Gersten K (2000) Boundary-layer theory, 8th edn. Springer, New York
Synolakis CE, Badeer HS (1989) On combining the Bernoulli and Poiseuille equation—a plea to authors of college physics texts. Am J Phys 57(11):1013–1019
Trowbridge EA (1982) The fluid mechanics of blood: equilibrium and sedimentation. Clin Phys Physiol Meas 3(4):249–265
Trowbridge EA (1983) The physics of arteriole blood flow. I. General theory. Clin Phys Physiol Meas 4(2):151–175
Trowbridge EA, Meadowcroft PM (1983) The physics of arteriole blood flow. II. Comparison of theory with experiment. Clin Phys Physiol Meas 4(2):177–196
U.S. Department of Agriculture (1955) Wood handbook. Handbook No. 72. U.S. Government Printing Office, Washington, DC, p. 74.
Vogel SV (1992) Vital circuits: on pumps, pipes, and the workings of circulatory systems. Oxford University Press, Oxford
Vogel SV (1994) Life in moving fluids. Princeton University Press, Princeton
Wick GL, Tooby PF (1977) Centrifugal buoyancy forces. Am J Phys 45:1074–1076
Williams M, Lissner HR (1962) Biomechanics of human motion. Saunders, Philadelphia

Exponential Growth and Decay

2

The exponential function is one of the most important and widely occurring functions in physics and biology. In biology it may describe the growth of bacteria or animal populations, the decrease of the number of bacteria in response to a sterilization process, the growth of a tumor, or the absorption or excretion of a drug. (Exponential growth cannot continue forever because of limitations of nutrients, etc.) Knowledge of the exponential function makes it easier to understand birth and death rates, even when they are not constant. In physics, the exponential function describes the decay of radioactive nuclei, the emission of light by atoms, the absorption of light as it passes through matter, the change of voltage or current in some electrical circuits, the variation of temperature with time as a warm object cools, and the rate of some chemical reactions.

In this book, the exponential function will be needed to describe certain probability distributions, the concentration ratio of ions across a cell membrane, the flow of solute particles through membranes, the decay of a signal traveling along a nerve axon, and the return of some physiologic variables to their equilibrium values after they have been disturbed.

Because the exponential function is so important, and because we have seen many students who did not understand it even after having been exposed to it, the chapter starts with a gentle introduction to exponential growth (Sect. 2.1) and decay (Sect. 2.2). Section 2.3 shows how to analyze exponential data using semilogarithmic graph paper. The next section shows how to use semilogarithmic graph paper to find instantaneous growth or decay rates when the rate varies. Some would argue that the availability of computer programs that automatically produce logarithmic scales for plots makes these sections unnecessary. We feel that intelligent use of semilogarithmic and logarithmic (log–log) plots requires an understanding of the basic principles.

Variable rates are described in Sect. 2.4. Clearance, discussed in Sect. 2.5, is an exponential decay process that is important in physiology. Microbiologists often grow cells in a chemostat, described in Sect. 2.6. Sometimes there are competing paths for exponential removal of a substance: multiple decay paths are introduced in Sect. 2.7. A very basic and simple model for many processes is the combination of input at a fixed rate accompanied by exponential decay, described in Sect. 2.8. Sometimes a substance exists in two forms, each with its own decay rate. One then must fit two or more exponentials to the set of data, as shown in Sect. 2.9.

Section 2.10 discusses the logistic equation, one possible model for a situation in which the growth rate decreases as the amount of substance increases. The chapter closes with a section on power–law relationships. While not exponential, they are included because data analysis can be done with log–log graph paper, a technique similar to that for semilog paper. If you feel mathematically secure, you may wish to skim the first four sections, but you will probably find the rest of the chapter worth reading.

2.1 Exponential Growth

An exponential growth process is one in which the rate of increase of a quantity is proportional to the present value of that quantity. The simplest example is a savings account. If the interest rate is 5 % and if the interest is credited to the account once a year, the account increases in value by 5 % of its present value each year. If the account starts out with $ 100, then at the end of the first year, $ 5 is credited to the account and the value becomes $ 105. At the end of the second year, 5 % of $ 105 is credited to the account and the value grows by $ 5.25 to 110.25. The growth of such an account is shown in Table 2.1 and Fig. 2.1. These amounts can be calculated as follows: At the end of the first year, the original amount, y_0, has been augmented by $(0.05)y_0$:

$$y_1 = y_0(1 + 0.05).$$

During the second year, the amount y_1 increases by 5 %, so

$$y_2 = y_1(1.05) = y_0(1.05)(1.05) = y_0(1.05)^2.$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_2, © Springer International Publishing Switzerland 2015

Table 2.1 Growth of a savings account earning 5 % interest compounded annually, when the initial investment is $ 100

Year	Amount ($)	Year	Amount ($)	Year	Amount ($)
1	105.00	10	162.88	100	1.31×10^4
2	110.25	20	265.33	200	1.73×10^6
3	115.76	30	432.19	300	2.27×10^8
4	121.55	40	704.00	400	2.99×10^{10}
5	127.63	50	1146.74	500	3.93×10^{12}
6	134.01	60	1867.92	600	5.17×10^{14}
7	140.71	70	3042.64	700	6.80×10^{16}
8	147.75	80	4956.14	800	8.94×10^{18}
9	155.13	90	8073.04	900	1.18×10^{21}

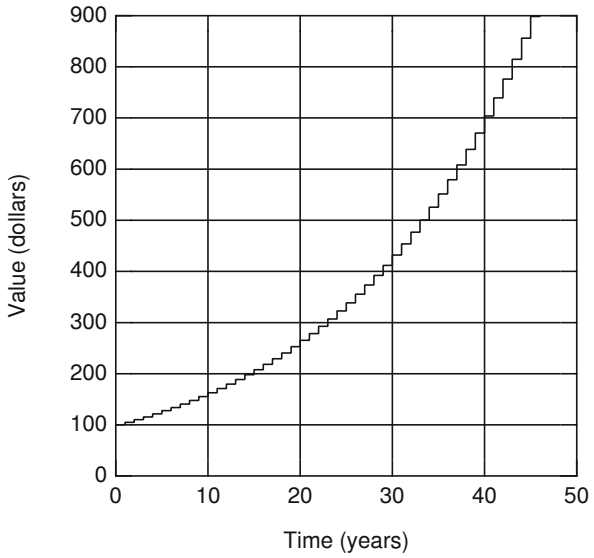


Fig. 2.1 The amount in a savings account after t years, when the amount is compounded annually at 5 % interest

After t years, the amount in the account is

$$y_t = y_0(1.05)^t.$$

In general, if the growth rate is b per compounding period, the amount after t periods is

$$y_t = y_0(1+b)^t. \tag{2.1}$$

It is possible to keep the same annual growth (interest) rate, but to compound more often than once a year. Table 2.2 shows the effect of different compounding intervals on the amount, when the interest rate is 5 %. The last two columns, for monthly compounding and for "instant interest," are listed to the nearest tenth of a cent to show the slight difference between them.

The table entries were calculated in the following way: Suppose that compounding is done N times a year. In t years, the number of compoundings is Nt. If the annual fractional rate of increase is b, the increase per compounding is b/N. For 6 months at 5 % ($b = 0.05$), the increase is 2.5, for 3 months it is 1.25, etc. The amount after t units of time (years) is, in analogy with Eq. 2.1,

$$y = y_0\,(1 + b/N)^{Nt}\,. \tag{2.2}$$

Recall (refer to Appendix C) that $(a)^{bc} = (a^b)^c$. The expression for y can be written as

$$y = y_0\left[(1 + b/N)^N\right]^t. \tag{2.3}$$

Most calculus textbooks show that the quantity

$$(1 + b/N)^N \to e^b$$

as N becomes very large. (Rather than proving this fact here, we give numerical examples in Table 2.3 for two different values of b.) Therefore, Eq. 2.3 can be rewritten as

$$y = y_0 e^{bt} = y_0 \exp(bt). \tag{2.4}$$

(The exp notation is used when the argument is complicated.) To calculate the amount for instant interest, it is necessary only to multiply the fractional growth rate per unit time b by the length of the time interval and then look up the exponential function of this amount in a table or evaluate it with a computer or calculator. The number e is approximately equal to 2.71828 . . . and is called the *base of the natural logarithms*. Like π (3.14159 . . .), e has a long history (Maor 1994).

Table 2.2 Amount of an initial investment of $ 100 at 5 % annual interest, with different methods of compounding

Month	Annual ($)	Semiannual ($)	Quarterly ($)	Monthly ($)	Instant ($)
0	100.00	100.00	100.00	100.000	100.000
1	100.00	100.00	100.00	100.417	100.418
2	100.00	100.00	100.00	100.835	100.837
3	100.00	100.00	101.25	101.255	101.258
4	100.00	100.00	101.25	101.677	101.681
5	100.00	100.00	101.25	102.101	102.105
6	100.00	102.50	102.52	102.526	102.532
7	100.00	102.50	102.52	102.953	102.960
8	100.00	102.50	102.52	103.382	103.390
9	100.00	102.50	103.80	103.813	103.821
10	100.00	102.50	103.80	104.246	104.255
11	100.00	102.50	103.80	104.680	104.690
12	105.00	105.06	105.09	105.116	105.127

Table 2.3 Numerical examples of the convergence of $(1 + b/N)^N$ to e^b as N becomes large

N	$b = 1$	$b = 0.05$
10	2.594	1.0511
100	2.705	1.0513
1000	2.717	1.0513
e^b	2.718	1.0513

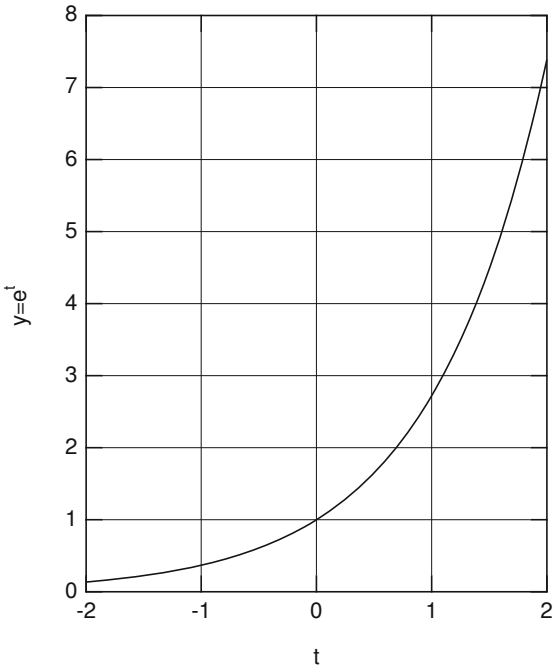


Fig. 2.2 A graph of the exponential function $y = e^t$

The exponential function is plotted in Fig. 2.2. (The meaning of negative values of t will be considered in the next section.) This function increases more and more rapidly as t increases. This is expected, since the rate of growth is always proportional to the present amount. This is also reflected in the following property of the exponential function:

$$\frac{d}{dt}\left(e^{bt}\right) = be^{bt}. \tag{2.5}$$

This means that the function $y = y_0 e^{bt}$ has the property that

$$\frac{dy}{dt} = by. \tag{2.6}$$

Any constant multiple of the exponential function e^{bt} has the property that its rate of growth is b times the function itself. Whenever we see the exponential function, we know that it satisfies Eq. 2.6. Equation 2.6 is an example of a *differential equation*. If you learn how to solve only one differential equation, let it be Eq. 2.6. Whenever we have a problem in which the growth rate of something is proportional to the present amount, we can expect to have an exponential solution. Notice that for time intervals t that are not too large, Eq. 2.6 implies that $\Delta y = (b\Delta t)y$. This again says that the increase in y is proportional to y itself.

The independent variable in this discussion has been t. It can represent time, in which case b is the fractional growth rate per unit time; distance, in which case b is the fractional growth per unit distance; or something else. We could, of course, use another symbol such as x for the independent variable, in which case we would have $dy/dx = by$, $y = y_0 e^{bx}$.

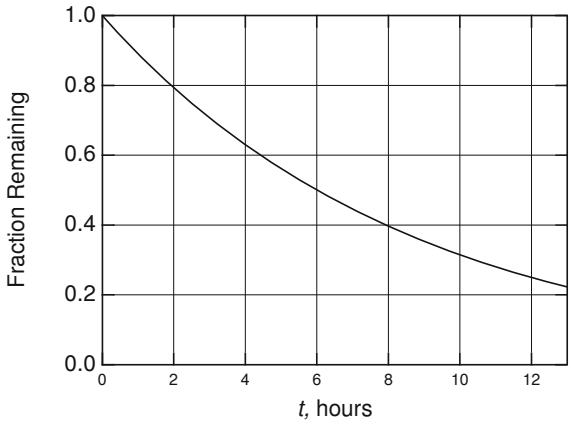


Fig. 2.3 A plot of the fraction of nuclei of ^{99m}Tc surviving at time t

2.2 Exponential Decay

Figure 2.2 shows the exponential function for negative values of t as well as positive ones. (Remember that $e^{-t} = 1/e^t$.) To see what this means, consider a bank account in which no interest is credited, but from which 5 % of what remains is taken each year. If the initial balance is \$ 100, \$ 5 is removed the first year to leave \$ 95.00. In the second year, 5 % of \$ 95 or \$ 4.75 is removed. In the third year, 5 % of \$ 90.25 or \$ 4.51 is removed. The annual decrease in y becomes less and less as y becomes less and less. The equations developed in the preceding section also describe this situation. It is only necessary to call b the fractional decay and allow it to have a negative value, $-|b|$. Equation 2.1 then has the form $y = y_0(1 - |b|)^t$ and Eq. 2.4 is

$$y = y_0 e^{-|b|t}. \tag{2.7}$$

Often b is regarded as being intrinsically positive, and Eq. 2.7 is written as

$$y = y_0 e^{-bt}. \tag{2.8}$$

One could equally well write $y = y_0 e^{bt}$ and regard b as being negative, but this can cause confusion, for example with Eq. 2.10 below.

The radioactive isotope ^{99m}Tc (read as technetium-99) has a fractional decay rate $b = 0.1155\,\mathrm{h}^{-1}$. If the number of atoms at $t = 0$ is y_0, the fraction $f = y/y_0$ remaining at later times decreases as shown in Fig. 2.3. The equation that describes this curve is

$$f = \frac{y}{y_0} = e^{-bt}, \tag{2.9}$$

where t is the elapsed time in hours and $b = 0.1155\,\mathrm{h}^{-1}$. The product bt must be dimensionless, since it is in the exponent.

People often talk about the *half-life* $T_{1/2}$, which is the length of time required for f to decrease to one-half. From

inspection of Fig. 2.3, the half-life is 6 h. This can also be determined from Eq. 2.9:

$$0.5 = e^{-bT_{1/2}}.$$

From a table of exponentials, one finds that $e^{-x} = 0.5$ when $x = 0.69315$. This leads to the very useful relationship $bT_{1/2} = 0.693$ or

$$T_{1/2} = \frac{0.693}{b}. \tag{2.10}$$

For the case of ^{99m}Tc, the half-life is $T_{1/2} = 0.693/0.1155 = 6$ h.

One can also speak of a *doubling time* if the exponent is positive. In that case, $2 = e^{bT_2}$, from which

$$T_2 = \frac{0.693}{b}. \tag{2.11}$$

2.3 Semilog Paper

A special kind of graph paper, called *semilog paper*, makes the analysis of exponential growth and decay problems much simpler. If one takes logarithms (to any base) of Eq. 2.4, one has

$$\log y = \log y_0 + bt \log e. \tag{2.12}$$

If the dependent variable is considered to be $u = \log y$, and since $\log y_0$ and $\log e$ are constants, this equation is of the form

$$u = c_1 + c_2 t. \tag{2.13}$$

The graph of u vs t is a straight line with positive slope if b is positive and negative slope if b is negative.

On semilog paper the vertical axis is marked in a logarithmic fashion. The graph can be plotted without having to calculate any logarithms. Figure 2.4 shows a plot of the exponential function of Fig. 2.2, for both positive and negative values of t. First, note how to read the vertical axis. A given distance along the axis always corresponds to the same multiplicative factor. Each cycle represents a factor of ten. To use the paper, it is necessary first to mark off the decades with the desired values. In Fig. 2.4, the decades have been marked 0.1, 1, 10, and 100. The 6 that lies between 0.1 and 1 is 0.6; the 6 between 1 and 10 is 6.0; the 6 between 10 and 100 represents 60; and so forth. The paper can be imagined to go vertically forever in either direction; one never reaches zero. Figure 2.4 has two examples marked on it with dashed lines. The first shows that for $t = -1.0$, $y = 0.36$; the second shows that for $t = +1.5$, $y = 4.5$.

Semilog paper is most useful for plotting data that you suspect may have an exponential relationship. If the data plot as a straight line, your suspicions are confirmed. From the

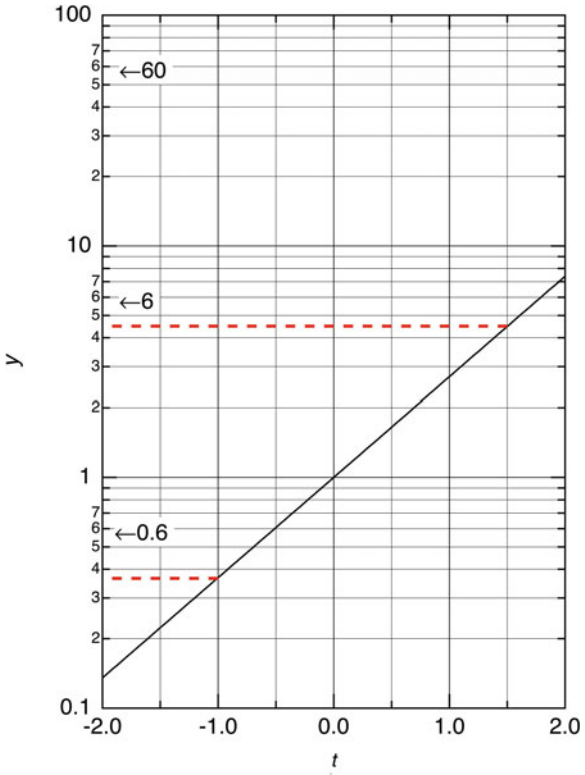


Fig. 2.4 A plot of the exponential function on semilog paper

straight line, you can determine the value of b. Figure 2.5 is a plot of the intensity of light that passed through an absorber in a hypothetical example. The independent variable is absorber thickness x. The decay is exponential, except for the last few points, which may be high because of experimental error. (As the intensity of the light decreases, it becomes harder to measure accurately.) We wish to determine the decay constant in $y = y_0 e^{-bx}$. One way to do it would be to note (dashed line A in Fig. 2.5) that the half-distance is 0.145 cm, so that, from Eq. 2.10,

$$b = \frac{0.693}{0.145} = 4.8 \text{ cm}^{-1}.$$

This technique can be inaccurate because it is difficult to read the graph accurately. It is more accurate to use a portion of the curve for which y changes by a factor of 10 or 100. The general relationship is $y = y_0 e^{bx}$, where the value of b can be positive or negative. If two different values of x are selected, one can write

$$\frac{y_2}{y_1} = \frac{y_0 e^{bx_2}}{y_0 e^{bx_1}} = e^{b(x_2 - x_1)}.$$

If $y_2/y_1 = 10$, then this equation has the form $10 = e^{bX_{10}}$ where $X_{10} = x_2 - x_1$ when $y_2/y_1 = 10$. From a table of

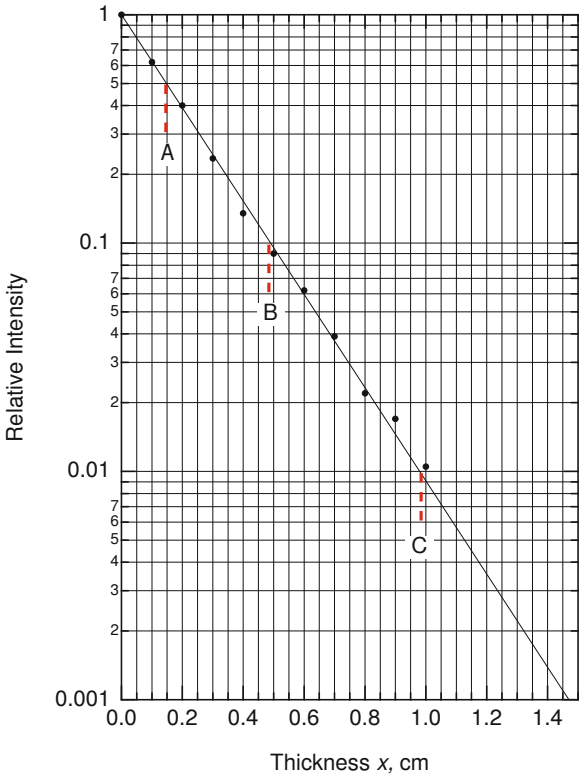


Fig. 2.5 A semilogarithmic plot of the intensity of light after it has passed through an absorber of thickness x

exponentials, $bX_{10} = 2.303$, so that

$$b = \frac{2.303}{X_{10}}. \tag{2.14}$$

The same procedure can be used to find b using a factor of 100 change in y:

$$b = \frac{4.605}{X_{100}}. \tag{2.15}$$

If the curve represents a decaying exponential, then $y_2/y_1 = 10$ when $x_2 < x_1$, so that $X_{10} = x_2 - x_1$ is negative. Equation 2.14 then gives a negative value for b. It is customary to state separately that we are dealing with decay and regard b as positive.

As an example, consider the exponential decay in Fig. 2.5. Using points B and C, we have $x_1 = 0.97$, $y_1 = 10^{-2}$, $x_2 = 0.48$, $y_2 = 10^{-1}$, $X_{10} = 0.480 - 0.97 = -0.49$. Therefore, $b = 2.303/(0.49) = 4.7\ \text{cm}^{-1}$, which is a more accurate determination than the one we made using the half-life.

When we are dealing with real data, we must consider the fact that each measurement has an experimental error associated with it. If we make several measurements of y for a particular value of the independent variable x, the values of y will be scattered. We indicate this by the error bars in

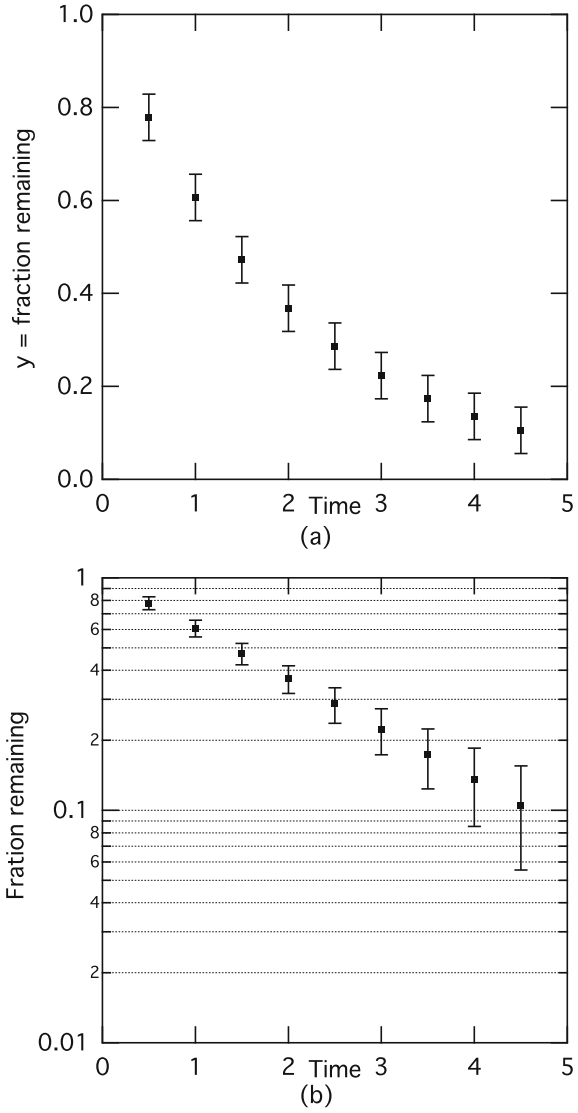


Fig. 2.6 Plot of $y = e^{-0.5t}$ with error bars ± 0.05 on linear (**a**) and semilog paper (**b**)

Fig. 2.6. (Determining the size of these error bars is discussed in Chap. 11.) The data points in Fig. 2.6 are given exactly by $y = e^{-0.5x}$, where y is the fraction remaining at time x. There is no data point for $x = 0$, but we must make sure that our fitting line passes through the point (0,1). The error bars show an error of ± 0.09. The error bars on the semilog plot are not all the same length, being much larger for long times (small values of y). If we do not plot the error bars before drawing our line, we will give too much emphasis to the data points for small y.

Equal error bars for all the points on a semilog plot correspond to the same percentage error for each point, as shown in Fig. 2.7.

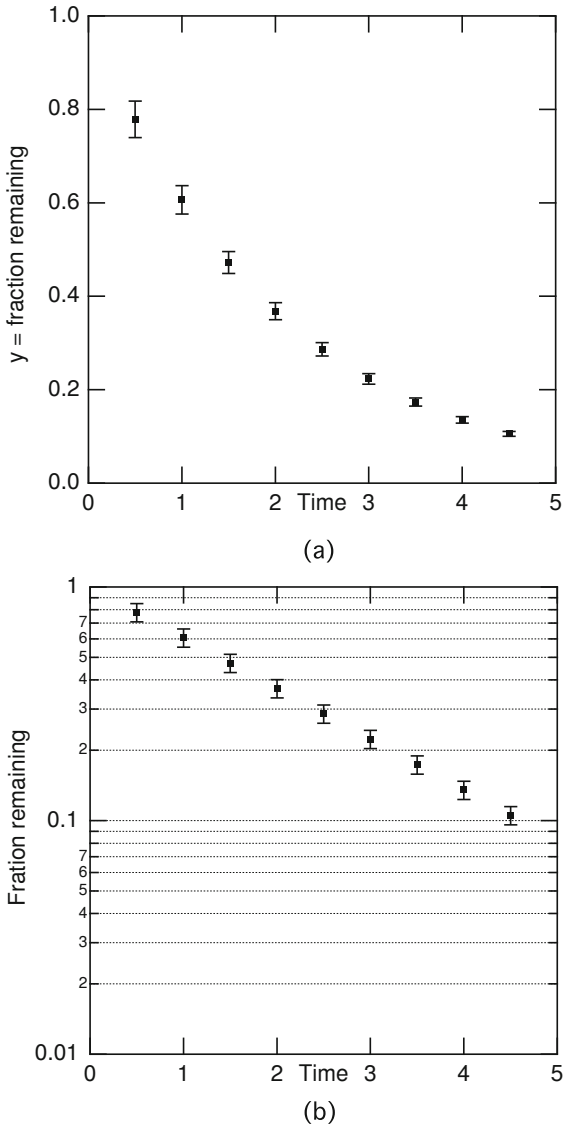


Fig. 2.7 Plot of $y = e^{-0.5t}$ with 5 % error bars in linear (**a**) and semilog paper (**b**)

2.4 Variable Rates

The equation $dy/dx = by$ (or $dy/dt = by$) says that y grows or decays at a rate that is proportional to y. The constant b is the *fractional rate of growth or decay*. It is possible to define the fractional rate of growth or decay even if it is not constant but is a function of x:

$$b(x) = \frac{1}{y}\frac{dy}{dx}. \tag{2.16}$$

Semilogarithmic graph paper can be used to analyze the curve even if b is not constant. Since $d(\ln y)/dy = 1/y$, the chain rule for evaluating derivatives gives

$$\frac{d}{dx}(\ln y) = \frac{1}{y}\frac{dy}{dx} = b.$$

This means that $b(x)$ is the slope of a plot of $\ln y$ vs x. A semilogarithmic plot of y vs x is shown in Fig. 2.8. The straight lines are tangent to the curve and decay with a constant rate equal to $b(x)$ at the point of tangency. The ordinate in Fig. 2.8 can be the log of y to any base; the value of b for the tangent line is determined using the methods in the previous section.

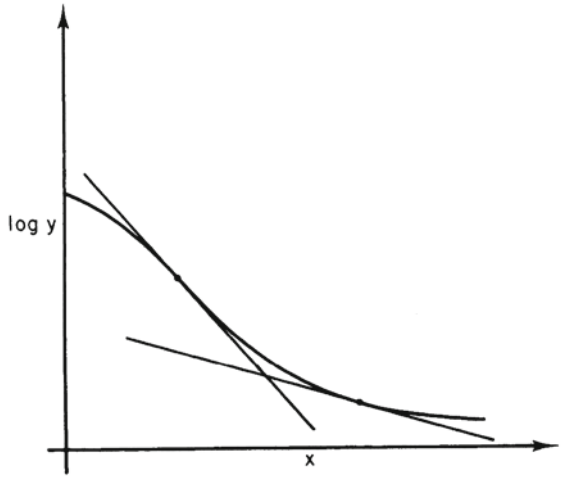


Fig. 2.8 A semilogarithmic plot of y vs x when the decay rate is not constant. Each *tangent line* represents the instantaneous decay rate for that value of x

If finite changes Δx and Δy have been measured, they may be used to estimate $b(x)$ directly from Eq. 2.16. For example, suppose that y=100,000 people and that in $\Delta x = 1$ year there is a change $\Delta y = -37$. In this case, Δy is very small compared to y, so we can say that $b = (1/y)(\Delta y/\Delta x) = -37 \times 10^{-5}\,\mathrm{y}^{-1}$. If the only cause of change in this population is deaths, the absolute value of b is called the *death rate*.

A plot of the number of people surviving in a population, all of whom have the same disease, can provide information about the prognosis for that disease. The death rate is equivalent to the decay constant. An example of such a plot is shown in Fig. 2.9. Curve A shows a disease for which the death rate is constant. Curve B shows a disease with an initially high death rate that decreases with time; if the patient survives the initial period, the prognosis is much better. Curve C shows a disease for which the death rate increases with time.

Surprisingly, there are a few diseases that have death rates independent of the duration of the disease (Zumoff et al. 1966). Any discussion of mortality should be made in terms

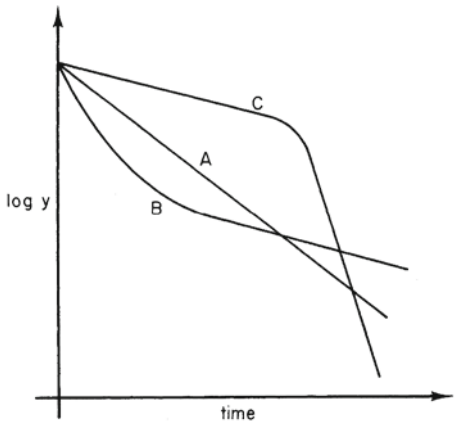


Fig. 2.9 Semilogarithmic plots of the fraction of a population surviving in three different diseases. The death rates (decay constants) depend on the duration of the disease

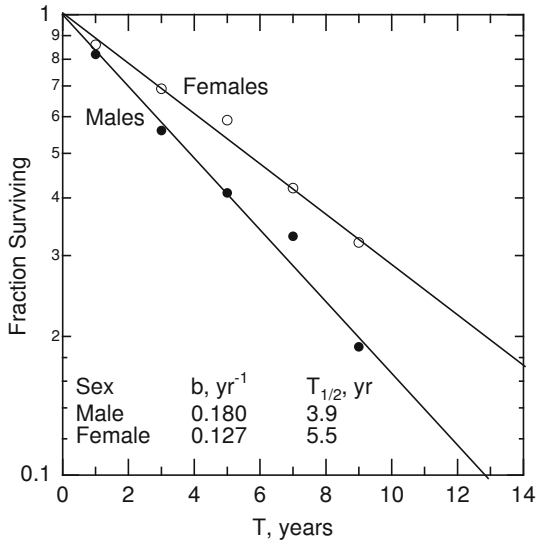


Fig. 2.10 Survival of patients with congestive heart failure. (Data are from McKee et al. 1971)

of the surviving population, since any further deaths must come from that group. Nonetheless, one often finds results in the literature reported in terms of the cumulative fraction of patients who have died. Figure 2.10 shows the survival of patients with congestive heart failure for a period of 9 years. The data are taken from the Framingham study (McKee et al. 1971; Levy and Brink 2005); the death rate is constant during this period. For a more detailed discussion of various possible survival distributions, see Clark (1975).

As long as b has a constant value, it makes no difference what time is selected to be $t = 0$. To see this, suppose that the value of y decays exponentially with constant rate: $y = y_0 e^{-bt}$. Consider two different time scales, shifted with respect to each other so that $t' = t_0 + t$. In terms of the shifted

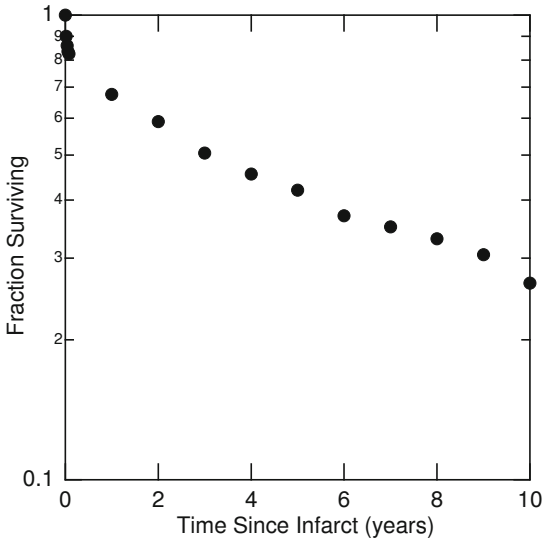


Fig. 2.11 The fraction of patients surviving after a myocardial infarction (heart attack) at $t = 0$. The mortality rate decreases with time. (From data in Bland and White 1941)

time t', the value of y is

$$y = y_0 e^{-bt} = y_0 e^{-b(t'-t_0)} = \left(y_0 e^{bt_0}\right) e^{-bt'}.$$

This has the same form as the original expression for $y(t)$. The value of y_0' is $y_0 e^{bt_0}$, which reflects the fact that $t' = 0$ occurs at an earlier time than $t = 0$, so $y_0' > y_0$.

If the decay rate is not constant, then the origin of time becomes quite important. Usually there is something about the problem that allows $t = 0$ to be determined. Figure 2.11 shows survival after a heart attack (myocardial infarct). The time of the initial infarct defines $t = 0$; if the origin had been started 2 or 3 years after the infarct, the large initial death rate would not have been seen.

As long as the rate of increase can be written as a function of the independent variable, Eq. 2.16 can be rewritten as $dy/y = b(x)dx$. This can be integrated:

$$\begin{aligned} \int_{y_1}^{y_2} \frac{dy}{y} &= \int_{x_1}^{x_2} b(x)\,dx, \\ \ln(y_2/y_1) &= \int_{x_1}^{x_2} b(x)\,dx, \\ \frac{y_2}{y_1} &= \exp\left(\int_{x_1}^{x_2} b(x)\,dx\right). \end{aligned} \tag{2.17}$$

If we can integrate the right-hand side analytically, numerically, or graphically, we can determine the ratio y_2/y_1.

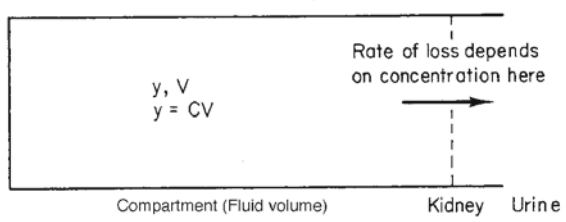


Fig. 2.12 A case in which the rate of removal of a substance from the a fluid compartment depends on the concentration, not on the total amount of substance in the compartment. Increasing the compartment volume with the same concentration of the substance would not change the rate of removal

2.5 Clearance

In some cases in physiology, the amount of a substance may decay exponentially because the rate of removal is proportional to the concentration of the substance (amount per unit volume) instead of to the total amount. For example, the rate at which the kidneys excrete a substance may be proportional to the concentration in the blood that passes through the kidneys, while the total amount depends on the total fluid volume in which the substance is distributed. This is shown schematically in Fig. 2.12. The large box on the left represents the total fluid volume V. It contains a total amount of some substance, y. If the fluid is well mixed, the concentration is $C = y/V$. The removal process takes place only at the dashed line, at a rate proportional to C. The equation describing the change of y is

$$\frac{dy}{dt} = -KC = -K\left(\frac{y}{V}\right). \tag{2.18}$$

The proportionality constant K is called the *clearance*. Its units are $\mathrm{m}^3\,\mathrm{s}^{-1}$. The equation is the same as Eq. 2.6 if K/V is substituted for b. The solution is

$$y = y_0 e^{-(K/V)t}. \tag{2.19}$$

The basic concept of clearance is best remembered in terms of Fig. 2.12. Other definitions are found in the literature. It sometimes takes considerable thought to show that the definitions are equivalent. A common definition in physiology books is "*clearance is the volume of plasma from which y is completely removed per unit time*." To see that this definition is equivalent, imagine that y is removed from the body by removing a volume V of the plasma in which the concentration of y is C. The rate of loss of y is the concentration times the rate of volume removal:

$$\frac{dy}{dt} = -\left|\frac{dV}{dt}\right| C. \tag{2.20}$$

(dV/dt is negative for removal.) Comparison with Eq. 2.18 shows that $|dV/dt| = K$.

As long as the compartment containing the substance is well mixed, the concentration will decrease uniformly throughout the compartment as y is removed. The concentration also decreases exponentially:

$$C = C_0 e^{-(K/V)t}. \tag{2.21}$$

An example may help to clarify the distinction between b and K. Suppose that the substance is distributed in a fluid volume $V = 18\,\mathrm{l}$. The substance has an initial concentration $C_0 = 3\,\mathrm{mg\,l^{-1}}$ and the clearance is $K = 21\,\mathrm{h^{-1}}$. The total amount is $y_0 = C_0 V = 3 \times 18 = 54\,\mathrm{mg}$. The fractional decay rate is $b = K/V = 1/9\,\mathrm{h^{-1}}$. The equations for C and y are $C = (3\,\mathrm{mg\,l^{-1}})e^{-t/9}$, $y = (54\,\mathrm{mg})e^{-t/9}$. At $t = 0$, the initial rate of removal is $-dy/dt = 54/9 = 6\,\mathrm{mg\,h^{-1}}$.

Now double the fluid volume to $V = 36\,\mathrm{l}$ without adding any more of the substance. The concentration falls to $1.5\,\mathrm{mg\,l^{-1}}$ although y_0 is unchanged. The rate of removal is also cut in half, since it is proportional to K/V and the clearance is unchanged. The concentration and amount are now $C = 1.5e^{-t/18}$, $y = 54e^{-t/18}$. The initial rate of removal is $dy/dt = 54/18 = 3\,\mathrm{mg\,h^{-1}}$. It is half as large as above, because C is now half as large.

If more of the substance were added along with the additional fluid, the initial concentration would be unchanged, but y_0 would be doubled. The fractional decay rate would still be $K/V = 1/18\,\mathrm{h^{-1}}$: $C = 3.0e^{-t/18}$, $y = 108e^{-t/18}$. The initial rate of disappearance would be $dy/dt = 108/18 = 6\,\mathrm{mg\,h^{-1}}$. It is the same as in the first case, because the initial concentration is the same.

2.6 The Chemostat

The *chemostat* is used by bacteriologists to study the growth of bacteria (Hagen 2010). It allows the rapid growth of bacteria to be observed over a longer time scale. Consider a container of bacterial nutrient of volume V. It is well stirred and contains y bacteria with concentration $C = y/V$. Some of the nutrient solution is removed at rate Q and replaced by fresh nutrient. The bacteria in the solution are reproducing at rate b. The rate of change of y is

$$\frac{dy}{dt} = by - QC = by - \frac{Qy}{V}. \tag{2.22}$$

Therefore the growth rate is slowed to

$$b - \frac{Q}{V}$$

and can be adjusted by varying Q.

2.7 Multiple Decay Paths

It is possible to have several independent paths by which y can disappear. For example, there may be several competing ways by which a radioactive nucleus can decay, a radioactive isotope given to a patient may decay radioactively and be excreted biologically at the same time, a substance in the body can be excreted in the urine and metabolized by the liver, or patients may die of several different diseases.

In such situations the total decay rate b is the sum of the individual rates for each process, as long as the processes act independently and the rate of each is proportional to the present amount (or concentration) of y:

$$\frac{dy}{dt} = -b_1 y - b_2 y - b_3 y - \cdots = -(b_1 + b_2 + b_3 + \cdots)y = -by. \tag{2.23}$$

The equation for the disappearance of y is the same as before, with the total decay rate being the sum of the individual rates. The rate of disappearance of y by the ith process *is not* dy/dt but is $-b_i y$. Instead of decay rates, one can use half-lives. Since $b = b_1 + b_2 + b_3 + \cdots$, the total half-life T is given by

$$\frac{0.693}{T} = \frac{0.693}{T_1} + \frac{0.693}{T_2} + \frac{0.693}{T_3} + \cdots$$

or

$$\frac{1}{T} = \frac{1}{T_1} + \frac{1}{T_2} + \frac{1}{T_3} + \cdots. \tag{2.24}$$

2.8 Decay Plus Input at a Constant Rate

Suppose that in addition to the removal of y from the system at a rate $-by$, y enters the system at a constant rate a, independent of y and t. The net rate of change of y is given by

$$\frac{dy}{dt} = a - by. \tag{2.25}$$

It is often easier to write down a differential equation describing a problem than it is to solve it. In this case the solution to the equation and the techniques for solving it are well known. However, a good deal can be learned about the solution by examining the equation itself. Suppose that $y(0) = 0$. Then the equation at $t = 0$ is $dy/dt = a$, and y initially grows at a constant rate a. As y builds up, the rate of growth decreases from this value because of the $-by$ term. Finally when $a - by = 0$, dy/dt is zero and y stops growing. This is enough information to make the sketch in Fig. 2.13.

The equation is solved in Appendix F. The solution is

$$y = \frac{a}{b}\left(1 - e^{-bt}\right). \tag{2.26}$$

The derivative of y is $dy/dt = \left(\frac{a}{b}\right)(-1)(-b)e^{-bt} = ae^{-bt}$.

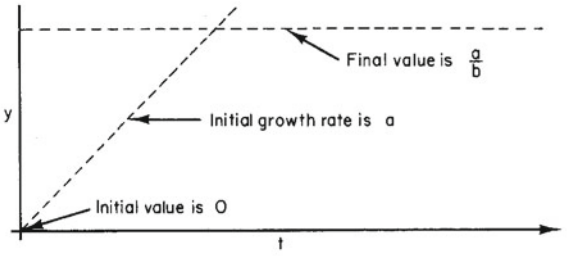


Fig. 2.13 Sketch of the initial slope a and final value a/b of y when $y(0) = 0$

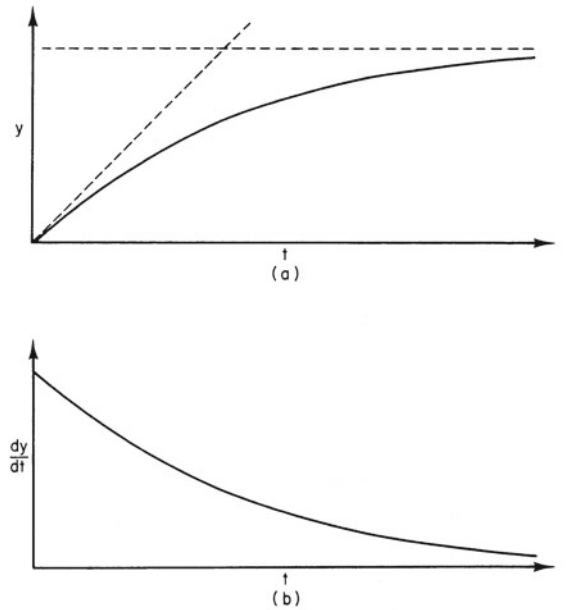


Fig. 2.14 **a** Plot of $y(t)$. **b** Plot of dy/dt

You can verify by substitution that Eq. 2.26 satisfies Eq. 2.25. The solution does have the properties sketched in Fig. 2.13, as you can see from Fig. 2.14. The initial value of dy/dt is a, and it decreases exponentially to zero. When t is large, the exponential term in y vanishes, leaving $y = a/b$.

2.9 Decay With Multiple Half-Lives and Fitting Exponentials

Sometimes y is a mixture of two or more quantities, each decaying at a constant rate. It might represent a mixture of radioactive isotopes, each decaying at its own rate. A biological example is the survival of patients after a myocardial infarct (Fig. 2.11). The death rate is not constant, and many models can be proposed to explain why. One possible model is that there are two distinct classes of patients immediately after the infarct. Each class has an associated death rate that

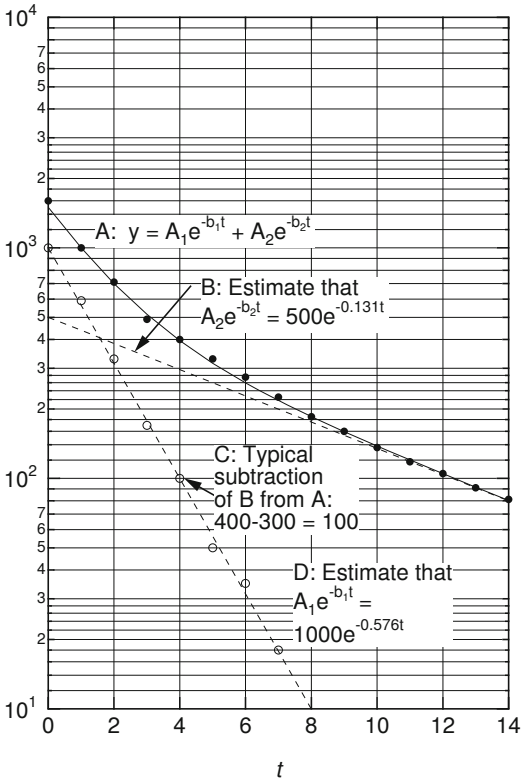


Fig. 2.15 Fitting a curve with two exponentials

is constant. After 3 years, virtually none of the subgroup with the higher death rate remains. Another model is that the death rate is higher right after the infarct for all patients. This higher death rate is due to causes associated with the myocardial injury: irritability of the muscle, arrhythmias in the heartbeat, the weakening of the heart wall at the site of the infarct, and so forth. After many months, the heart has healed, scar tissue has replaced the necrotic (dead) muscle, and deaths from these causes no longer occur.

Whatever the cause, it is sometimes useful to fit a set of experimental data with a sum of exponentials. It should be clear from the discussion of survival after myocardial infarction that simply fitting with an exponential or a sum of exponentials does not prove anything about the decay mechanism.

If y consists of two quantities, y_1 and y_2, each with its own decay rate, then

$$y = y_1 + y_2 = A_1 e^{-b_1 t} + A_2 e^{-b_2 t}. \tag{2.27}$$

Suppose that $b_1 > b_2$, so that y_1 decays more rapidly than y_2. After enough time has elapsed, y_1 will be much less than y_2, and its effect on a semilog plot will be negligible. A typical plot of y is curve A in Fig. 2.15. Line B can then be drawn through the data and used to determine A_2 and b_2. This line is extrapolated back to earlier times, so that y_2 can be subtracted from y to give an estimate for y_1. For example, at point C ($t = 4$), $y = 400$, $y_2 = 300$, and $y_1 = 100$. At $t = 0$, $y_1 = 1500 - 500 = 1000$. For times greater than 5 s, the curves for y and y_2 are close together, and error in reading the graph produces considerable scatter in y_1. Once several values of y_1 have been determined, line D is drawn, and parameters A_1 and b_1 are estimated.

This technique can be extended to several exponentials. However it becomes increasingly difficult to extract meaningful parameters as more exponentials are used, because the estimated parameters for the short-lived terms are very sensitive to the initial guess for the parameters of the longest-lived term. Fig. 2.6 suggests that estimating the parameters for the longest-lived term may be difficult because of the potentially large error bars associated with the data for small values of y. For a discussion of this problem, see Riggs (1970, pp. 146–163). A more modern and better way to fit multiple exponentials is the technique of nonlinear least squares. This is discussed in Sect. 11.2.

2.10 The Logistic Equation

Exponential growth cannot go on forever. This fact is often ignored by economists and politicians. Albert Bartlett has written extensively on this subject. You can find several references in *The American Journal of Physics* and *The Physics Teacher*. See the summary in Bartlett (2004).

Sometimes a growing population will level off at some constant value. Other times the population will grow and then crash. One model that exhibits leveling off is the *logistic model*, described by the differential equation

$$\frac{dy}{dt} = b_0 y \left(1 - \frac{y}{y_\infty}\right), \tag{2.28}$$

where b_0 and y_∞ are constants. This equation has constant solutions $y = 0$ and $y = y_\infty$. If $y \ll y_\infty$, then the equation is approximately $dy/dt = b_0 y$ and y grows exponentially. As y becomes larger, the term in parentheses reduces the rate of increase of y, until y reaches the saturation value y_∞. This might happen, for example, as the population begins to consume a significant fraction of the food supply, causing the birth rate to decrease or the mortality rate to increase.

If the initial value of y is y_0, the solution of Eq. 2.28 is

$$\begin{aligned} y(t) &= \frac{1}{\dfrac{1}{y_\infty} + \left(\dfrac{1}{y_0} - \dfrac{1}{y_\infty}\right) e^{-b_0 t}} \\ &= \frac{y_0 y_\infty}{y_0 + (y_\infty - y_0) e^{-b_0 t}}. \end{aligned} \tag{2.29}$$

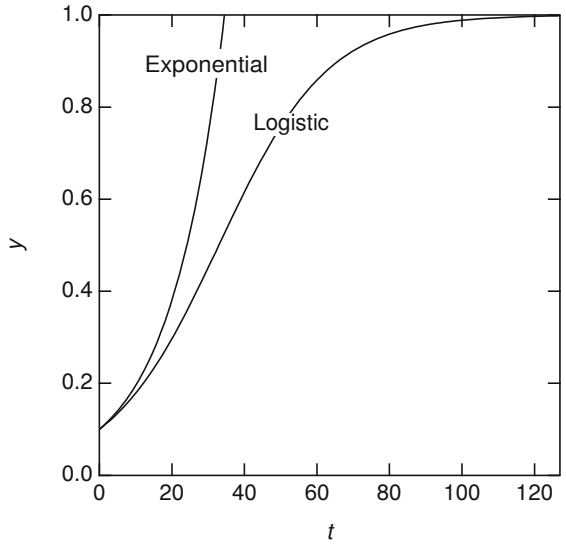


Fig. 2.16 Plot of the solution of the logistic equation when $y_0 = 0.1$, $y_\infty = 1.0$, $b_0 = 0.0667$. Exponential growth with the same values of y_0 and b_0 is also shown

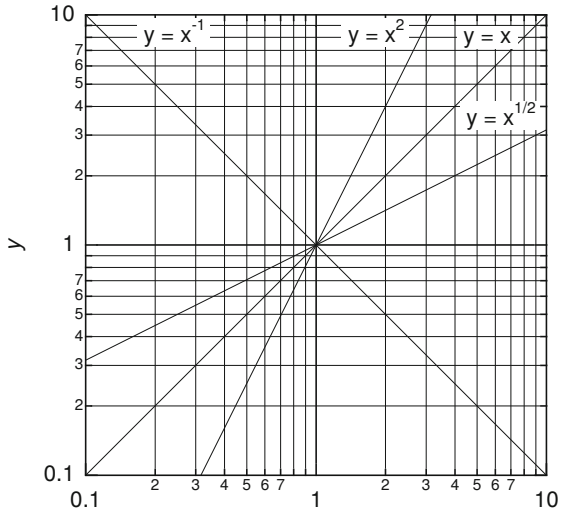


Fig. 2.17 Log–log plots of $y = x^n$ for different values of n. When $x = 1$, $y = 1$ in every case

You can easily verify that $y(0) = y_0$ and $y(\infty) = y_\infty$. A plot of the solution is given in Fig. 2.16, along with exponential growth with the same value of b_0.

Another way to think of Eq. 2.28 is that it has the form $dy/dt = b(y)y$, where $b(y) = b_0(1 - y/y_\infty)$ is now a function of the dependent variable y instead of the independent variable t. As y grows toward the asymptotic value, the growth rate $b(y)$ decreases linearly to zero. The logistic model was an early and very important model for population growth. It provides good fits in a few cases, but there are now many more sophisticated models in population biology (Murray 2001) and bacterial growth (Hagen 2010).

2.11 Log–log Plots, Power Laws, and Scaling

This section considers the use of plots in which both scales are logarithmic: log–log plots. They are useful when x and y are related by the power law

$$y = Bx^n. \tag{2.30}$$

Notice the difference between this and the exponential function: here the independent variable x is *raised to a constant power*, while in the exponential case, x (or t) is *in the exponent*. It also leads to a discussion of *scaling*, whereby simple physical arguments lead to important conclusions about the variations between species in size, shape, metabolic rate, and the like.

2.11.1 Log–log Plots and Power Laws

By taking logarithms of both sides of Eq. 2.30, we get

$$\log y = \log B + n \log x. \tag{2.31}$$

This is a linear relationship between $u = \log y$ and $v = \log x$:

$$u = \text{const} + nv. \tag{2.32}$$

Therefore a plot of u vs v is a straight line with slope n. The slope can be positive or negative and need not be an integer. Figure 2.17 shows plots of $y = x$, $y = x^2$, $y = x^{1/2}$, and $y = x^{-1}$. The slope can be determined from the graph by taking $\Delta u/\Delta v$. The value of B is determined either by substituting particular values of y and x in Eq. 2.30 after n is known, or by determining the value of y when $x = 1$, in which case $x^n = 1$ for any value of n, so n need not be known.

Figure 2.18 shows how the curves change when B is changed while $n = 1$. The curves are all parallel to each other. Multiplying by B is equivalent to adding a constant to $\log y$.

If the expression is not of the form $y = Bx^n$ but has an added term, it will not plot as a straight line on log–log paper. Figure 2.18 also shows a plot of $y = x + 1$, which is not a straight line. (Of course, for very large values of x, $\log(x + 1)$ becomes nearly indistinguishable from $\log x$, and the line appears straight.)

When the slope is constant, n can be determined from the slope $\Delta u/\Delta v$ measured with a ruler on the log–log paper. When determining the slope in this way *one must be sure that the length of a cycle is the same in each direction on the*

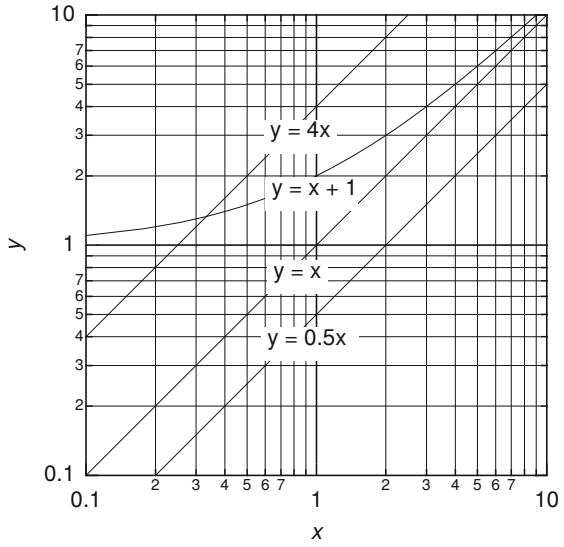


Fig. 2.18 Log–log plots of $y = Bx$, showing how the curves shift on the paper as B changes. Since $n = 1$ for all the curves, they all have the same slope. There is also a plot of $y = x + 1$ to show that a polynomial does not plot as a straight line

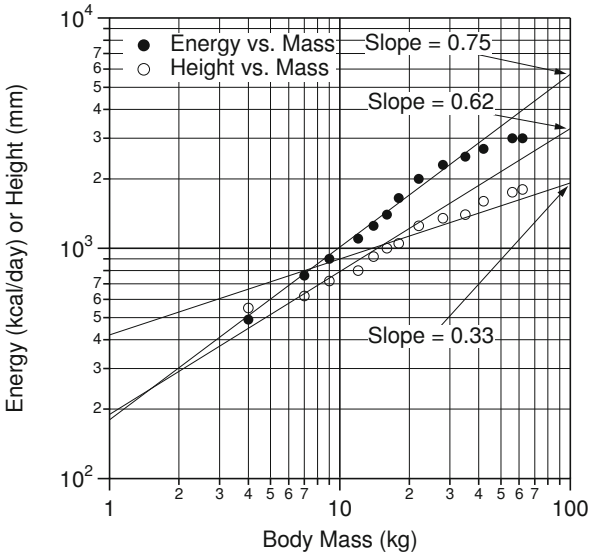


Fig. 2.19 Plot of daily food requirement F and height H vs mass M for growing children. (Data are from Kempe et al. 1970, p. 90)

graph paper. To repeat the warning: it is easy to get a rough idea of the exponent from inspection of the slope of the log–log plot in Fig. 2.17 because on commercial log–log graph paper, the distance spanned by a decade or cycle is the same on both axes. Some magazines routinely show log–log plots in which the distance spanned by a decade is not the same on both axes. Moreover, commercial graphing software does not impose this constraint on log–log plots, so it is becoming less and less likely that you can determine the exponent by glancing at the plot. Be careful!

When using a spreadsheet or other graphing software, it is often useful to make an extra column that contains the calculated variable $y_{\text{calc}} = Ax^m$ with the values for A and m stored in two cells of the spreadsheet. If you plot this column as a line, and your real data as points without a line, then you can change the parameters while inspecting the graph to find the values that give the best fit.

An example of the use of a log–log plot is Poiseuille flow of fluid through a tube vs tube radius when the pressure gradient along the tube is constant (Problem 39). It was shown in Chap. 1 that an r^4 dependence is expected.

2.11.2 Food Consumption, Basal Metabolic Rate, and Scaling

Consider the relation of daily food consumption to body mass. This will introduce us to simple scaling arguments. As a first model, we might suppose that each kilogram of tissue has the same metabolic requirement, so that food consumption should be proportional to body mass. However, there is a problem with this argument. Most of the food that we consume is converted to heat. The various mechanisms to lose heat—radiation, convection, and perspiration—are all roughly proportional to the surface area of the body rather than its mass. (This statement neglects the fact that considerable evaporation takes place through the lungs and that the body can control the rate of heat loss through sweating and shivering.) If all persons were the same shape, then the total surface area would be proportional to H^2, where H is the height. The total volume and mass would be proportional to H^3, so H would be proportional to $M^{1/3}$. Therefore the surface area would be proportional to $(M^{1/3})^2$ or $M^{2/3}$. (See Problem 44 for a discussion of other possible dependences of surface area on mass.) Figure 2.19 plots H and the total daily food requirement F vs body mass M for growing children (Kempe et al. 1970, p. 90).

Neither of the models proposed above fits the data very well. At early ages, H is more nearly proportional to $M^{0.62}$ than to $M^{1/3}$. For older children, when the shape of the body has stopped changing, an $M^{0.33}$ dependence does fit better. This better fit occurs for masses greater than 23 kg, which correspond to ages over 6 years. The slope of the $F(M)$ curve is 0.75. This is less than the 1.0 of the model that food consumption is proportional to the mass and greater than the 0.67 of the model that food consumption is proportional to surface area.

This $\frac{3}{4}$-power dependence is remarkable because it is seen across many species, from one-celled organisms to

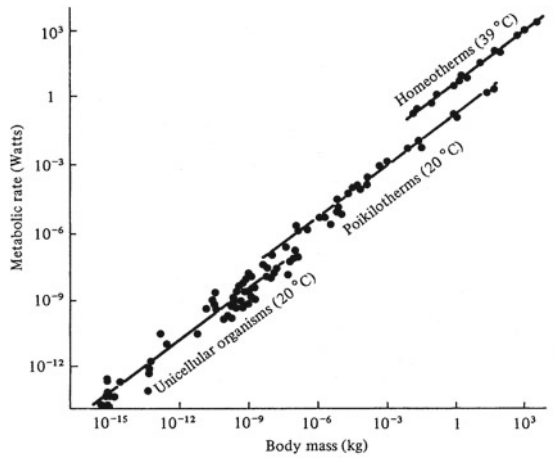


Fig. 2.20 Plot of resting metabolic rate vs. body mass for many different organisms. (Graph is from R. H. Peters 1983. Modified from A. M. Hemmingsen 1960). Used with permission

large mammals. It is called *Kleiber's law*. Peters (1983) quotes work by Hemmingsen (1960) that shows the standard metabolic rates for many species can be fitted by the following. The standard metabolic rate is in watts and mass in kilograms. (*Standard* means as close to resting or basal as possible.) For unicellular organisms at 20 °C,

$$R_{\text{unicellular}} = 0.018M^{0.751}. \tag{2.33a}$$

The range of masses extended from 10^{-15} to 10^{-6} kg. For *poikilotherms* (organisms such as fish whose body temperature is the same as the surroundings) at 20 °C (masses from 10^{-8} to 10^{2} kg),

$$R_{\text{poikilotherm}} = 0.14M^{0.751}, \tag{2.33b}$$

and for *homeotherms* (animals that can maintain their body temperature independent of the surroundings) at 39 °C (masses from 10^{-2} to 10^{3} kg),

$$R_{\text{homeotherm}} = 4.1M^{0.751}. \tag{2.33c}$$

Peters' graph is shown in Fig. 2.20.

A number of models have been proposed to explain a $\frac{3}{4}$-power dependence (McMahon 1973; Peters 1983; West et al. 1999; Banavar et al. 1999). West and his coworkers argue that the $\frac{3}{4}$-power law is universal (Brown et al. 2004; West and Brown 2004). They derive it from a model that supplies nutrients through a branching network that reaches all parts of the organism, minimizes the energy required for distribution, and ends in capillaries (or terminal xylem in plants) that are all the same size. Whether it is universal is still debated (White and Seymour 2003; Glazier 2005).

Symbols Used in Chap. 2

Symbol	Use	Units	First used page
a	Rate of input of a substance	s^{-1}	41
b, b_0	Rate of growth or decay	s^{-1}, h^{-1}	33
c_1, c_2	Constants		36
f	Fraction		35
m, n	Exponent in power–law relationship		43
t	Time	s	34
u	Logarithm of dependent variable		36
v	Logarithm of independent variable		43
x	General independent variable		35
y	General dependent variable		33
y	Amount of substance in plasma	kg, mg	40
x_0, y_0	Initial value of x or y		33
y_∞	Saturation value of y		42
A	Constant		42
B	Constant		43
C	Concentration	$kg\ m^{-3}$, etc.	40
F	Food requirement	$kcal\ day^{-1}$	44
H	Body height	m	44
K	Clearance	$m^3\ s^{-1}$	40
M	Body mass	kg	44
N	Number of compoundings per year		34
Q	Flow through chemostat	$m^3\ s^{-1}$	40
R	Standard metabolic rate	W	45
$T_{1/2}$	Half-life	s, etc.	35
T_2	Doubling time	s	36
V	Volume	m^3	40
X_{10}	Change in x for a factor-of-10 change in y		36
X_{100}	Change in x for a factor-of-100 change in y		37

Problems

Section 2.1

Problem 1. Suppose that you are 20 years old and have an annual income of $20,000. You plan to work for 40 years. If inflation takes place at a rate of 3 % per year, what income would you need at age 60 to have the same buying power you have now? Ignore taxes. Make the calculation assuming that (a) inflation is 3 % and occurs once a year and (b) inflation is continuous but at a 3 % annual rate.

Problem 2. The number e is defined by $\lim_{n\to\infty}(1+1/n)^n$.

(a) Calculate values of $(1+1/n)^n$ for $n = 1, 2, 4, 8$, and 16.

(b) Use the binomial formula $(1+a)^n = 1 + na + \frac{n(n-1)}{2!}a^2 + \frac{n(n-1)(n-2)}{3!}a^3 + \cdots$ to obtain a series for $e^x = \lim_{n\to\infty}(1+x/n)^n$. [See also Appendix D, Eq. D.3.]

Problem 3. A child with acute lymphocytic leukemia (ALL) has approximately 10^{12} leukemic cells when the disease is clinically apparent.

(a) If a cell is about 8 μm in diameter, estimate the total mass of leukemic cells.

(b) Cure requires killing every single cell. The doubling time for the cells is about 5 days. If all cells were killed except for one, how long would it take for the disease to become apparent again?
(c) Suppose that chemotherapy reduces the number of cells to 10^9 and there are no changes of ALL cell properties (no mutations). How long a remission would you expect? What if the number were reduced to 10^6?

Problem 4. Suppose that tumor cells within the body reproduce at rate r, so that the number is given by $y = y_0 e^{rt}$. Each time a chemotherapeutic agent is given, it destroys a fraction f of the cells then existing. Make a semilog plot showing y as a function of time for several administrations of the drug, separated by time T. What different cases must you consider for the relation among f, T, and r?

Problem 5. An exponentially growing culture of bacteria increases from 10^6 to 5×10^8 cells in 6 h. What is the time between successive cell divisions if there is no cell mortality?

Problem 6. The following data on railroad tracks were obtained from R. H. Romer (1991).

Year	Miles of track
1860	30,626
1870	52,922
1880	93,262
1890	166,703

(a) What is the doubling time?
(b) Estimate the surface area of the contiguous USA. Assume that a railroad roadbed is 7-m wide. In what year would an extrapolation predict that the surface of the USA would be completely covered with railroad track?

Section 2.2

Problem 7. A dose D of drug is given that causes the plasma concentration to rise from 0 to C_0. The concentration then falls according to $C = C_0 e^{-bt}$. At time T, what dose must be given to raise the concentration to C_0 again? What will happen if the original dose is administered over and over again at intervals of T?

Problem 8. Consider the atmosphere to be at constant temperature but to have a pressure p that varies with height y. A slab between y and $y + dy$ has a different pressure on the top than on the bottom because of the weight of the air in the slab. (The weight of the air is the number of molecules N times mg, where m is the mass of a molecule and g is the gravitational acceleration.) Use the ideal gas law, $pV = Nk_BT$ (where k_B is the Boltzmann constant and T, the absolute temperature, is constant), and the fact that the air is in equilibrium to write a differential equation for p as a function of y. The equation should be familiar. Show that $p(y) = Ce^{-mgy/k_BT}$.

Problem 9. The mean life of a radioactive substance is defined by the equation

$$\tau = \frac{-\int_0^\infty t\,(dy/dt)\,dt}{-\int_0^\infty (dy/dt)\,dt}.$$

Show that if $y = y_0 e^{-bt}$, then $\tau = 1/b$.

Section 2.3

Problem 10. R. Guttman (1966) measured the temperature dependence of the current pulse necessary to excite the squid axon. She found that for pulses shorter than a certain length τ, a fixed amount of electric charge was necessary to make the nerve fire; for longer pulses, the current was fixed. This suggests that the axon integrates the current for a time τ but no longer. The following data are for the integrating time τ vs temperature T (°C). Find an empirical exponential relationship between T and τ.

T (°C)	τ (ms)
5	4.1
10	3.4
15	1.9
20	1.4
25	0.7
30	0.6
35	0.4

Problem 11. A normal rabbit was injected with 1 cm^3 of *Staphylococcus aureus* culture containing 10^8 organisms. At various later times, 0.2 cm^3 of blood was taken from the rabbit's ear. The number of organisms per cm^3 was calculated by diluting the material, smearing it on culture plates, and counting the number of colonies formed. The results are shown below. Plot these data and see if they can be fit by a single exponential. Can you also estimate the blood volume of the rabbit?

t (min)	Bacteria (cm^{-3})
0	5×10^5
3	2×10^5
6	5×10^4
10	7×10^3
20	3×10^2
30	1.7×10^2

Section 2.4

Problem 12. All members of a certain population are born at $t = 0$. The death rate in this population (deaths per unit

population per unit time) is found to increase linearly with age t: (death rate) $= a+bt$. Find the population as a function of time if the initial population is y_0.

Problem 13. The accompanying table gives death rates (in yr^{-1}) as a function of age. Plot these data on linear graph paper and on semilog paper. Find a region over which the death rate rises approximately exponentially with age, and determine parameters to describe that region.

Age	Death rate	Age	Death rate
0	0.000 863	45	0.005 776
5	0.000 421	50	0.008 986
10	0.000 147	55	0.013 748
15	0.001 027	60	0.020 281
20	0.001 341	65	0.030 705
25	0.001 368	70	0.046 031
30	0.001 697	75	0.066 196
35	0.002 467	80	0.101 443
40	0.003 702	85	0.194 197

Problem 14. Suppose that the amount of a resource at time t is $y(t)$. At $t = 0$, the amount is y_0. The rate at which it is consumed is $r = -dy/dt$. Let $r = r_0 e^{bt}$, that is, the rate of use increases exponentially with time. (For example, until recently the world use of crude oil had been increasing about 7 % per year since 1890.)

(a) Show that the amount remaining at time t is $y(t) = y_0 - (r_0/b)(e^{bt} - 1)$.

(b) If the present supply of the resource were used up at constant rate r_0, it would last for a time T_c. Show that when the rate of consumption grows exponentially at rate b, the resource lasts a time $T_b = (1/b)\ln(1 + bT_c)$.

(c) An advertisement in *Scientific American*, September 1978, p. 181, said, "There's still twice as much gas underground as we've used in the past 50 years—at our present rate of use, that's enough to last about 60 years." Calculate how long the gas would last if it were used at a rate that increases 7 % per year.

(d) If the supply of gas were doubled, how would the answer to part (c) change?

(e) Repeat parts (c) and (d) if the growth rate is 3 % per year.

Problem 15. When we are dealing with death or component failure, we often write Eq. 2.17 in the form $y(t) = y_0 \exp\left[-\int_0^t m(t')dt'\right]$ and call $m(t)$ the *mortality function*. Various forms for the mortality function can represent failure of computer components, batteries in pacemakers, or the death of organisms. (This is not the most general possible mortality model. For example, it ignores any interaction between organisms, so it cannot account for effects such as overcrowding or a limited supply of nutrients.)

(a) For human populations, the mortality function is often written as $m(t) = m_1 e^{-b_1 t} + m_2 + m_3 e^{+b_3 t}$. What sort of processes does each of these terms represent?

(b) Assume that m_1 and m_2 are zero. Then $m(t)$ is called the *Gompertz mortality function*. Obtain an expression for $y(t)$ with the Gompertz mortality function. Time $t_{\max}$ is sometimes defined to be the time when $y(t) = 1$. It depends on y_0. Obtain an expression for $t_{\max}$.

Problem 16. The *incidence* of a disease is the number of new cases per unit time per unit population (or per 100,000). The *prevalence* of the disease is the number of cases per unit population. For each situation below, the size of the general population remains fixed at the constant value y, and the disease has been present for many years.

(a) The incidence of the disease is a constant, i cases per year. Each person has the disease for a fixed time of T years, after which the person is either cured or dies. What is the prevalence p? Hint: the number who are sick at time t is the total number who became sick between $t - T$ and t.

(b) The patients in part (a) who are sick die with a constant death rate b. What is the prevalence?

(c) A new epidemic begins at $t = 0$, and the incidence increases exponentially with time: $i = i_0 e^{kt}$. What is the prevalence if each person has the disease for T years?

Section 2.5

Problem 17. The creatinine clearance test measures a patient's kidney function. Creatinine is produced by muscle at a rate p g h^{-1}. The concentration in the blood is C g l^{-1}. The volume of urine collected in time T (usually 24 h) is V l. The creatinine concentration in the urine is U g l^{-1}. The clearance is K. The plasma volume is V_p. Assume that creatinine is stored only in the plasma.

(a) Draw a block diagram for the process and write a differential equation for C.

(b) Find an expression for the creatinine clearance K in terms of p and C when C is not changing with time.

(c) If C is constant, all creatinine produced in time T appears in the urine. Find K in terms of C, V, U, and T.

(d) If p were somehow doubled, what would be the new steady-state value of C? What would be the time constant for change to the new value?

Problem 18. A liquid is injected in muscle and spreads throughout a spherical volume $V = 4\pi r^3/3$. The volume is well supplied with blood, so that the liquid is removed at a rate proportional to the remaining mass per unit volume. Let the mass be m and assume that r remains fixed. Find a differential equation for $m(t)$ and show that m decays exponentially.

Problem 19. A liquid is injected as in Problem 18, but this time a cyst is formed. The rate of removal of mass is proportional to both the pressure of liquid within the cyst, and

to the surface area of the cyst, which is $4\pi r^2$. Assume that the cyst shrinks so that the pressure of liquid within the cyst remains constant. Find a differential equation for the rate of mass removal and show that dm/dt is proportional to $m^{2/3}$.

Problem 20. The following data showing ethanol concentration in the blood vs time after ethanol ingestion are from Bennison and Li (1976, pp. 9–13). Plot the data and discuss the process by which alcohol is metabolized.

t (min)	Ethanol concentration (mg dl^{-1})
90	134
120	120
150	106
180	93
210	79
240	65
270	50

Problem 21. Consider the following two-compartment model. Compartment 1 is damaged myocardium (heart muscle). Compartment 2 is the blood of volume V. At $t = 0$, the patient has a heart attack and compartment 1 is created. It contains q molecules of some chemical that was released by the dead cells. Over the next several days, the chemical moves from compartment 1 to compartment 2 at a rate $i(t)$, such that $q = \int_0^\infty i(t)dt$. The amount of substance in compartment 2 is $y(t)$ and the concentration is $C(t)$. The only mode of removal from compartment 2 is clearance with clearance constant K.

(a) Write a differential equation for $C(t)$ that may also involve $i(t)$.
(b) Integrate the equation and show that q can be determined by numerical integration if $C(t)$ and K are known.
(c) Show that volume V need not be known if $C(0) = C(\infty)$.

Section 2.7

Problem 22. The radioactive nucleus ^{64}Cu decays independently by three different paths. The relative decay rates of these three modes are in the ratio 2:2:1. The half-life is 12.8 h. Calculate the total decay rate b, and the three partial decay rates b_1, b_2, and b_3.

Problem 23. The following data were taken from Berg et al. (1982). At $t = 0$, a 70-kg subject was given an intravenous injection of 200 mg of phenobarbital. The initial concentration in the blood was 6 mg l^{-1}. The concentration decayed exponentially with a half-life of 110 h. The experiment was repeated, but this time the subject was fed 200 g of activated charcoal every 6 h. The concentration of phenobarbital again fell exponentially, but with a half-life of 45 h.

(a) What was the volume in which the phenobarbital was distributed?
(b) What was the clearance in the first experiment?
(c) What was the clearance due to charcoal?

Section 2.8

Problem 24. You are treating a severely ill patient with an intravenous antibiotic. You give a loading dose D mg, which distributes immediately through blood volume V to give a concentration C mg dl^{-1} (1 dl = 0.1 l). The half-life of this antibiotic in the blood is T h. If you are giving an intravenous glucose solution at a rate R ml h^{-1}, what concentration of antibiotic should be in the glucose solution to maintain the concentration in the blood at the desired value?

Problem 25. The solution to the differential equation $dy/dt = a - by$ for the initial condition $y(0) = 0$ is $y = (a/b)(1 - e^{-bt})$. Plot the solution for $a = 5$ g min^{-1} and for $b = 0.1$, 0.5, and 1.0 min^{-1}. Discuss why the final value and the time to reach the final value change as they do. Also make a plot for $b = 0.1$ and $a = 10$ to see how that changes the situation.

Problem 26. Derive an approximate expression for $(a/b)\left(1 - e^{-bt}\right)$ which is accurate for small times ($t \ll 1/b$). Use the Taylor expansion for an exponential given in Appendix D.

Problem 27. We can model the repayment of a mortgage with a differential equation. Suppose that $y(t)$ is the amount still owed on the mortgage at time t, the rate of repayment per unit time is a, b is the interest rate, and the initial amount of the mortgage is y_0.

(a) Find the differential equation for $y(t)$.
(b) Try a solution of the form $y(t) = a/b + Ce^{bt}$, where C is a constant to be determined from the initial conditions. Find C, plot the solution, and determine the time required to pay off the mortgage.

Problem 28. When an animal of mass m falls in air, two forces act on it: gravity, mg, and a force due to air friction. Assume that the frictional force is proportional to the speed v.

(a) Write a differential equation for v based on Newton's second law, $F = m(dv/dt)$.
(b) Solve this differential equation (hint: compare your equation to Eq. 2.25).
(c) Assume that the animal is spherical, with radius a and density ρ. Also, assume that the frictional force is proportional to the surface area of the animal. Determine the terminal speed (speed of descent in steady state) as a function of a.
(d) Use your result in part (c) to interpret the following quote by J. B. S. Haldane (1985): "You can drop a mouse down a thousand-yard mine shaft; and arriving at the bottom, it gets a slight shock and walks away. A rat is killed, a man is broken, a horse splashes."

Problem 29. In Problem 28, we assumed that the force of air friction is proportional to the speed v. For flow at high Reynolds numbers, a better approximation is that the force is $-kv^2$.

(a) Write the differential equation for v as a function of t.

(b) This differential equation is nonlinear because of the v^2 term and thus difficult to solve analytically. However, the terminal speed can easily be obtained directly from the differential equation by setting $dv/dt = 0$. Find the terminal speed as a function of a (defined in Problem 28).

(c) Verify that $v(t) = \sqrt{mg/k}\tanh\left(\sqrt{kg/m}t\right)$ is a solution.

Problem 30. A drug is infused into the body through an intravenous drip at a rate of $100\,\mathrm{mg\,h^{-1}}$. The total amount of drug in the body is y. The drug distributes uniformly and instantaneously throughout the body in a compartment of volume $V = 18\,\mathrm{l}$. It is cleared from the body by a single exponential process. In the steady state, the total amount in the body is 200 mg.

(a) At noon ($t = 0$), the intravenous line is removed. What is $y(t)$ for $t > 0$?

(b) What is the clearance of the drug?

Section 2.9

Problem 31. You are given the following data:

x	y	x	y
0	1.000	5	0.444
1	0.800	6	0.400
2	0.667	7	0.364
3	0.571	8	0.333
4	0.500	9	0.308
		10	0.286

Plot these data on semilog graph paper. Is this a single exponential? Is it two exponentials? Plot $1/y$ vs x. Does this alter your answer?

Problem 32. Cells can repair DNA damage caused by x-ray exposure (see Sect. 16.9). Wang et al. (2001) found that the amount of damage is characterized by two time constants. Assume the DNA damage, D, as a function of time, t, is given by the following data

t (h)	$D(\%)$	t (h)	$D(\%)$
0	100	1.5	16
0.25	46	2	14
0.50	28	4	9.0
0.75	21	6	5.8
1.0	18	8	3.7

Plot the data on semilog paper. Fit the data to Eq. 2.27 by eye or using a spreadsheet and determine A_1, A_2, b_1, and b_2. Note that the data are normalized to 100 % at $t = 0$. What does this mean in terms of A_1 and A_2?

Section 2.10

Problem 33. Suppose that the rate of consumption of a resource increases exponentially. (This might be petroleum, or the nutrient in a bacterial culture.) During the first doubling time, the amount used is 1 unit. During the second doubling time, it is 2 units, the next 4, etc. How does the amount consumed during a doubling time compare to the total amount consumed during all previous doubling times?

Problem 34. Suppose that the rate of growth of y is described by $dy/dt = b(y)y$. Expand $b(y)$ in a Taylor's series and relate the coefficients to the terms in the logistic equation.

Problem 35. Verify that the solution $y(t)$ in Eq. 2.29 obeys the differential Eq. 2.28.

Problem 36. In the logistic model (Eq. 2.28), what value of y corresponds to the maximum rate of change of y?

Problem 37. The consumption of a finite resource is often modeled using the logistic equation. Let $y(t)$ be the cumulative amount of a resource consumed and y_∞ be the total amount that was initially available at $t = -\infty$. Model the rate of consumption using Eq. 2.29 over the range $-\infty < t < \infty$.

(a) Set $y_0 = y_\infty/2$, so that the zero of the time axis correponds to when half the resource has been used. Show that this simplifies Eq. 2.29.

(b) Differentiate $y(t)$ to find an expression for the rate of consumption. Sketch plots of dy/dt vs t on linear and semilog graph paper. When does the peak rate of consumption occur?

When this model is applied to world oil consumption, the maximum is called *Hubbert's peak* (Deffeyes 2008).

Problem 38. Consider a classic predator–prey problem. Let the number of foxes be F and the number of rabbits be R. The rabbits eat grass, which is plentiful. The foxes eat only rabbits. The number of foxes and rabbits can be modeled by the *Lotka–Volterra equations*

$$\frac{dR}{dt} = aR - bRF$$
$$\frac{dF}{dt} = -cF + dRF.$$

(a) Describe the physical meaning of each term on the right-hand side of each equation. What does each of the constants a, b, c, and d denote?

(b) Solve for the steady-state values of F and R.

These differential equations are difficult to solve because they are nonlinear (see Chap. 10). Typically, R and F oscillate about the steady-state solutions found in (b). For more information, see Murray (2001).

Section 2.11

Problem 39. Plot the following data for Poiseuille flow on log–log graph paper. Fit the equation $i = CR_p^n$ to the data by eye (or by trial and error using a spread sheet), and determine C and n.

R_p (μm)	i (μm^{3}s^{-1})
5	0.000 10
7	0.000 38
10	0.001 6
15	0.008 1
20	0.026
30	0.13
50	1.0

Problem 40. Below are the molecular weights and radii of some molecules. Use log–log graph paper to develop an empirical relationship between them.

Substance	M	R (nm)
Water	18	0.15
Oxygen	32	0.20
Glucose	180	0.39
Mannitol	180	0.36
Sucrose	390	0.48
Raffinose	580	0.56
Inulin	5 000	1.25
Ribonuclease	13,500	1.8
β-lactoglobin	35,000	2.7
Hemoglobin	68,000	3.1
Albumin	68,000	3.7
Catalase	250,000	5.2

Problem 41. How well does Eq. 2.33c explain the data of Fig. 2.19? Discuss any differences.

Problem 42. Compare the mass and metabolic requirements (and hence waste output, including water vapor) of 180 people each weighing 70 kg with 12,600 chickens of average mass 1 kg.

Problem 43. Figure 2.19 shows that in young children, height is more nearly proportional to $M^{0.62}$ than to $M^{1/3}$. Find pictures of children and adults and compare ratios of height to width, to see what the differences are.

Problem 44. Consider three models of an organism. The first is a sphere of radius R. The second is a cube of length L. These are crude models for animals. The third is a broad leaf of surface area A on each side and thickness t. Assume all have density ρ. In each case, calculate the surface area S as a function of mass, M. Ignore the surface area of the edge of the leaf. (For a comparison of scaling in leaves and animals, see Reich (2001). He shows that for broad leaves, $S \propto M^{1.1}$.)

Problem 45. If food consumption is proportional to $M^{3/4}$ across species, how does the food consumption per unit mass scale with mass? Qualitatively compare the eating habits of hummingbirds to eagles and mice to elephants. (See Schmidt-Nielsen 1984, pp. 62–64.)

Problem 46. In Problem 45, you found how the specific metabolic rate (food consumption per unit mass) varies with mass. If all animal heart volumes and blood volumes are proportional to M, then the only way for the heart to increase the oxygen delivery to the body is by increasing the frequency of the heart rate (Schmidt-Nielsen 1984, pp. 126–150).

(a) Using the result from Problem 45, if a 70 kg man has a heart rate of 80 beats min^{-1}, determine the heart rate of a guinea pig ($M = 0.5$ kg).

(b) To a first approximation, all hearts beat about 800,000,000 times in a lifetime. A 30-g mouse lives about 3 years. Estimate the life span of a 3000-kg elephant.

(c) Humans live longer than what their mass would indicate. Calculate the life span of a 70-kg human based on scaling, and compare it to a typical human life span.

Problem 47. Let us examine how high animals can jump (Schmidt-Nielsen 1984, pp. 176–179). Assume that the energy output of the jumping muscle is proportional to the body mass, M. The gravitational potential energy gained upon jumping to a height h is Mgh ($g = 9.8\,\mathrm{m\,s^{-2}}$). If a 3-g locust can jump 60 cm, how high can a 70-kg human jump? Use scaling arguments.

Problem 48. In Problem 47, you should have found that all animals can jump to about the same height (approximately 0.6 m), independent of their mass M.

(a) Equate the kinetic energy at the bottom of the jump ($Mv^2/2$, where v is the"take-off speed") to the potential energy Mgh at the top of the jump to find how the take-off speed scales with mass.

(b) Calculate the take-off speed.

(c) In order to reach this speed, the animal must accelerate upward over a distance L. If we assume a constant acceleration a, then $a = v^2/(2L)$. Assume L scales as the linear size of the animal (and assume all animals are basically the same shape but different size). How does the acceleration scale with mass?

(d) For a 70-kg human, L is about 1/3 m. Calculate the acceleration (express your answer in terms of g).

(e) Use your result from part (c) to estimate the acceleration for a 0.5-mg flea (again, express your answer in terms of g).

(f) Speculate on the biological significance of the result in part (e) (See Schmidt-Nielsen 1984, pp. 180–181).

References

Banavar JR, Maritan A, Rinaldo A (1999) Size and form in efficient transportation networks. Nature 399:130–132

Bartlett A (2004) The essential exponential! For the future of our planet. Center for Science, Mathematics & Computer Education, Lincoln

Bennison LJ, Li TK (1976) Alcohol metabolism in American Indians and whites—lack of racial differences in metabolic rate and liver alcohol dehydrogenase. N Engl J Med 294:9–13

Berg MJ, Berlinger WG, Goldberg MJ, Spector R, Johnson GF (1982) Acceleration of the body clearance of phenobarbital by oral activated charcoal. N Engl J Med 307:642–644

Bland EF, White PD (1941) Coronary thrombosis (with myocardial infarction) ten years later. J Am Med Assoc 114(7):1171–1173

Brown JH, Gillooly JF, Allen AP, Savage VM, West GB (2004) Toward a metabolic theory of ecology. Ecology 85:1771–1789

Clark VA (1975) Survival distributions. Ann Rev Biophys Bioeng 4:431–438

Deffeyes KS (2008) Hubbert's peak: the impending world oil shortage. Princeton University Press, Princeton

Glazier DS (2005) Beyond the '3/4-power law': Variation in the intra- and interspecific scaling of metabolic rate in animals. Biol Rev 80:611–662

Guttman R (1966) Temperature characteristics of excitation in space-clamped squid axons. J Gen Physiol 49:1007–1018. doi:10.1085/jgp.49.5.1007

Hagen SJ (2010) Exponential growth of bacteria: constant multiplication through division. Am J Phys 78(12):1290–1296

Haldane JBS (1985) On being the right size and other essays. Oxford University Press, Oxford

Hemmingsen AM (1960) Energy metabolism as related to body size and respiratory surfaces, and its evolution. Rep Steno Meml Hosp Nordisk Insul Lab 9:6–110

Kempe CH, Silver HK, O'Brien D (1970) Current pediatric diagnosis and treatment, 2nd edn. Lange, Los Altos

Levy D, Brink S (2005) A change of heart: how the people of Framingham, Massachusetts helped unravel the mysteries of cardiovascular disease. Knopf, New York

Maor E (1994) e, The story of a number. Princeton University Press, Princeton

McKee PA, Castelli WP, McNamara PM, Kannel WB (1971) The natural history of congestive heart failure: the Framingham study. N Engl J Med 285:1441–1446

McMahon T (1973) Size and shape in biology. Science 179:1201–1204

Murray JD (2001) Mathematical biology. Springer, New York

Peters RH (1983) The ecological implications of body size. Cambridge University Press, Cambridge

Reich P (2001) Body size, geometry, longevity and metabolism: do plant leaves behave like animal bodies? Trends Ecol Evol 16(12):674–680

Romer RH (1991) The mathematics of exponential growth—keep it simple. Phys Teach 9:344–345

Riggs DS (1970) The mathematical approach to physiological problems. MIT Press, Cambridge

Schmidt-Nielsen K (1984) Scaling: why is animal size so important? Cambridge University Press, Cambridge

Wang H, Zeng Z-C, Bui T-A, Sonoda E, Takata M, Takeda S, Iliakis G (2001) Efficient rejoining of radiation-induced DNA double-strand breaks in vertebrate cells deficient of genes of the RAD52 epistasis group. Oncogene 20:2212–2224

West GB, Brown JH, Enquist BJ (1999) The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 284:1677–1679. (see also Mackenzie's accompanying editorial on page 1607 of the same issue of Science)

West GB, Brown JH (2004). Life's universal scaling laws. Phys Today 57(9):36–42

White CR, Seymour RS (2003) Mammalian basal metabolic rate is proportional to body mass$^{2/3}$. Proc Natl Acad Sci U S A 100(7):4046–4049

Zumoff B, Hart H, Hellman L (1966) Considerations of mortality in certain chronic diseases. Ann Intern Med 64:595–601

Systems of Many Particles

3

It is possible to identify all the external forces acting on a simple system and use Newton's second law ($\mathbf{F} = m\mathbf{a}$) to calculate how the system moves. (Applying this technique in a complicated case such as the femur may require the development of a simplified model, because so many muscles, other bones, and ligaments apply forces at so many different points.) In an atomic-size system consisting of a single atom or molecule, it is possible to use the quantum-mechanical equivalent of $\mathbf{F} = m\mathbf{a}$, the Schrödinger equation, to do the same thing. (The Schrödinger equation takes into account the wave properties that are important in small systems.)

In systems of many particles, such calculations become impossible. Consider, for example, how many particles there are in a cubic millimeter of blood. Table 3.1 shows some of the constituents of such a sample. To calculate the translational motion in three dimensions, it would be necessary to write three equations for each particle[1] using Newton's second law. Suppose that at time t the force on a molecule is $\mathbf{F}$. Between t and $t + \Delta t$, the velocity of the particle changes according to the three equations

$$v_i(t + \Delta t) = v_i(t) + F_i \Delta t/m, \ (i = x, y, z).$$

The three equations for the change of position of the particle are of the form $x(t + \Delta t) = x(t) + v_x(t)\Delta t + F_x(t)(\Delta t)^2/(2m)$. If Δt is small enough the last term can be neglected. Solving these equations requires at least six multiplications and additions for each particle. For 10^{19} particles, this means about 10^{20} arithmetic operations per time interval. If a computer can do 10^{12} operations/s, then the complete calculation for a single time interval will require 10^8 s or 3 years!

Another limitation arises in the physics of the processes. Relatively simple systems can exhibit deterministic chaos: a collection of identical systems differing in their initial conditions by an infinitesimally small amount can become completely different in their subsequent behavior in a surprisingly short period of time. It is impossible to trace the behavior of this many molecules on an individual basis.

Table 3.1 Some constituents of 1 mm^3 of blood

Constituent	Concentration in customary units	Number in 1 mm^3
Water	1 g cm^{-3}	3.3×10^{19}
Sodium	3.2 mg cm^{-3}	8.3×10^{16}
Albumin	4.5 g dl^{-1}	3.9×10^{14}
Cholesterol	200 mg dl^{-1}	3.1×10^{15}
Glucose	100 mg dl^{-1}	3.3×10^{15}
Hemoglobin	15 g dl^{-1}	1.4×10^{15}
Erythrocytes	5×10^6 mm^{-3}	5×10^6

Nor is it necessary. We do not care which water molecule is where. The properties of a system that are of interest are averages over many molecules: pressure, concentration, average speed, and so forth. These average macroscopic properties are studied in *statistical* or *thermal physics* or *statistical mechanics*.

Unfortunately, this chapter relies heavily on your ability to accept delayed gratification. It has only a few biological examples, but the material developed here is necessary for understanding some topics in most of the later chapters, especially Chaps. 4–9 and 14–18. In addition to developing a statistical understanding of pressure, temperature, and concentration, this chapter derives four quantities or concepts that are used later:

1. The *Boltzmann factor*, which tells how concentrations of particles vary with potential energy (Sect. 3.7).
2. The *principle of equipartition of energy*, which underlies the diffusion process that is so important in the body (Sect. 3.10).
3. The *chemical potential*, which describes the condition for equilibrium of two systems for the exchange of particles, and how the particles flow when the systems are not in equilibrium (Sects. 3.12, 3.13, and 3.18).

[1] In computational biology, a mole of differential equations is sometimes called a leibniz (Huang and Wikswo 2006). Solving for the motion of each water molecule in a cubic millimeter of blood requires solving 0.16 millileibniz of equations.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_3, © Springer International Publishing Switzerland 2015

4. The *Gibbs free energy*, which tells the direction in which a chemical reaction proceeds and allows us to understand how the cells in the body use energy (Sect. 3.17).

The first six sections form the basis for the rest of the chapter, developing the concepts of microstates, heat flow, temperature, and entropy. Sections 3.7 and 3.8 develop the Boltzmann factor and its corollary, the Nernst equation. Section 3.9 applies the Boltzmann factor to the air molecules in the atmosphere. Section 3.10 discusses the very important equipartition of energy theorem. Section 3.11 discusses heat capacity—the energy required to increase the temperature of a system.

The transport of particles between two systems is described most efficiently using the chemical potential. The chemical potential is introduced in Sect. 3.12, and an example of its use is shown in Sect. 3.13.

Section 3.14 considers systems that can exchange volume. An idealized example is two systems separated by a flexible membrane or a movable piston. The next two sections extend the idea of systems that exchange energy, particles, or volume to the exchange of other variables such as electric charge.

The Gibbs free energy, introduced in Sect. 3.17, is used to describe chemical reactions that take place at constant temperature and pressure. It is closely related to the chemical potential. The chemical potential of an ideal solution is derived in Sect. 3.18 and is used extensively in Chap. 5.[2]

3.1 Gas Molecules in a Box

Statistical physics or *statistical mechanics* deals with average quantities such as pressure, temperature, and particle concentration and with probability distributions of variables such as velocity. Some of the properties of these averages can be illustrated by considering a simple example: the number of particles in each half of a box containing a fixed number of gas molecules. (This is a simple analog for the concentration.) We will not be concerned with the position and velocity of each molecule, since we have already decided not to use Newtonian mechanics. Nor will we ask for the velocity distribution at this time. This simplified example will describe only how many molecules are in the volume of interest. The number will fluctuate with time. We will deal with probabilities:[3] if the number of particles in the volume is measured repeatedly, what values are obtained, and with what relative frequency?

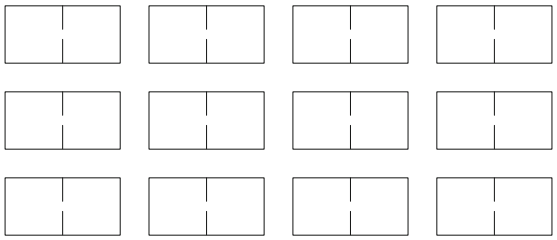


Fig. 3.1 An ensemble of boxes, each divided in half by an imaginary partition

If we were willing to use Newtonian mechanics, we could count periodically how many molecules are in the volume of interest. (This has actually been done for small numbers of particles. See Reif (1964), pp. 8–9.) For larger numbers of particles, it is easier to use statistical arguments to obtain the probabilities. The particles travel back and forth, colliding with the walls of the box and occasionally with one another. After some time has elapsed, all memory of the particles' original positions and velocities has been lost because of collisions with the walls of the box, which have microscopic inhomogeneities. Therefore, the result can be obtained by imagining a whole succession of completely different boxes, in which the particles have been placed at random. We can count the number of molecules in the volume of interest in each box. Such a collection of similar boxes is called an *ensemble*. Ensembles of similar systems will be central to the ideas of this chapter.

Imagine an ensemble of boxes, each divided in half as in Fig. 3.1. We want to know how often a certain number of particles is found in the left half. If one particle is in a box ($N = 1$), two cases can be distinguished, depending on which half the particle is in. Call them L and R. Each case is equally likely to occur, since nothing distinguishes one half of a box from the other. If n is the number of particles in the left half, then case L corresponds to $n = 1$ and case R corresponds to $n = 0$.

The probability of having a particular value of n is defined to be

$$P(n) = \frac{(\text{number of systems in the ensemble in which } n \text{ is found})}{(\text{total number of systems})} \tag{3.1}$$

in the limit as the number of systems becomes very large.

As there are only two possible values of n, 0 or 1, and because each corresponds to one of the equally likely configurations, $P(0) = 0.5$, $P(1) = 0.5$. The sum of the probabilities is 1. A histogram of $P(n)$ for $N = 1$ is given in Fig. 3.2a. To recapitulate: n is the number of molecules in the left half of the box, and N is the total number of molecules in the entire box. Since N will change in the discussion below,

[2] Many excellent introductory textbooks on thermodynamics and statistical mechanics exist, such as those by Reif (1964) and Schroeder (2000). To learn more about how thermodynamics is applied to biological problems, see Haynie (2008).

[3] A good book on probability is Weaver (1963).

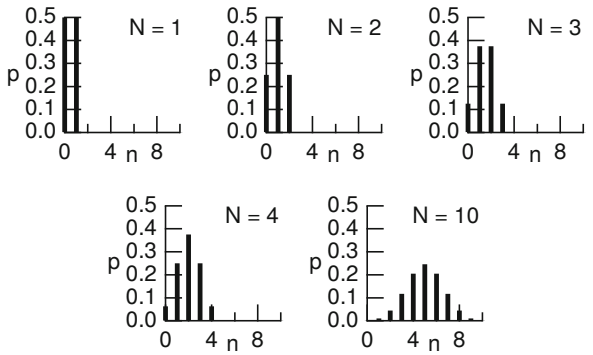


Fig. 3.2 Histograms of $P(n; N)$ for different values of N

we will call the probability $P(n; N)$. (The fixed parameters that determine the probability distribution are located after the semicolon.)

Now let $N = 2$. Each molecule can be on the left or the right with equal probability. The possible outcomes are listed in the following table, along with the corresponding values of n and $P(n; 2)$.

Molecule 1	Molecule 2	n	$P(n; 2)$
R	R	0	$\frac{1}{4}$
R	L	1	$\frac{1}{2}$
L	R	1	
L	L	2	$\frac{1}{4}$

Each of the four outcomes is equally probable. To see this, note that L or R is equally likely for each molecule. In half of the boxes in the ensemble, the first molecule is found on the left. In half of these, the second molecule is also on the left. Therefore LL occurs in one-fourth of the systems in the ensemble. (This is not strictly true, because there can be fluctuations. If we throw a coin six times, we cannot say that heads will always occur three times. If we repeat the experiment many times, the average number of heads will be three.)

If three molecules are placed in each box, there are two possible locations for the first particle, two for the second, and two for the third. If the three particles are all independent, then there are $2^3 = 8$ different ways to locate the particles in a box. If a box is divided in half, each of these ways has a probability of 1/8.

Molecule 1	Molecule 2	Molecule 3	n	$P(n; 3)$
R	R	R	0	$\frac{1}{8}$
R	R	L	1	
R	L	R	1	$\frac{3}{8}$
L	R	R	1	
L	L	R	2	
L	R	L	2	$\frac{3}{8}$
R	L	L	2	
L	L	L	3	$\frac{1}{8}$

The cases of two and three molecules in the box are also plotted in Fig. 3.2.

In each case, $P(n; N)$ has been determined by listing all the ways that the N particles can go into a box. This can become tedious if the number of particles is large. Furthermore, it does not provide a way to calculate P if the two volumes of the box are not equal. We will now introduce a more general technique that can be used for any number of particles and for any fractional volume of the box.

Each box is divided into two volumes, v and v', with total volume $V = v + v'$. Call p the probability that a single particle is in volume v. The probability that the particle is in the remainder of the box, v', is q:

$$p + q = 1. \tag{3.2}$$

As long as there is nothing to distinguish one part of a box from the other, p is the ratio of v to the total volume:

$$p = \frac{v}{V}. \tag{3.3}$$

By the same argument, $q = v'/V$. These values satisfy Eq. 3.2. If N particles are distributed between the two volumes of the box, the number in v is n and the number in v' is $n' = N - n$. The probability that n of the N particles are found in volume v is given by the *binomial probability distribution* (Appendix H):

$$P(n; N) = P(n; N, p) = \frac{N!}{n!\,(N-n)!} p^n (1-p)^{N-n}. \tag{3.4}$$

Table 3.2 shows the calculation of $P(n; 10)$ using this equation. Histograms for $N = 4$ and 10 are also plotted in Fig. 3.2. In each case there is a value of n for which P is a maximum. When N is even, this value is $N/2$; when N is odd, the values on either side of $N/2$ share the maximum value. The probability is significantly different from zero only for a few values of n on either side of the maximum.

A probability distribution, in the form of an expression, a table of values, or a histogram, usually gives all the information that is needed about the number of molecules in v; it is not necessary to ask which molecules are in v. The number of molecules in v is not fixed but fluctuates about the number for which P is a maximum. For example, if $N = 10$, and we measure the number of molecules in the left half many times, we find $n = 5$ only about 25 % of the time. On the other hand, we find that $n = 4$, 5, or 6 about 65 % of the time, while $n = 3$, 4, 5, 6, or 7 about 90 % of the time.

Table 3.2 Calculation of $P(n; 10)$ using the binomial probability distribution. Note that $0! = 1$

$$P(0;10) = \frac{10!}{0!10!}\left(\tfrac{1}{2}\right)^0\left(\tfrac{1}{2}\right)^{10} = \left(\tfrac{1}{2}\right)^{10} = 0.001$$
$$P(1;10) = \frac{10!}{1!9!}\left(\tfrac{1}{2}\right)^1\left(\tfrac{1}{2}\right)^9 = 10\left(\tfrac{1}{2}\right)^{10} = 0.010$$
$$P(2;10) = \frac{10!}{2!8!}\left(\tfrac{1}{2}\right)^2\left(\tfrac{1}{2}\right)^8 = 45\left(\tfrac{1}{2}\right)^{10} = 0.044$$
$$P(3;10) = \frac{10!}{3!7!}\left(\tfrac{1}{2}\right)^3\left(\tfrac{1}{2}\right)^7 = 120\left(\tfrac{1}{2}\right)^{10} = 0.117$$
$$P(4;10) = \frac{10!}{4!6!}\left(\tfrac{1}{2}\right)^4\left(\tfrac{1}{2}\right)^6 = 210\left(\tfrac{1}{2}\right)^{10} = 0.205$$
$$P(5;10) = \frac{10!}{5!5!}\left(\tfrac{1}{2}\right)^5\left(\tfrac{1}{2}\right)^5 = 252\left(\tfrac{1}{2}\right)^{10} = 0.246$$
$$P(6;10) = \frac{10!}{6!4!}\left(\tfrac{1}{2}\right)^6\left(\tfrac{1}{2}\right)^4 = 210\left(\tfrac{1}{2}\right)^{10} = 0.205$$
$$P(7;10) = \frac{10!}{7!3!}\left(\tfrac{1}{2}\right)^7\left(\tfrac{1}{2}\right)^3 = 120\left(\tfrac{1}{2}\right)^{10} = 0.117$$
$$P(8;10) = \frac{10!}{8!2!}\left(\tfrac{1}{2}\right)^8\left(\tfrac{1}{2}\right)^2 = 45\left(\tfrac{1}{2}\right)^{10} = 0.044$$
$$P(9;10) = \frac{10!}{9!1!}\left(\tfrac{1}{2}\right)^9\left(\tfrac{1}{2}\right)^1 = 10\left(\tfrac{1}{2}\right)^{10} = 0.010$$
$$P(10;10) = \frac{10!}{10!0!}\left(\tfrac{1}{2}\right)^{10}\left(\tfrac{1}{2}\right)^0 = \left(\tfrac{1}{2}\right)^{10} = 0.001$$

3.2 Microstates and Macrostates

If we know "enough" about the detailed properties (such as position and momentum) of every particle in a system,[4] then we say that the *microstate* of the system is specified. (The criterion for "enough" will be discussed shortly.) We may know less than this but know the *macrostate* of the system. (In an ideal gas, for example, the macrostate would be defined by knowing the number of molecules and volume, and the pressure, temperature, or total energy.) Usually there are many microstates corresponding to each macrostate. The large-scale average properties (such as pressure and number of particles per unit volume in the ideal gas) fluctuate slightly about well-defined mean values.

In the problem of how many molecules are in half of a box, the macrostate is specified if we know how many molecules there are, while a microstate would specify the position and momentum of every molecule. In other cases, internal motions of the molecule may be important, and it will be necessary to know more than just the position and momentum of each particle.

The relation between microstates and macrostates may be clarified by the following example, which contains the essential features, although it is oversimplified and somewhat artificial. A room is empty except for some toys on the floor. Specifying the location of each of the toys on the floor would

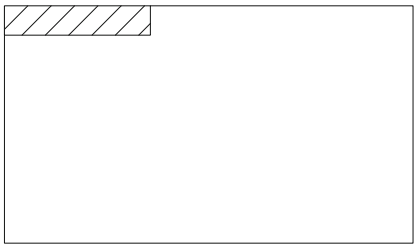


Fig. 3.3 A room with toys. If all the toys are in the shaded area, the macrostate is "picked up." Otherwise, the macrostate is "mess"

specify the microstate of the system. If the toys are in the shaded corner in Fig. 3.3, the macrostate is "picked up." If the toys are in any place else in the room, the macrostate is "mess." There are many more microstates corresponding to the macrostate "mess" than there are corresponding to the macrostate "picked up." We know from experience that children tend to regard any microstate as equally satisfactory; the chances of spontaneously finding the macrostate "picked up" are relatively small.

A situation in which P is small is called *ordered* or *nonrandom*. A situation in which P is large is called *disordered* or *random*. Macrostate "mess" is more probable than macrostate "picked up" and is disordered or random.

The same idea can be applied to a box of gas molecules. Initially, the molecules are all kept in the left half of the box by a partition. If the partition is suddenly removed, a large number of additional microstates are suddenly available to the molecules. The macrostate in which they find themselves—all in the left half of the box, even though the partition has been removed—is very improbable or highly ordered. The molecules soon fill the entire box; it is quite unlikely that they will all be in the left half again if the number of molecules is very large. (Suppose that there are 80 molecules in the box. The probability that all are in the left half is $\left(\frac{1}{2}\right)^{80} = 10^{-24}$. If samples were taken 10^6 times/s, it would take 10^{18} s to sample 10^{24} boxes, one of which, on the average, would have all of the molecules in the left half. This is greater than the age of the universe.)

Just after the partition in the box was removed, the situation was very ordered. The system spontaneously approached a much more random situation in which nearly half the molecules were in each half of the box. The actual number n fluctuates about $N/2$, but in such a way that the average $\langle n\rangle$ (taken, say, over several seconds) no longer changes with time. Typical fluctuations with a constant $\langle n\rangle$ are shown in Fig. 3.4a. When the average[5] of the macroscopic parameters is not changing with time, we say that the system is in

[4] A *system* is that part of the universe that we choose to examine. The *surroundings* are the rest of the universe. The system may or may not be isolated from the surroundings.

[5] There is a subtlety about the meaning of average that we are glossing over here. If we take a whole ensemble of identical systems, which were all prepared the same way, and measure n in each one, we have the *ensemble average* $\bar{n}$. This is calculated in the way described in Appendix G. If we watch one system over some long time interval, as in

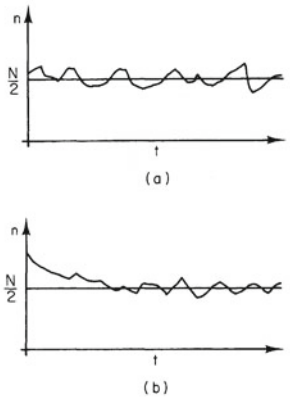


Fig. 3.4 **a** Fluctuations of n about $N/2$. **b** The approach of the system to the equilibrium state after the partition is removed

an *equilibrium state*. Figure 3.4b shows the system moving toward the equilibrium state after the partition is removed.

An equilibrium state is characterized by macroscopic parameters whose average values remain constant with time, although the parameters may fluctuate about the average value. It is also the most random (i.e., most probable) macrostate possible under the prescribed conditions. It is independent of the past history of the system and is specified by a few macroscopic parameters.[6]

The definition of a microstate of a system has so far been rather vague; we have not said precisely what is required to specify it. It is actually easier to specify the microstate of a system when using quantum mechanics than when using classical mechanics. When the energy of an individual particle in a system (such as one of the molecules in the box) is measured with sufficient accuracy, it is found that only certain discrete values of the energy occur. This is because of the wave nature of the particles. The allowed values of the energy are called *energy levels*. You are probably familiar with the idea of energy levels from a previous physics or chemistry course; for example, the spectral lines of atoms are due to the emission of light when an atom changes from one energy level to another. Because the energy levels are well defined, the energy difference, and hence the frequency or color of the light, is also well defined (see Chap. 14).

A particle in a box has a whole set of energy levels at energies determined by the size and shape of the box. Compared to macroscopic measurements of energy, these levels are very close together. The particle can be in any one of these levels; which energy the particle has is specified by a set of *quantum numbers*. If the particle moves in three dimensions, three quantum numbers are needed to specify the energy level. If there are N particles, it will be necessary to specify three quantum numbers for each particle or $3N$ numbers in all. (If there are M molecules, each made up of a atoms, then $N = aM$. The number of quantum numbers is less than $3N$ because the atoms cannot all move independently. If the molecules were thought of as single particles, there would be $3M$ quantum numbers. But the molecules can rotate and vibrate, so that the number of quantum numbers is greater than $3M$ and less than $3N$.)

The total number of quantum numbers required to specify the state of all the particles in the system is called the number of *degrees of freedom* of the system, f.

A microstate of a system is specified if all the quantum numbers for all the particles in the system are specified.

In most of this chapter, it will not be necessary to consider the energy levels in detail. The important fact is that each particle in a system has discrete energy levels, and a microstate is specified if the energy level occupied by each particle is known.

3.3 The Energy of a System: The First Law of Thermodynamics

Figure 3.5 shows some energy levels in a system occupied by a few particles. The total energy of the system U is the sum of the energy of each particle. In making this drawing, we have assumed that all the particles are the same and that they do not interact with one another very much. Then each particle has the same set of energy levels, and the presence of other particles does not change them. In that case, we can say that there is a certain set of energy levels in the system and that each level can be occupied by any number of particles. The energy of the ith level, occupied or not, will be called

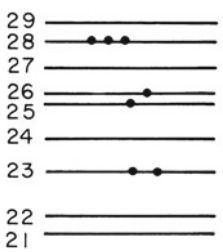


Fig. 3.5 A few of the energy levels in a system. If a particle has a particular energy, a dot is drawn on the level. More than one particle in this system can have the same quantum numbers

Fig. 3.4, we can take the *time average* $\langle n \rangle$. It is taken by recording values of n for a large number of discrete times in some interval. Strictly speaking, an equilibrium state is one in which the ensemble average is not changing with time.

[6] A more detailed discussion of equilibrium states is found in Reif (1964).

u_i. For the example of Fig. 3.5, the total energy is

$$U = 2u_{23} + u_{25} + u_{26} + 3u_{28}.$$

Suppose that the system is isolated so that it does not gain or lose energy. It is still possible for particles within the system to exchange energy and move to different energy levels, as long as the total energy does not change. (Classically, two particles could collide, so that one gains and one loses energy.) Therefore the number of particles occupying each energy level can change, as long as the total energy remains constant. For a system in equilibrium, the average number of particles in each level does not change with time.

There are two ways in which the total energy of a system can change. *Work* can be done on the system by the surroundings, or *heat* can flow from the surroundings to the system. The meaning of work and heat in terms of the energy levels of the system is quite specific and is discussed shortly. First, we define the sign conventions associated with them.

It is customary to define Q to be the heat flow *into* a system. If no work is done, the energy change in the system is

$$\Delta U = Q.$$

It is also customary to call W the work done *by* the system *on* the surroundings. When W is positive, energy flows from the system to the surroundings. If there is no accompanying heat flow, the energy change of the system is

$$\Delta U = -W.$$

The most general way the energy of a system can change is to have both work done by the system and heat flow into the system. The statement of the conservation of energy in that case is called the *first law of thermodynamics*:

$$\Delta U = Q - W. \tag{3.5}$$

The joule is the SI unit for energy, work and heat flow. The calorie (1 cal = 4.184 J) is sometimes used. The dietary Calorie is 1000 cal.

The positions of the energy levels in a system are determined by some macroscopic properties of the system. For a gas of particles in a box, for example, the positions of the levels are determined by the size and shape of the box. For charged particles in an electric field, the positions of the levels are determined by the electric field. If the macroscopic parameters that determine the positions of the energy levels are not changed, the only way to change the total energy of a system is to change the average number of particles occupying each energy level, as in Fig. 3.6. This energy change is called *heat flow*.

Work is associated with the change in the macroscopic parameters (such as volume) that determine the positions of the energy levels. If the energy levels are shifted by doing work without an accompanying heat flow, the change is called *adiabatic*. An adiabatic change is shown in Fig. 3.7. In general, there is also a shift of the populations of the levels in an adiabatic change; the average occupancy of each level can be calculated using the Boltzmann factor, described in Sect. 3.7. There is no heat flow, but work is done on or by the system, and its energy changes.

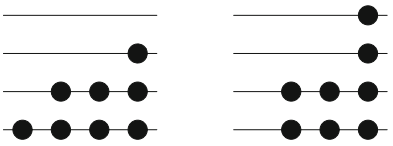


Fig. 3.6 No work is done on the system, but heat is added. The positions of the levels do not change; their average population does change

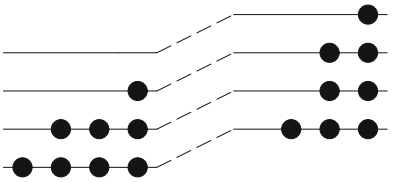


Fig. 3.7 Work is done on the system, but no heat flows. Each level has been shifted to a higher energy

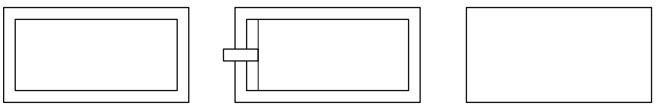


Fig. 3.8 Symbols used to indicate various types of isolation in a system. **a** This system is completely isolated. **b** There is no heat flow through the double wall, but work can be done (symbolized by a piston). **c** No work can be done, but there can be heat flow through the single wall

To summarize: Pure heat flow involves a change in the average number of particles in each level without a change in the positions of the levels. Work involves a change in the macroscopic parameters, which changes the positions of at least some of the energy levels. In general, this means that there is also a shift in the average population of each level. The most general energy change of a system involves both work and heat flow. In that case the total energy change is the sum of the changes due to work and to heat flow.

It is customary in drawing systems to use the symbols in Fig. 3.8 to describe how the system can interact with the surroundings. A double-walled box means that no heat flows, and any processes that occur are adiabatic. This is shown in Fig. 3.8a. If work can be done on the system, a piston is shown as in Fig. 3.8b. If heat can flow to or from the system, a single wall is used as in Fig. 3.8c.

3.4 Ensembles and the Basic Postulates

In the next few sections we will develop some quite remarkable results from statistical mechanics. Making the postulate that when a system is in equilibrium each microstate is equally probable, and arguing that as the energy, volume, or number of particles in the system is increased the number of microstates available to the system increases, we will obtain several well-known results from thermodynamics: heat flows from one system to another in thermal contact until their temperatures are the same; if their volumes can change they adjust themselves until the pressures are the same; and the systems exchange particles until their chemical potentials are the same. We will also obtain the concept of entropy; the Boltzmann factor; the theorem of equipartition of energy; and the Gibbs free energy, which is useful in chemical reactions in living systems where the temperature and pressure are constant.

The initial postulates are deceptively simple. Unfortunately, a fair amount of mathematics is required to get from them to the final results. We start with the basic postulates.

The microstate of a system is determined by specifying the quantum numbers of each particle in the system. The total number of quantum numbers is the number of degrees of freedom. The macrostate of a system is determined by specifying two things:

1. All of the external parameters, such as the volume of a box of gas or any external electric or magnetic field, on which the positions of the energy levels depend. (Classically, all the external parameters that affect the motion of the particles in the system.)
2. The total energy of the system, U.

The external parameters determine a set of energy levels for the particles in the system; the total energy determines which energy levels are accessible to the system.

Statistical physics deals with average quantities and probabilities. We imagine a whole set or ensemble of "identical" systems, as we did in Fig. 3.1. The systems are identical in that they all are in the same macrostate. Different systems within the ensemble will be in different microstates. Imagine that at some instant of time we "freeze" all the systems in the ensemble and examine which microstate each is in. From this we can determine the probability that a system in the ensemble is in microstate i:

$$P(\text{of being in microstate } i) = \frac{\text{number of systems in microstate } i}{\text{total number of systems in the ensemble}}.$$

Imagine that we now "unfreeze" all the systems in the ensemble and let the particles move however they want. At some later time we freeze them again and examine the probability that a system is in each microstate. These probabilities may have changed with time. For example, if the system is a group of particles in a box, and if the initial "freeze" was done just after a partition confining all the particles to the left half of the box had been removed, we would have found many systems in the ensemble in microstates for which most of the particles are on the left-hand side. Later, this would not be true. We would find microstates corresponding to particles in both halves of the box.

We will make two basic *postulates* about the systems in the ensemble.[7]

1. If an isolated system (really, an ensemble of isolated systems) is found with equal probability in each one of its accessible microstates, it is in equilibrium.[8] Conversely, if it is in equilibrium, it is found with equal probability in each one of its accessible microstates.
2. If it is not in equilibrium, it tends to change with time until it is in equilibrium. Therefore the equilibrium state is the most random, most probable state.

For the rest of this chapter, we deal with equilibrium systems. According to our first postulate, each microstate that is accessible to the system (that is, consistent with the total energy that the system has) is equally probable. We will discover that this statement has some far-reaching consequences.

Suppose that we want to consider some variable x, which takes on various values. This variable might be the pressure of a gas, the number of gas molecules in some volume of the box, or the energy that one of the molecules has. For each value of x, there will be some number of microstates in which the system could be that are consistent with that value of x. There will also be some total number of microstates in which the system could be, consistent with its initial preparation. We will use the Greek letter Ω to denote the number of microstates. The total number of accessible microstates (for all possible values of x) is Ω; the number for which x has some particular value is Ω_x. It is consistent with the first assumption to say that the probability that the variable has a value x when the system is in equilibrium is

$$P_x = \frac{\Omega_x}{\Omega}. \tag{3.6}$$

We have been considering ensemble averages. For example, the variable of interest might be the pressure, and we

[7] For a more detailed discussion of these assumptions, see Reif (1964, Chap. 3).

[8] In thermodynamics and statistical mechanics, *equilibrium* and *steady state* do not mean the same thing. Steady state means that some variable is not changing with time. The concentration of sodium in a salt solution flowing through a pipe could be in steady state as the solution flowed through, but the system would not be in equilibrium. Only a few microstates corresponding to bulk motion of the fluid are occupied. In other areas, such as feedback systems, the words equilibrium and steady state are used almost interchangeably.

could find the ensemble average by calculating $\bar{p} = \sum P_p p$, where P_p is the probability of having pressure p. In equilibrium P_p is given by Eq. 3.6, and $\bar{p}$ does not change with time. We could also consider a single system, measure $p(t)$ M times, and compute the time average, $\langle p(t) \rangle = \sum_i p(t_i)/M$. (The equivalence of the time average and the ensemble average for systems in equilibrium is called the *ergodic hypothesis*.)

3.5 Thermal Equilibrium

A system that never interacts with its surroundings is an idealization. The adiabatic walls of Fig. 3.8a can never be completely realized. However, much can be learned by considering two systems that can exchange heat, work, or particles, but that, taken together, are isolated from the rest of the universe. Once we have learned how these two systems interact, the second system can be taken to be the rest of the universe. Eventually, we will allow all three exchanges—heat flow, work, and particles—to take place; for now, it will be convenient to consider only exchanges of energy by heat flow. Figure 3.9 shows the two systems, A and A', isolated from the rest of the universe. The total system will be called A^*. The total number of particles is $N^* = N + N'$. For now N and N' are fixed. The total energy is $U^* = U + U'$. The two systems can exchange energy by heat flow, so that U and U' may change, as long as their sum remains constant.

The number of microstates accessible to the total system is Ω^*. The combined system was originally given a total energy U^* before it was sealed off from the rest of the universe. The barrier between A and A' prevents exchange of particles or work. The total number of microstates depends on how much energy is in each system: when system A has energy U, the total number of microstates is $\Omega^*(U)$.[9]

There are many microstates accessible to the system, with U and U' having different values, subject always to $U^* = U + U'$. Let the total number of microstates, including all possible values of U, be Ω^*_{tot}. Then, according to the postulate, the probability of finding system A with energy U is

$$P(U) = \frac{\Omega^*(U)}{\Omega^*_{\text{tot}}} = C\,\Omega^*(U). \tag{3.7}$$

$C = 1/\Omega^*_{\text{tot}}$ is a constant (independent of U).

If the meaning of Eq. 3.7 is obscure, consider the following example. Systems A and A' each consist of two particles, the energy levels for each particle being at u, $2u$, $3u$, and so forth. The total energy available to the combined system is $U^* = 10u$. The smallest possible energy for system A is $U = 2u$, both particles having energy u. If $U = 3u$, there are two states: in one, the first particle has energy u and the other $2u$; in the second, the particles are reversed. Label these states $(u, 2u)$ and $(2u, u)$. For $U = 4u$, there are three possibilities: $(u, 3u)$, $(2u, 2u)$, and $(3u, u)$. In general, if $U = nu$, there are $n - 1$ states, corresponding to the first particle having energy u, $2u$, $3u$, ..., $(n-1)u$. Table 3.3 shows values for U, U', Ω, and Ω'.

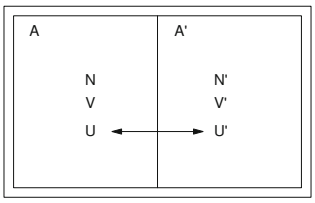


Fig. 3.9 Two systems are in thermal contact with each other but are isolated from the rest of the universe. They can exchange energy only by heat flow

Table 3.3 An example of two systems that can exchange heat energy. The total energy is $U^* = 10u$. Each system contains two particles for which the energy levels are u, $2u$, $3u$, etc

System A		System A'		System A^*
U	Ω	U'	Ω'	Ω^*
$2u$	1	$8u$	7	7
$3u$	2	$7u$	6	12
$4u$	3	$6u$	5	15
$5u$	4	$5u$	4	16
$6u$	5	$4u$	3	15
$7u$	6	$3u$	2	12
$8u$	7	$2u$	1	7
				$\Omega^*_{\text{tot}} = 84$

It is now necessary to consider Ω^* in more detail. If there are two microstates available to system A and 6 available to system A^*, there are $2 \times 6 = 12$ states available to the total system. $\Omega^* = \Omega\Omega'$ is also given in Table 3.3. In a more general case, *the number of microstates for the total system is the product of the number for each subsystem*:

$$\Omega^*(U) = \Omega(U)\,\Omega'(U'). \tag{3.8}$$

For the specific example, there are a total of 84 microstates accessible to the system when $U^* = 10u$. Equation 3.7 says that since each microstate is postulated to be equally probable, the probability that the energy of system A is $3u$ is $12/84 = 0.14$. The most probable state of the combined system is that for which A has energy $5u$ and A' has energy $5u$.

The next question is how Ω and Ω' depend on energy in the general case. In the example, Ω is proportional to U. For three particles, one can show that Ω increases as U^2 (See

[9] If Ω is a continuous function of U, then $\Omega(U)dU$ is actually the number of states wih energy between U and $U + dU$. We ignore this distinction. For a discussion of it, see Chap. 3 of Reif (1964).

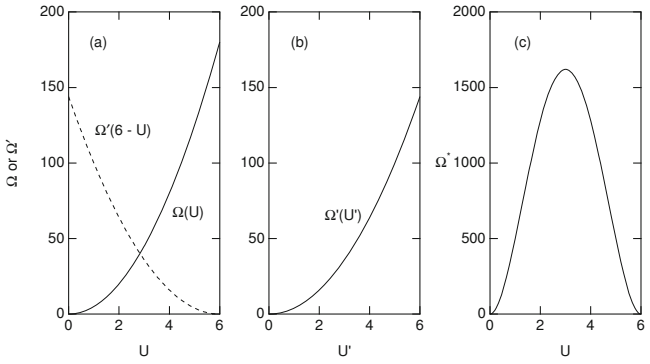


Fig. 3.10 Example of the behavior of Ω, Ω', and Ω^*. In this case, the values used are $\Omega(U) = 5U^2$ and $\Omega'(U') = 4(U')^2$. (These functions give $\Omega = 0$ when $U = 0$, which is not correct. But they are simple and behave properly at higher energies.) The total energy is 6, so only values of U between 0 and 6 are allowed. **a** Plot of $\Omega(U)$. The dashed line is $\Omega'(6 - U)$. **b** Plot of $\Omega'(U')$. **c** Plot of $\Omega^* = \Omega\Omega'$

Problem 19). In general, the more particles there are in a system, the more rapidly Ω increases with U. For a system with a large number of particles, *increasing the energy drastically increases the number of microstates accessible to the system*.

As more energy is given to system A and $\Omega(U)$ increases, there is less energy available for system A' and $\Omega'(U')$ decreases. The product $\Omega^* = \Omega\Omega'$ goes through a maximum at some value of U, and that value of U is therefore the most probable. These features are shown in Fig. 3.10, which assumes that U and Ω are continuous variables. The continuous approximation becomes excellent when we deal with a large number of particles and very closely spaced energy levels. The solid line in Fig. 3.10a represents $\Omega(U)$; $\Omega'(U')$ is the solid line in Fig. 3.10b. The function Ω' is also plotted against U, rather than U', as the dashed line in Fig. 3.10a. As more energy is given to A, Ω increases but Ω' decreases. The product, $\Omega^* = \Omega\Omega'$, shown in Fig. 3.10c, reaches a maximum at $U = 3$.

The most probable value of U is that for which $P(U)$ is a maximum. Since P is proportional to Ω^*, $\Omega^*(U)$ is also a maximum. Therefore,

$$\frac{d}{dU}\left[\Omega^*(U)\right] = 0 \tag{3.9}$$

at the most probable value of U. This derivative can be evaluated using Eq. 3.8. Since $U + U' = U^*$, Eq. 3.8 can be rewritten as

$$\Omega^*(U) = \Omega(U)\,\Omega'(U^* - U). \tag{3.10}$$

The derivative is

$$\frac{d\Omega^*}{dU} = \frac{d\Omega}{dU}\,\Omega' + \Omega\,\frac{d\Omega'}{dU}.$$

By the chain rule for taking derivatives,

$$\frac{d\Omega'}{dU} = \left(\frac{d\Omega'}{dU'}\right)\left(\frac{dU'}{dU}\right).$$

Since $U' = U^* - U$, $dU'/dU = -1$. Therefore

$$\frac{d\Omega^*}{dU} = \Omega'\,\frac{d\Omega}{dU} - \Omega\,\frac{d\Omega'}{dU'}. \tag{3.11}$$

Factoring out $\Omega\,\Omega'$ gives

$$\frac{d\Omega^*}{dU} = \Omega\,\Omega'\left(\frac{1}{\Omega}\,\frac{d\Omega}{dU} - \frac{1}{\Omega'}\,\frac{d\Omega'}{dU'}\right). \tag{3.12}$$

In equilibrium, this must be zero by Eq. 3.9. Since $\Omega^* = \Omega\Omega'$ cannot be zero, the most probable state or the equilibrium state exists when

$$\frac{1}{\Omega}\,\frac{d\Omega}{dU} = \frac{1}{\Omega'}\,\frac{d\Omega'}{dU'}. \tag{3.13}$$

It is convenient to define the quantity τ as

$$\frac{1}{\tau} \equiv \frac{1}{\Omega}\,\frac{d\Omega}{dU}$$

for any system. We must remember that this derivative was taken when the number of particles and the parameters that determine the energy levels were held fixed. These parameters are such things as volume and electric and magnetic fields. To remind ourselves that everything *but* U is being held fixed, it is customary to use the notation for a *partial derivative*: ∂ instead of d (Appendix N). Therefore, we write

$$\frac{1}{\tau} \equiv \frac{1}{\Omega}\left(\frac{\partial\Omega}{\partial U}\right)_{N,V,\text{etc.}}. \tag{3.14}$$

Often we will be careless and just write $\partial\Omega/\partial U$.

The quantity τ defined by Eq. 3.14 depends only on the variables of one system, system A. It is therefore a property of that system. Thermal equilibrium occurs when $\tau = \tau'$. Since Ω is just a number, Eq. 3.14 shows that τ has the dimensions of energy.

Systems A and A', which are in thermal contact, will be in equilibrium (the state of greatest probability) when $\tau = \tau'$. This is reminiscent of something that is familiar to all of us: if a hot system is placed in contact with a cold one, the hotter one cools off and the cooler one gets warmer. The systems come to equilibrium when they are both at the same temperature. This suggests that τ is in some way related to temperature, even though it has the dimensions of energy. We will not prove it, but many things work out right if the absolute temperature T is defined by the relationship

$$\tau = k_B T. \tag{3.15}$$

The proportionality constant is called *Boltzmann's constant*. If T is measured in kelvin (K), k_B has the value

$$k_B = 1.380\,651 \times 10^{-23}\,\mathrm{J\,K^{-1}} = 0.861\,734 \times 10^{-4}\,\mathrm{eV\,K^{-1}}. \tag{3.16}$$

(The *electron volt* (eV) is a unit of energy commonly used when considering atoms or molecules. $1\,\mathrm{eV} = 1.602\,18 \times 10^{-19}\,\mathrm{J}$.) The most convincing evidence in this book that Eq. 3.15 is reasonable is the derivation of the thermodynamic identity in Sect. 3.16.

The *absolute temperature* T is related to the temperature in degrees centigrade or Celsius by

$$T = (\text{temperature in } ^\circ\mathrm{C}) + 273.15. \tag{3.17}$$

3.6 Entropy

The preceding section used the idea that the number of microstates accessible to a system increases as the energy of the system increases, to develop a condition for thermal equilibrium. There are two features of those arguments that suggest that there are advantages to working with the natural logarithm of the number of microstates. First, the total number of microstates is the product of the number in each subsystem: $\Omega^* = \Omega\Omega'$. Taking natural logarithms of this gives

$$\ln \Omega^* = \ln \Omega + \ln \Omega'. \tag{3.18}$$

The other feature is the appearance of $(1/\Omega)\,(\partial\Omega/\partial U)$ in the equilibrium condition. For any non-negative, differentiable function $y(x)$,

$$\frac{d}{dx}(\ln y) = \frac{1}{y}\frac{dy}{dx}.$$

Therefore, Eq. 3.14 can be written as

$$\frac{1}{\tau} = \frac{\partial}{\partial U}(\ln \Omega). \tag{3.19}$$

The *entropy* S is defined by

$$S = k_B \ln \Omega, \qquad \Omega = e^{S/k_B}. \tag{3.20}$$

If both sides of Eq. 3.19 are multiplied by k_B, it is seen that

$$\left(\frac{\partial S}{\partial U}\right)_{N,V,\text{etc.}} = \frac{k_B}{\tau} = \frac{1}{T}. \tag{3.21}$$

This is a fundamental property of entropy that may be familiar to you from other thermodynamics textbooks; if so, it forms a justification for defining temperature as we did.

Another important property of the entropy is that the entropy of system A^* is the sum of the entropy of A and the entropy of A':

$$S^* = S + S'. \tag{3.22}$$

This can be proved by multiplying Eq. 3.18 by k_B.

A third property of the entropy is that S^* is a maximum when systems A and A' are in thermal equilibrium. This result follows from the fact that Ω^* is a maximum at equilibrium, since $S^* = k_B \ln \Omega^*$ and the logarithm is a monotonic function.

Finally, the entropy change in the system can be related to the heat flow into it. Equation 3.21 shows that if there is an energy change in the system *when N and the parameters that govern the spacing of the energy levels are fixed*, then

$$dS = \left(\frac{\partial S}{\partial U}\right)_{N,V,\text{etc.}} dU = \left(\frac{dU}{T}\right)_{N,V,\text{etc.}}.$$

But the energy change when N, V, and any other parameters are fixed is the heat flow dQ:

$$dS = \frac{dQ}{T}. \tag{3.23}$$

3.7 The Boltzmann Factor

Section 3.5 considered the equilibrium state of two systems that were in thermal contact. It is often useful to consider systems in thermal contact when one of the systems is a single particle. This leads to an expression for the total number of microstates as a function of the energy in the single-particle system, known as the *Boltzmann factor*. The Boltzmann factor is used in many situations, as is its alternate form, the *Nernst equation* (Sect. 3.8).

Let system A be a single particle in thermal contact with a large system or *reservoir* A'. Transferring energy from A' to A decreases the number of microstates in A'. The number of microstates in A may change by some factor G or remain the same. We will discuss G at the end of this section.

To make this argument quantitative, consider system A when it has two different energies, U_r and U_s. Reservoir A' is very large so that its temperature T' remains constant, and it has many energy levels almost continuously distributed. Let $\Omega'(U')$ be the number of microstates in A' when it has energy U'. The relative probability that A has energy U_s compared to having energy U_r is given by the ratio of the total number of microstates accessible to the combined system:

$$\frac{P(U_s)}{P(U_r)} = \frac{\Omega^*(U = U_s)}{\Omega^*(U = U_r)} = \frac{\Omega(U_s)\,\Omega'(U^* - U_s)}{\Omega(U_r)\,\Omega'(U^* - U_r)}. \tag{3.24}$$

This probability is the product of two functions, one depending on system A and one on reservoir A':

$$G = \frac{\Omega(U_s)}{\Omega(U_r)},$$
$$R = \frac{\Omega'(U^* - U_s)}{\Omega'(U^* - U_r)}. \tag{3.25}$$

Ratio R is calculated most easily by using Eq. 3.14, remembering the definition $\tau = k_B T$. Since neither the volume nor number of particles is changed, we use an ordinary derivative. We write it in terms of the temperature of the reservoir:

$$\frac{1}{\Omega'}\left(\frac{d\Omega'}{dU'}\right) = \frac{1}{k_B T'},$$
$$\frac{d\Omega'}{dU'} = \left(\frac{1}{k_B T'}\right)\Omega'. \tag{3.26}$$

Since T' is constant, this is easily integrated:

$$\Omega'(U') = \text{const} \times e^{U'/k_B T'}.$$

Therefore the ratio is

$$R = \frac{\text{const} \times e^{(U^* - U_s)/k_B T'}}{\text{const} \times e^{(U^* - U_r)/k_B T'}}$$
$$= e^{-(U_s - U_r)/k_B T}. \tag{3.27}$$

Although the temperature T' is a property of the reservoir, we drop the prime. This ratio is called the *Boltzmann factor*. It gives the factor by which the number of microstates in the reservoir decreases when the reservoir gives up energy $U_s - U_r$ to the system A.

The relative probability of finding system A with energy U_r or U_s is then given by

$$\frac{P(U_s)}{P(U_r)} = G\, e^{-(U_s - U_r)/k_B T} = \left[\frac{\Omega(U_s)}{\Omega(U_r)}\right] e^{-(U_s - U_r)/k_B T}. \tag{3.28}$$

The exponential Boltzmann factor is a property of the reservoir. The factor G is called the *density of states factor*. It is a property of the system. If system A is a single atom with discrete energy levels and we want to know the relative probability that the atom has a particular value of its allowed energy, G may be unity. In other cases, there may be two or more sets of quantum numbers corresponding to the same energy, a situation called *degeneracy*. In that case G may be a small integer. We would have to know the details to calculate it.

3.8 The Nernst Equation

The Nernst equation is widely used in physiology to relate the concentration of ions on either side of a membrane to the electrical potential difference across the membrane. It is an example of the Boltzmann factor.

Suppose that certain ions can pass easily through a membrane. If the membrane has an electrical potential difference across it, the ions will have different energy on each side of the membrane. As a result, when equilibrium exists they will be at different concentrations. The ratio of the probability of finding an ion on either side of the membrane is the ratio of the concentrations on the two sides:

$$\frac{C_2}{C_1} = \frac{P(2)}{P(1)}.$$

The total energy of an ion is its kinetic energy plus its potential energy: $U = E_k + E_p$. Chapter 6 will show that when the electrical potential is v, the potential energy is $E_p = zev$. In this equation z is the valence of the ion ($+1, -1, +2$, etc.) and e is the *elementary charge* (1.6×10^{-19} C).

The concentration ratio is given by a Boltzmann factor, Eq. 3.28:

$$\frac{C_2}{C_1} = \left[\frac{\Omega(2)}{\Omega(1)}\right] e^{-(U_2 - U_1)/k_B T}. \tag{3.29}$$

We must now evaluate the quantity in square brackets. It is the ratio of the number of microstates available to the ion on each side of the membrane. The concentration is the number of ions per unit volume and is proportional to the probability that an ion is in volume $\Delta x \Delta y \Delta z$. We will state without proof that for a particle that can undergo translational motion in three dimensions, $\Omega(U)$ is $\alpha\, \Delta x \Delta y \Delta z$, where α is a proportionality constant. Therefore

$$\frac{\Omega(2)}{\Omega(1)} = \frac{\alpha\, \Delta x \Delta y \Delta z}{\alpha\, \Delta x \Delta y \Delta z} = 1.$$

The energy difference is

$$U_2 - U_1 = E_k(2) - E_k(1) + ze(v_2 - v_1).$$

It will be shown in Sect. 3.10 that the average kinetic energy on both sides of the membrane is the same if the temperature is the same. Therefore,

$$\frac{C_2}{C_1} = e^{-ze(v_2 - v_1)/k_B T}. \tag{3.30}$$

If the potential difference is $v_2 - v_1$, then the ions will be in equilibrium if the concentration ratio is as given by Eq. 3.30. If the ratio is not as given, then the ions, since they are free to move through the membrane, will do so until equilibrium is attained or the potential changes.

If the ions are positively charged and $v_2 > v_1$, then the exponent is negative and $C_2 < C_1$. If the ions are negatively charged, then $C_2 > C_1$.

The concentration difference is explained qualitatively by the electrical force within the membrane that causes the potential difference. If $v_2 > v_1$, the force within the membrane

on a positive ion acts from region 2 toward region 1. It slows positive ions moving from 1 to 2 and accelerates those moving from 2 to 1. Thus it tends to increase C_1.

The Nernst equation is obtained by taking logarithms of both sides of Eq. 3.30:

$$\ln\left(\frac{C_2}{C_1}\right) = -\frac{ze}{k_B T}(v_2 - v_1).$$

From this,

$$v_2 - v_1 = \frac{k_B T}{ze}\ln\left(\frac{C_1}{C_2}\right).$$

Multiplying both numerator and denominator of $k_B T/ze$ by *Avogadro's number* $N_A = 6.022\,141 \times 10^{23}$ molecule mol^{-1} gives the quantities $N_A k_B$ and $N_A e$. The former is the *gas constant*:

$$N_A k_B = R = 8.314\,46\,\text{J mol}^{-1}\,\text{K}^{-1}. \tag{3.31}$$

The latter is the Faraday constant:

$$N_A e = F = 96\,485.34\,\text{C mol}^{-1}. \tag{3.32}$$

The coefficient is therefore

$$\frac{k_B T}{ze} = \frac{RT}{zF}. \tag{3.33}$$

At body temperature, $T = 37\,^\circ\text{C} = 310\,\text{K}$, the value of RT/F is $0.0267\,\text{J C}^{-1} = 26.7\,\text{mV}$.

In the form

$$v_2 - v_1 = \frac{RT}{zF}\ln\left(\frac{C_1}{C_2}\right), \tag{3.34}$$

the Boltzmann factor is called the *Nernst equation*.

3.9 The Pressure Variation in the Atmosphere

It is well known that the atmospheric pressure decreases with altitude. This truth has medical significance because of the effects of lower oxygen at high altitudes. We will derive an approximate, constant-temperature model for the decrease using the Boltzmann factor, and then we will do it again using hydrostatic equilibrium.

The gravitational potential energy of an air molecule at height y is mgy, where m is the mass of the molecule and g is the gravitational acceleration. If the atmosphere has a constant temperature, there will be no change of kinetic energy with altitude. For a molecule to increase its potential energy, and therefore its total energy, by mgy, the energy of all the other molecules (the reservoir) must decrease, with a corresponding decrease in the number of accessible microstates. The number of particles per unit volume is given by a Boltzmann factor:

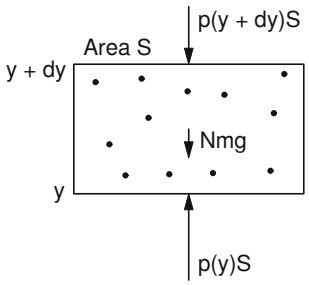


Fig. 3.11 Forces on a small volume element of the atmosphere

$$C(y) = C(0)e^{-mgy/k_B T}. \tag{3.35}$$

Since for an ideal gas $p = Nk_B T/V = Ck_B T$, the pressure also decreases exponentially with height.

The same result can be obtained without using statistical physics, by considering a small volume of the atmosphere that is in static equilibrium. Let the volume have thickness dy and horizontal cross-sectional area S, as shown in Fig. 3.11. The force exerted upward across the bottom face of the element is $p(y)S$. The force down on the top face is $p(y+dy)S$. The N molecules in the volume each experience the downward force of gravity. The total gravitational force is Nmg. In terms of the concentration, $N = CSdy$. Therefore, the condition for equilibrium is $p(y)S - p(y+dy)S - CSmg\,dy = 0$. Since $p(y) - p(y+dy) = -(dp/dy)\,dy$, this can be written as

$$\left[-\left(\frac{dp}{dy}\right) - Cgm\right] S\,dy = 0.$$

The next step is to use the ideal gas law to write $p = Ck_B T$:

$$-k_B T\frac{dC}{dy} - Cmg = 0.$$

If this is written in the form

$$\frac{dC}{dy} = -\frac{mg}{k_B T}C \tag{3.36}$$

it will be recognized as the equation for exponential decay. The solution is Eq. 3.35.

3.10 Equipartition of Energy and Brownian Motion

A very important application of the Boltzmann factor is the proof that the average translational kinetic energy per degree of freedom of a particle in thermal contact with a reservoir at

temperature T is $k_B T/2$. This result holds for any term in the total energy that depends on the square of one of the variables (such as a component of the position or the momentum).

The proof is done for the kinetic energy resulting from the x component of momentum. The same procedure can be used for the other components. When the x component of the momentum of a particle is between p_x and $p_x + dp_x$, the kinetic energy is $p_x^2/2m$. The relative probability that the particle has this energy is given by the Boltzmann factor, $e^{-p_x^2/2mk_BT}$. We assert that the probability that the particle has momentum in this interval is also proportional to dp_x.[10] The average kinetic energy associated with p_x is obtained by multiplying the energy by the Boltzmann factor and integrating over all values of p_x. We normalize the probability by dividing by the integral of the Boltzmann factor.

$$\overline{\left(\frac{p_x^2}{2m}\right)} = \frac{\int_{-\infty}^{\infty} (p_x^2/2m) e^{-p_x^2/2mk_BT}\, dp_x}{\int_{-\infty}^{\infty} e^{-p_x^2/2mk_BT}\, dp_x}. \tag{3.37}$$

The integral in the denominator is evaluated in Appendix K and is $(2\pi mk_BT)^{1/2}$. The integral in the numerator of Eq. 3.37 is

$$\left(\frac{1}{m}\right)\left(\frac{1}{4}\right)(2mk_BT)(2\pi mk_BT)^{1/2}.$$

Combining these gives

$$\overline{\left(\frac{p_x^2}{2m}\right)} = \frac{k_BT}{2}. \tag{3.38}$$

The average value of the kinetic energy corresponding to motion in the x direction is $k_BT/2$, independent of the mass of the particle. The only condition that went into this derivation was that the energy depended on the square of the variable. Any term in the total energy that is a *quadratic function of some variable* will carry through the same way, so that the average energy will be $k_BT/2$ for that variable. This result is called the *equipartition of energy*.

The total translational kinetic energy is the sum of three terms $(p_x^2 + p_y^2 + p_z^2)/2m$, so the total translational kinetic energy has average value $\frac{3}{2}k_BT$.

This result is true for particles of *any* mass: atoms, molecules, pollen grains, and so forth. Heavier particles will have a smaller velocity but the same average kinetic energy. Even heavy particles are continually moving with this average kinetic energy. The random motion of pollen particles in water was first seen by a botanist, Robert Brown, in 1827. This *Brownian motion* is an important topic in the next chapter.

[10] A more detailed justification of this is found in earlier editions of this book, in texts on statistical mechanics, or on the web site associated with this book.

3.11 Heat Capacity

Consider a system into which a small amount of heat Q flows. In many cases the temperature of the system rises. (An exception is when there is a change of state such as the melting of ice.) The *heat capacity* C of the system is defined as

$$C = \frac{Q}{\Delta T}. \tag{3.39}$$

Heat capacity has units of J K^{-1}. It depends on the size of the object and the substance it is made of. The *specific heat capacity*, c, is the heat capacity per unit mass (J K^{-1} kg^{-1}) or the heat capacity per mole (J K^{-1} mol^{-1}).

The heat capacity also depends on any changes in the macroscopic parameters that take place during the heat flow. Recall the first law of thermodynamics, Eq. 3.5: $\Delta U = Q - W$. Only part of the energy transferred to the system by the heat flow increases the internal energy. Some also goes to work done by the system. For example, if the volume changes, there will be pressure-volume work done by the system (Sect. 1.18).

One special case is the heat capacity at constant volume, C_V. In that case, no pdV work is done by the system and $\Delta U = Q$, so

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V. \tag{3.40}$$

Many processes in the body occur at constant pressure. The heat capacity at constant pressure, C_p, is not equal to C_V. If both the pressure and volume change during the process, the heat capacity depends on the details of the pressure and volume changes.

The simplest example is the heat capacity at constant volume of a monatomic ideal gas. The average kinetic energy of a gas molecule at temperature T moving in three dimensions is $\frac{3}{2}k_BT$, and the total energy of N molecules is $U = \frac{3}{2}Nk_BT$. Therefore at constant volume $C_V = \frac{3}{2}Nk_B$. For one mole of monatomic ideal gas the heat capacity is $\frac{3}{2}N_Ak_B = \frac{3}{2}R$. Molecules with two or more atoms can also have rotational and vibrational energy, and the heat capacity is larger. The heat capacity can also depend on the temperature.

As a biological example, consider the energy loss from breathing (Denny 1993). In each breath we inhale about $V = 0.5$ l of air. Our body warms this air from the surrounding temperature to body temperature. (The body has a much higher heat capacity and does not significantly cool. See Problem 49.) The specific heat of air under these conditions is $c \approx 1000$ J K^{-1} kg^{-1}, and the density of air is $\rho = 1.3$ kg m^{-3}. Therefore the heat flow required to raise the air temperature in each breath is

$$Q = c\rho V \left(T_{\text{body}} - T_{\text{surroundings}}\right). \tag{3.41}$$

For a body temperature of 37 °C and surroundings at 20 °C, the temperature difference is 17 °C = 17 K. From Eq. 3.41, $Q = 11$ J. We breathe about once every 5 s, so the average power lost to the air we breathe is 2.2 W. A typical basal metabolic rate is about 100 W, so this represents 2 % of our energy consumption.

3.12 Equilibrium When Particles Can Be Exchanged: the Chemical Potential

Section 3.5 considered two systems that could exchange heat. The most probable or equilibrium state was that in which energy had been exchanged so that the total number of microstates or total entropy was a maximum. This occurred when (Eq. 3.13)

$$\frac{1}{\Omega}\left(\frac{\partial\Omega}{\partial U}\right)_{N,V} = \frac{1}{\Omega'}\left(\frac{\partial\Omega'}{\partial U'}\right)_{N',V'},$$

which is equivalent to $T = T'$. Since $S = k_B \ln\Omega$ this is also equivalent to

$$\left(\frac{\partial S}{\partial U}\right)_{N,V} = \left(\frac{\partial S'}{\partial U'}\right)_{N',V'}.$$

This section considers the case in which the systems can exchange both energy by heat flow and particles; they are in thermal and diffusive contact (Fig. 3.12). The number of particles in each system is not fixed, but their sum is constant:

$$N + N' = N^*. \tag{3.42}$$

Equilibrium will exist for the most probable state, which means that there is heat flow until the two temperatures are the same and Eq. 3.13 is satisfied. The most probable state also requires a maximum in Ω^* or S^* vs N. The arguments used in the earlier section for heat exchange can be applied to obtain the equilibrium condition

$$\frac{1}{\Omega}\left(\frac{\partial\Omega}{\partial N}\right)_{U,V} = \frac{1}{\Omega'}\left(\frac{\partial\Omega'}{\partial N'}\right)_{U',V'}. \tag{3.43}$$

The condition in terms of entropy is

$$\left(\frac{\partial S}{\partial N}\right)_{U,V} = \left(\frac{\partial S'}{\partial N'}\right)_{U',V'}. \tag{3.44}$$

For thermal contact, the temperature was defined in terms of the derivative of S with respect to U, so that equilibrium occurred when $T = T'$. An analogous quantity, the *chemical potential*, is defined by

$$\mu \equiv -T\left(\frac{\partial S}{\partial N}\right)_{U,V}. \tag{3.45}$$

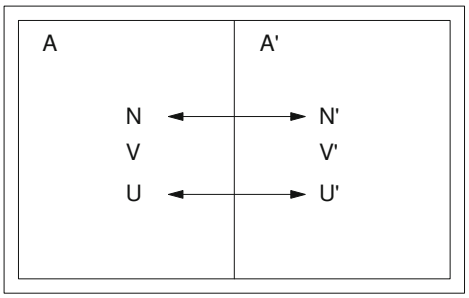


Fig. 3.12 Two systems can exchange energy by heat flow and paticles. The volume of each system remains fixed

(The reason T is included in the definition will become clear later.) Both thermal and diffusive equilibrium exist when

$$T = T', \qquad \mu = \mu'. \tag{3.46}$$

Two systems are in thermal and diffusive equilibrium when they have the same temperature and the same chemical potential.

Since the units of S are J K^{-1} and the units of N are dimensionless,[11] Eq. 3.45 shows that the units of chemical potential are J.

Consider next what happens to the entropy of the total system if particles are exchanged when the system is not in equilibrium. Let the number of particles in the unprimed system increase by ΔN and the number in the primed system change by $\Delta N' = -\Delta N$. The change of total entropy is

$$\Delta S^* = \left(\frac{\partial S^*}{\partial N}\right)\Delta N = \left(\frac{\partial S}{\partial N}\right)\Delta N + \left(\frac{\partial S'}{\partial N'}\right)\Delta N'.$$

Using the definition of the chemical potential we can rewrite this as

$$\Delta S^* = \left(-\frac{\mu}{T}\right)\Delta N - \left(-\frac{\mu'}{T'}\right)\Delta N.$$

If the two temperatures are the same, this is

$$\Delta S^* = \left(\frac{\mu' - \mu}{T}\right)\Delta N. \tag{3.47}$$

We see again that the entropy change will be zero for a small transfer of particles from one system to the other if $\mu = \mu'$. Suppose now that particles flow from A' to A, so that ΔN is positive. If $\mu' > \mu$, that is, the chemical potential of A' is greater than that of A, this will cause an increase in entropy of the combined system. *If particles move from a system of higher chemical potential to one of lower chemical potential, the entropy of the total system increases.*

[11] In this book, N represents the number of particles, and the chemical potential has units of energy per particle. In other books it may have units of energy per mole.

3.13 Concentration Dependence of the Chemical Potential

The change in chemical potential of an ideal gas (or a solute in an ideal solution)[12] when the concentration changes from C_0 to C and there is also a change in its potential energy has the form

$$\Delta\mu = k_B T \ln\left(\frac{C}{C_0}\right) + \Delta(\text{potential energy per particle}). \tag{3.48}$$

We will derive this in Sect. 3.18; for now we show that it is plausible and consistent with the Boltzmann factor.

We know from experience that particles tend to move from a region of higher to lower potential energy, thus increasing their kinetic energy, which can then be transferred as heat to other particles by collision. We also know that particles will move from a region of high concentration to a region of lower concentration. This process, called diffusion, is discussed in Chap. 4. Both processes cause a decrease in the chemical potential and therefore an increase in the entropy.

It is the combination of these two factors that causes the Boltzmann distribution of particles in the atmosphere. When the atmosphere is in equilibrium, the potential energy term increases with height and the concentration term decreases with height so that the chemical potential is the same at all heights.

To see the equivalence between Eq. 3.48 and the Boltzmann factor, suppose that particles can move freely from region 1 to region 2 and that the potential energy difference between the two regions is ΔE_p. The particles will be in equilibrium when $\mu_1 = \mu_2$. From Eq. 3.48 this means that

$$k_B T \ln C_1 + E_{p1} = k_B T \ln C_2 + E_{p2}.$$

This equation can be rearranged to give

$$\ln C_2 - \ln C_1 = -\frac{E_{p2} - E_{p1}}{k_B T}.$$

If exponentials are taken of each side, the result is

$$\frac{C_2}{C_1} = e^{-\Delta E_p / k_B T}.$$

If the temperature of each region is the same, the average kinetic energy will be the same in each system, and $\Delta E_p = \Delta U$. This is then the same as the Boltzmann factor, Eq. 3.29.

There is still another way to look at the concentration dependence. In an ideal gas, the pressure, volume, temperature, and number of particles are related by the equation of state $pV = Nk_B T$. In terms of the particle concentration $C = N/V$, this is $p = Ck_B T$. The work necessary to concentrate the gas from volume V_1 and concentration C_1 to V_2 and C_2 is (see Eq. 1.57)

$$W_{\text{on gas}} = -\int_{V_1}^{V_2} p(V)\, dV. \tag{3.49}$$

The concentration work at a constant temperature is

$$W = -Nk_B T \int_{V_1}^{V_2} \frac{dV}{V} = -Nk_B T \ln\frac{V_2}{V_1}.$$

If the final volume is smaller than the initial volume, the logarithm is negative and the concentration work is positive. In terms of the particle concentration $C = N/V$ or the molar concentration $c = n/V$, the concentration work is

$$W_{\text{conc}} = Nk_B T \ln\frac{C_2}{C_1} = nRT \ln\frac{c_2}{c_1}. \tag{3.50}$$

The last form was written by observing that $Nk_B = nR$ where R is the gas constant per mole.

Comparing Eq. 3.50 with Eq. 3.48, we see that the concentration work at constant temperature is proportional to the change in chemical potential with concentration. It is, in fact, just the number of molecules N times the change in μ: $W_{\text{conc}} = N\Delta\mu$.

The concentration work or change of chemical potential can be related to the Boltzmann factor in still another way. Particles are free to move between two regions of different potential energy at the same temperature. The work required to change the concentration is, by Eq. 3.50,

$$W_{\text{conc}} = N\Delta\mu = Nk_B T \ln\frac{C_2}{C_1}.$$

The concentration ratio is given by a Boltzmann factor:

$$C_2/C_1 = e^{-(E_{p2}-E_{p1})/k_B T},$$

so that $\ln(C_2/C_1) = -(E_{p2} - E_{p1})/k_B T$. Therefore, the concentration work is $W_{\text{conc}} = -N(E_{p2} - E_{p1})$.

If $C_2 < C_1$, W is negative and is equal in magnitude to the increase in potential energy of the molecules. The concentration energy lost by the molecules is precisely that required for them to move to the region of higher potential energy. If $C_2 > C_1$, the loss of potential energy going from region 1 to region 2 provides the energy necessary to concentrate the gas. Alternatively, one may say that the sum of the concentration energy and the potential energy is the same in the two regions. This was, in fact, the statement about the chemical potential at equilibrium: $\mu_2 = \mu_1$.

The same form for the chemical potential is obtained for a dilute solute. (We will present one way of understanding why in Sect. 3.18.) Therefore, the concentration work calculated for an ideal gas is the same as for an ideal solute. The

[12] An ideal solution is defined in Sect. 3.18.

work required to concentrate 1 mol of substance by a factor of 10 at 310 K is (1 mol)(8.31 J mol^{-1} K^{-1})(310 K) ln(10) or 5.93×10^3 J. The H^+ ion in gastric juice has a pH of 1. Since it was concentrated from plasma with a pH of about 7, the concentration ratio is 10^6. The work necessary to concentrate 1 mol is therefore $RT\ \ln(10^6) = (8.31)(310)(13.82) = 3.56 \times 10^4$ J.

3.14 Systems That Can Exchange Volume

We have considered two systems that can exchange energy or particles. Now consider the systems shown in Fig. 3.13. They are isolated from the rest of the universe. The vertical line that separates them is a piston that can move and conduct heat, so that energy and volume can be exchanged between the two systems. The piston prevents particles from being exchanged. The constraints are $V^* = V + V'$ and $U^* = U + U'$ from which $dV = -dV'$, $dU = -dU'$. As before, equilibrium exists when the total number of microstates or the total entropy is a maximum. The conditions for maximum entropy are

$$\left(\frac{\partial S^*}{\partial U}\right)_{N,V} = 0, \qquad \left(\frac{\partial S^*}{\partial V}\right)_{N,U} = 0.$$

The derivation proceeds as before. For example,

$$\left(\frac{\partial S^*}{\partial V}\right)_{N,U} = \left(\frac{\partial S}{\partial V}\right)_{N,U} + \left(\frac{\partial S'}{\partial V}\right)_{N,U} = \left(\frac{\partial S}{\partial V}\right)_{N,U} - \left(\frac{\partial S'}{\partial V'}\right)_{N',U'}.$$

Equilibrium requires that $T = T'$ so that there is no heat flow. The piston will stop moving and there will be no change of volume when

$$\left(\frac{\partial S}{\partial V}\right)_{N,U} = \left(\frac{\partial S'}{\partial V'}\right)_{N',U'}. \tag{3.51}$$

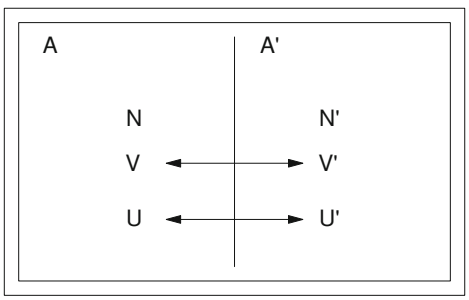


Fig. 3.13 Two systems that can exchange volume are separated by a movable piston. Heat can also flow through the piston

These derivatives can be evaluated in several ways. The method used here involves some manipulation of derivatives; a more detailed description, consistent with the microscopic picture of energy levels, is found in Reif (1964, pp. 267–273).

For a small exchange of heat and work, the first law can be written as $dU = dQ - dW$. In the present case the only form of work is that related to the change of volume, so $dU = dQ - pdV$. It was shown in Eq. 3.23 that $dQ = TdS$. Therefore $dU = TdS - pdV$. This equation can be solved for dS:

$$dS = \left(\frac{1}{T}\right) dU + \left(\frac{p}{T}\right) dV. \tag{3.52}$$

The entropy depends on U,V and N: $S = S(U, V, N)$. If N is not allowed to change, then

$$dS = \left(\frac{\partial S}{\partial U}\right)_{N,V} dU + \left(\frac{\partial S}{\partial V}\right)_{N,U} dV. \tag{3.53}$$

Comparison of this with Eq. 3.52 shows that

$$\left(\frac{\partial S}{\partial U}\right)_{N,V} = \frac{1}{T},$$

$$\left(\frac{\partial S}{\partial V}\right)_{N,U} = \frac{p}{T}. \tag{3.54}$$

The first of these equations was already seen as Eq. 3.21. The second gives the condition for equilibrium under volume change. Referring to Eq. 3.51 we see that at equilibrium

$$\frac{p}{T} = \frac{p'}{T'}.$$

Therefore, equilibrium requires both $T = T'$ and

$$p = p'. \tag{3.55}$$

This agrees with common experience. The piston does not move when the pressure on each side is the same.

3.15 Extensive Variables and Generalized Forces

The number of microstates and the entropy of a system depend on the number of particles, the total energy, and the positions of the energy levels of the system. The positions of the energy levels depend on the volume and may also depend on other macroscopic parameters. For example, they may depend on the length of a stretched muscle fiber or a protein molecule. For charged particles in an electric field, they depend on the charge. For a thin film such as the fluid lining the alveoli of the lungs, the entropy depends on the surface area

Table 3.4 Examples of extensive variables and the generalized force associated with each of them

x	X	$dU = -dW$
Volume V	$-$pressure $-p$	$-p\,dV$
Length L	Force F	$F\,dL$
Area a	Surface tension σ	$\sigma\,da$
Charge q	Potential v	$v\,dq$

of the film. The number of particles, energy, volume, electric charge, surface area, and length are all *extensive variables*: if a homogeneous system is divided into two parts, the value of the variable for the total system (volume, charge, etc.) is the sum of the values for each part. A general extensive variable will be called x.

An adiabatic energy change is one in which no heat flows to or from the system. The energy change is due to work done on or by the system as a macroscopic parameter changes, shifting at least some of the energy levels. For each extensive variable x we can define a *generalized force* X such that the energy change in an adiabatic process is

$$dU = -dW = X dx. \tag{3.56}$$

(Remember that dU is the increase in energy of the system and dW is the work done by the system on the surroundings.) Examples of extensive variables and their associated forces are given in Table 3.4.

3.16 The General Thermodynamic Relationship

Suppose that a system has N particles, total energy U, volume V, and another macroscopic parameter x on which the positions of the energy levels may depend. The number of microstates, and therefore the entropy, will depend on these four variables: $S = S(U, N, V, x)$. If each variable is changed by a small amount, there is a change of entropy

$$dS = \left(\frac{\partial S}{\partial U}\right)_{N,V,x} dU + \left(\frac{\partial S}{\partial N}\right)_{U,V,x} dN + \left(\frac{\partial S}{\partial V}\right)_{U,N,x} dV + \left(\frac{\partial S}{\partial x}\right)_{U,N,V} dx. \tag{3.57}$$

Now consider the change of energy of the system. If only heat flow takes place, there is an increase of energy $dQ = TdS$. If an adiabatic process with a constant number of particles takes place, the energy change is $-dW = Xdx - pdV$. If particles flow into the system without an accompanying flow of heat or work, the energy change is dU_N. It seems reasonable that this energy change, due solely to the movement of the particles, is proportional to dN: $dU_N = a\,dN$. (It will turn out that the proportionality constant is the chemical potential.) For the total change of energy resulting from all these processes, we can write a statement of the conservation of energy: $dU = TdS + Xdx - pdV + adN$. This is an extension of Eq. 3.5 to the additional variables on which the energy can depend. It can be rearranged as

$$dS = \left(\frac{1}{T}\right) dU - \left(\frac{a}{T}\right) dN + \left(\frac{p}{T}\right) dV - \left(\frac{X}{T}\right) dx. \tag{3.58}$$

Comparison of Eqs. 3.57 and 3.58 shows that

$$\left(\frac{\partial S}{\partial U}\right)_{N,V,x} = \frac{1}{T}, \tag{3.59a}$$

$$\left(\frac{\partial S}{\partial N}\right)_{U,V,x} = -\frac{a}{T}, \tag{3.59b}$$

$$\left(\frac{\partial S}{\partial V}\right)_{U,N,x} = \frac{p}{T}, \tag{3.59c}$$

$$\left(\frac{\partial S}{\partial x}\right)_{U,N,V} = -\frac{X}{T}. \tag{3.59d}$$

Comparison of Eq. 3.59b with Eq. 3.45 shows that $a = \mu$. This is why the factor of T was introduced in Eq. 3.45.

Equation 3.58, with the correct value inserted for a, is

$$TdS = dU - \mu\,dN + p\,dV - X\,dx. \tag{3.60}$$

This is known as the *thermodynamic identity* or the *fundamental equation of thermodynamics*. It is a combination of the conservation of energy with the relationship between entropy change and heat flow in a reversible process. (A reversible process is one that takes place so slowly that all parts of the system have the same temperature, pressure, etc.) This equation and derivative relations such as Eqs. 3.59 form the basis for the usual approach to thermodynamics.

Finally, let us consider the addition of a particle to a system when the volume is fixed. If we do this without changing the energy, it increases the number of ways the existing energy can be shared and hence the number of microstates. Therefore the entropy increases. If we want to restore the entropy to its original value, we must remove some energy. Exactly the same argument can be made mathematically. We have seen in Eqs. 3.45 and 3.59b that

$$\mu = -T\left(\frac{\partial S}{\partial N}\right)_{U,V}.$$

Since adding the particle at constant energy increases the entropy, $(\partial S/\partial N)_{U,V}$ is positive and the chemical potential is negative. Next, we rearrange Eq. 3.60 as $dU = TdS + \mu\,dN - p\,dV$ and by inspection see that

$$\mu = \left(\frac{\partial U}{\partial N}\right)_{S,V}.$$

Therefore adding a particle at constant volume while keeping the entropy constant requires that energy be removed from the system.

3.17 The Gibbs Free Energy

A conventional course in thermodynamics develops several functions of the entropy, energy, and macroscopic parameters that are useful in certain special cases. One of these is the *Gibbs free energy*, which is particularly useful in describing changes that occur in a system while the temperature and pressure remain constant. Most changes in a biological system occur under such conditions.

3.17.1 Gibbs Free Energy

Imagine a system A in contact with a much larger reservoir as in Fig. 3.14. The reservoir has temperature T' and pressure p'. A movable piston separates A and A'. (At equilibrium, $T = T'$ and $p = p'$.) The reservoir is large enough so that a change of energy or volume of system A does not change T' or p'.

Consider the change of entropy of the total system that accompanies an exchange of energy or volume between A and A'. Above, this entropy change was set equal to zero to obtain the condition for equilibrium. In this case, however, we will express the total entropy change of system plus reservoir in terms of the changes in system A alone. The total entropy is $S^* = S + S'$, so the total entropy change is $dS^* = dS + dS'$.

If reservoir A' exchanges energy with system A, the energy change is

$$dU' = T'\,dS' - dW' = T'\,dS' - p'\,dV'.$$

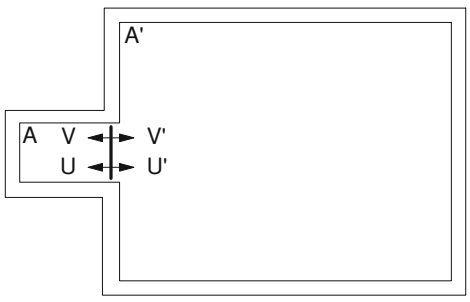


Fig. 3.14 System A is in contact with reservoir A'. Heat can flow through the piston, which is also free to move. The reservoir is large enough to ensure that anything that happens to system A takes place at constant temperature and pressure

This can be solved for dS', and the result can be put in the expression for the total entropy change:

$$dS^* = dS + \frac{dU'}{T'} + \frac{p'\,dV'}{T'}.$$

We are trying to get dS^* in terms of changes in system A alone. Since A and A' together constitute an isolated system, $dU = -dU'$ and $dV = -dV'$. Therefore,

$$dS^* = -\frac{-T'\,dS + dU + p'\,dV}{T'}. \tag{3.61}$$

(Note that a minus sign was introduced in front of this equation.) This expresses the total entropy change in terms of changes of S, U, and V in system A and the pressure and temperature of the reservoir.

The Gibbs free energy is *defined* to be

$$G \equiv U - T'S + p'V. \tag{3.62}$$

If the reservoir is large enough so that interaction of the system and reservoir does not change T' and p', then the change of G as system A changes is

$$dG = dU - T'dS + p'dV. \tag{3.63}$$

Comparison of Eqs. 3.61 and 3.63 shows that

$$dS^* = -\frac{dG}{T'}. \tag{3.64}$$

The change in entropy of system plus reservoir is related to the change of G, which is a property of the system alone, as long as the pressure and temperature are maintained constant by the reservoir.

To see why G is called a free energy, consider the conservation of energy in the following form:

$$\begin{aligned}\text{(work done by the system)} = \text{(energy lost by the system)} \\ + \text{(heat added to the system)},\end{aligned}$$

$$dW = -dU + T\,dS.$$

Subtracting pdV from both sides of this equation gives

$$dW - p\,dV = -dU + T\,dS - p\,dV = -dG.$$

The right-hand side is the decrease of Gibbs free energy of the system. The work done in any isothermal, isobaric (constant pressure) reversible process, *exclusive of pdV work*, is equal to the decrease of Gibbs free energy of the system. This non–$p\,dV$ work is sometimes called *useful work*. It may represent contraction of a muscle fiber, the transfer of particles from one region to another, the movement of charged particles in an electric field, or a change of concentration of particles. It differs from the change in energy of the system,

dU, for two reasons. The volume of the system can change, resulting in $p\,dV$ work, and there can be heat flow during the process. For example, let the system be a battery at constant temperature and pressure which decreases its internal (chemical) energy and supplies electrical energy. From a chemical energy change dU we subtract $T\,dS$, the heat flow to the surroundings, and $-p\,dV$, the work done on the atmosphere as the liquid in the battery changes volume. What is left is the energy available for electrical work.

3.17.2 An Example: Chemical Reactions

As an example of how the Gibbs free energy is used, consider a chemical reaction that takes place in the body at constant temperature and pressure. System A, the region in the body where the reaction takes place, is in contact with a reservoir A' that is large enough to maintain constant temperature and pressure. Suppose that there are four species of particles that interact. Capital letters represent the species and small letters represent the number of atoms or molecules of each that enter in the reaction:

$$aA + bB \longleftrightarrow cC + dD.$$

An example is 1 glucose$+6O_2 \longleftrightarrow 6CO_2+6H_2O$, where $a = 1, b = 6, c = 6, d = 6$. The state of the system depends on U, V, N_A, N_B, N_C, and N_D.

We begin with the definition of G, Eq. 3.62, and we call the pressure and temperature of the system and reservoir p and T:

$$G = U - TS + pV.$$

Differentiating, we obtain

$$dG = dU - T\,dS - S\,dT + p\,dV + V\,dp.$$

Generalize Eq. 3.60 for the case of four chemical species:

$$T\,dS = dU - \mu_A\,dN_A - \mu_B\,dN_B - \mu_C\,dN_C - \mu_D\,dN_D + p\,dV.$$

Insert this in the equation for dG and remember that since the process takes place at constant temperature and pressure, dT and dp are both zero. The result is

$$dG = \mu_A\,dN_A + \mu_B\,dN_B + \mu_C\,dN_C + \mu_D\,dN_D.$$

In Sect. 3.13 we saw that the concentration dependence of the chemical potential is given by a logarithmic term. Equation 3.48 can be used to write

$$\mu_A = \mu_{A0} + k_B T\,\ln(C_A/C_0),$$

where μ_{A0} is the chemical potential at a standard concentration (usually 1 molar, that is, 1 mol l^{-1}) and depends on temperature, pH, etc. Note that C_0 is the same reference concentration for all species. As the reaction takes place to the right, we can write the number of molecules gained or lost as $dN_A = -a\,dN$, $dN_B = -b\,dN$, $dN_C = c\,dN$, $dN_D = d\,dN$, so that we have

$$\begin{aligned} dG = &\left[\mu_{A0} + k_B T\,\ln(C_A/C_0)\right](-a\,dN) \\ &+ \left[\mu_{B0} + k_B T\,\ln(C_B/C_0)\right](-b\,dN) \\ &+ \left[\mu_{C0} + k_B T\,\ln(C_C/C_0)\right](c\,dN) \\ &+ \left[\mu_{D0} + k_B T\,\ln(C_D/C_0)\right](d\,dN). \end{aligned}$$

This can be rearranged as (letting $C_A = [A]$, etc.)

$$\begin{aligned} dG = &[c\mu_{C0} + d\mu_{D0} - a\mu_{A0} - b\mu_{B0} \\ &+ k_B T\,\ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right) - k_B T\,\ln\left(\frac{[C_0]^a[C_0]^b}{[C_0]^c[C_0]^d}\right)\Bigg]\,dN. \end{aligned}$$

The two logarithm terms together represent logs of concentration ratios. Therefore concentrations $[A]$, $[B]$, $[C]$, $[D]$, and C_0 must all be measured in the same units. The last term can be made to vanish if the units are such that C_0 is unity (for example 1 mol per liter). Then

$$\begin{aligned} dG = &[c\mu_{C0} + d\mu_{D0} - a\mu_{A0} - b\mu_{B0} \\ &+ k_B T\,\ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right)\Bigg]\,dN. \end{aligned}$$

Multiplying the expression in square brackets by Avogadro's number converts the chemical potential per molecule to the standard Gibbs free energy per mole, and $k_B T$ to RT. To compensate, the change in number of molecules dN is changed to moles dn or Δn:

$$\begin{aligned} \Delta G = &[(cG_{C0} + dG_{D0} - aG_{A0} - bG_{B0}) \\ &+ RT\,\ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right)\Bigg]\,\Delta n. \end{aligned} \tag{3.65}$$

The term in small parentheses is the standard free energy change for this reaction, ΔG^0, which can be found in tables. At equilibrium $\Delta G = 0$, so

$$0 = \Delta G^0 + RT\,\ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right) = \Delta G^0 + RT\,\ln K_{\text{eq}}.$$

The *equilibrium constant* K_{eq} is related to the standard (1 molar) free-energy change by

$$\Delta G^0 = -RT\,\ln K_{\text{eq}},$$

$$K_{\text{eq}} = \frac{[C]^c[D]^d}{[A]^a[B]^b}.$$

Many biochemical processes in the body receive free energy from the change of adenosine triphosphate (ATP) to

adenosine diphosphate (ADP) plus inorganic phosphate (P_i). This reaction involves a decrease of free energy. The energy is provided initially by forcing the reaction to go in the other direction to make an excess of ATP. One way this is done is through a very complicated series of chemical reactions known as the *respiration of glucose*. The net effect of these reactions is[13]

$$\text{glucose} + 6O_2 \rightarrow 6CO_2 + 6H_2O, \quad \Delta G^0 = -680\,\text{kcal},$$

$$36\text{ADP} + 36P_i \rightarrow 36\text{ATP} + 36H_2O, \quad \Delta G^0 = +263\,\text{kcal}\,.$$

The decrease in free energy of the glucose more than compensates for the increase in free energy of the ATP. The creation of glucose or other sugars is the reverse of the respiration process and is called *photosynthesis*. The free energy required to run the reaction the other direction is supplied by light energy.

3.18 The Chemical Potential of a Solution

We now consider a binary solution of *solute* and *solvent* and how the chemical potential changes as these two substances are intermixed.[14] This is a very fundamental process that will lead us to the logarithmic dependence of the chemical potential on solute concentration that we saw in Sect. 3.13, as well as to an expression for the chemical potential of the solvent that we will need in Chap. 5.

To avoid having the subscript s stand for both solute and solvent, we call the solvent water. The distinction between solute and water is artificial; the distinction is usually that the concentration of solute is quite small. We need the entropy change in a solution when N_s solute molecules, which initially were segregated, are mixed with N_w water molecules. We make the calculation for an ideal solution—one in which the total volume of water molecules does not change on mixing and in which there is no heat evolved or absorbed on mixing. This is equivalent to saying that the solute and water molecules are the same size and shape, and that the force between a water molecule and its neighbors is the same as the force between a solute molecule and its neighbors.[15] The resulting entropy change is called the *entropy of mixing*.

To calculate the entropy of mixing, imagine a system with N sites, all occupied by particles. The number of microstates is the number of different ways that particles can be placed in the sites. The first particle can go in any site. The second can go in any of $N-1$ sites, and so forth. The total number of different ways to arrange the particles is $N!$ But if the particles are identical, these states cannot be distinguished, and there is actually only one microstate. The number of microstates is $N!/N!$, where the $N!$ in the numerator gives the number of arrangements and the $N!$ in the denominator divides by the number of indistinguishable states.[16]

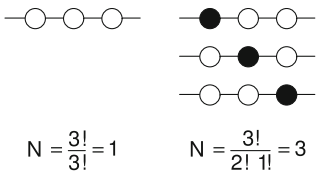


Fig. 3.15 The system on the *left* contains three water molecules. Because they are indistinguishable there is only one way they can be arranged. The system on the *right* contains two water molecules and one solute molecule. Three different arrangements are possible. In each case the number of arrangements in given by $(N_w + N_s)!/(N_w!N_s!)$

Suppose now that we have two different kinds of particles. The total number is $N = N_w + N_s$, and the total number of ways to arrange them is $(N_w+N_s)!$. The N_w water molecules are indistinguishable, so this number must be divided by $N_w!$. Similarly it must be divided by $N_s!$. Therefore, purely because of the ways of arranging the particles, the number of microstates Ω in the mixture is $(N_w + N_s)!/\,(N_w!\,N_s!)$. An example of counting microstates is shown in Fig. 3.15.

There could also be dependence on volume and energy; in fact, the dependence on volume and energy may also contain factors of N_w and N_s. However, our assumption that the molecules of water and solute have the same size, shape, and forces of interaction ensures that these dependencies will not change as solute molecules are mixed with water molecules. The only entropy change will be the entropy of mixing.

The entropy change of the mixture relative to the entropy of N_w molecules of pure water and N_s molecules of pure solute is

$$S_{\text{solution}} - S_{\substack{\text{pure water,}\\ \text{pure solute}}} = k_B \ln\left(\frac{\Omega_{\text{solution}}}{\Omega_{\substack{\text{pure water,}\\ \text{pure solute}}}}\right). \tag{3.66}$$

Since with our assumptions Ω is unity for the pure solute and the pure water, the entropy difference is

$$\begin{aligned} &S_{\text{solution}} - S_{\substack{\text{pure water,}\\ \text{pure solute}}} \\ &= k_B \ln\left(\frac{(N_w + N_s)!}{N_w!\,N_s!}\right) \\ &= k_B \left\{\ln\left[(N_w + N_s)!\right] - \ln(N_w!) - \ln(N_s!)\right\}. \end{aligned} \tag{3.67}$$

This is symmetric in water and solute, and it is valid for any number of molecules.

[13] There are multiple pathways in glucose respiration. The 36 is approximate.

[14] See also Hildebrand and Scott (1964), p. 17 and Chap. 6.

[15] Extensive work has been done on solutions for which these assumptions are not true. See Hildebrand and Scott (1964); Hildebrand et al. (1970).

[16] The fact that there is only one microstate because of the indistinguishability of the particles is called the *Gibbs paradox* (Reif 1965).

Since we usually deal with large numbers of molecules and factorials are difficult to work with, let us use Stirling's approximation (Appendix I) to write

$$S_{\text{solution}} - S_{\substack{\text{pure water,}\\ \text{pure solute}}} = k_B\left[(N_w + N_s)\ln(N_w + N_s) - N_w \ln N_w - N_s \ln N_s\right]. \tag{3.68}$$

The next step is to relate the entropy of mixing to the chemical potential. This is done by recalling the definition of the Gibbs free energy, (Eq. 3.62): $G = U + pV - TS$. The sum of the first two terms, $H = U + pV$, is called the *enthalpy*. Any change of the enthalpy is the heat of mixing; in our case it is zero. (The present case is actually more restrictive: p, V, and U are all constant.) Therefore, since T is also constant, the change in Gibbs free energy is due only to the entropy change:

$$\Delta G = -T\,\Delta S = k_B T\left[N_w \ln\left(\frac{N_w}{N_w + N_s}\right) + N_s \ln\left(\frac{N_s}{N_w + N_s}\right)\right].$$

This is still symmetric with water and solute, but it diverges if either N_w or N_s is zero, because of our use of Stirling's approximation.

We now need an expression that relates the change in G to the chemical potential. This can be derived for the general case using the following thermodynamic arguments. We use Eq. 3.62 to write the most general change in G:

$$dG = dU + p\,dV + V\,dp - T\,dS - S\,dT.$$

The fundamental equation of thermodynamics, Eq. 3.60, generalized to two molecular species, is

$$T\,dS = dU - \mu_w\,dN_w - \mu_s\,dN_s + p\,dV,$$

so

$$dG = \mu_w\,dN_w + \mu_s\,dN_s + V\,dp - S\,dT. \tag{3.69}$$

This can be used to write down some partial derivatives by inspection that are valid in general:

$$\mu_w = \left(\frac{\partial G}{\partial N_w}\right)_{N_s,p,T}, \tag{3.70a}$$

$$\mu_s = \left(\frac{\partial G}{\partial N_s}\right)_{N_w,p,T}, \tag{3.70b}$$

$$V = \left(\frac{\partial G}{\partial p}\right)_{N_s,N_w,T}, \tag{3.70c}$$

$$S = -\left(\frac{\partial G}{\partial T}\right)_{N_s,N_w,p}. \tag{3.70d}$$

To find the chemical potential, we differentiate our expression for G, Eq. 3.69, with respect to N_w and N_s to obtain

$$\mu_w = k_B T\,\ln x_w, \qquad \mu_s = k_B T\,\ln x_s. \tag{3.71}$$

These have been written in terms of the *mole fractions* or *molecular fractions*

$$x_w = \frac{N_w}{N_w + N_s}, \qquad x_s = \frac{N_s}{N_w + N_s}. \tag{3.72}$$

Each chemical potential is zero when the mole fraction for that species is one (i.e., the pure substance). The expressions for μ diverge for x_w or x_s close to zero because of the failure of Stirling's approximation for small values of x.

The last step is to write the chemical potential in terms of the more familiar concentrations instead of mole fractions. We can write the change in chemical potential of the *solute* as the concentration changes from a value C_1 to C_2 as

$$\Delta\mu_s = \mu_s(2) - \mu_s(1) = k_B T\,\ln(x_2/x_1).$$

As long as the solute is dilute, $N_w + N_s \approx N_w$, so $x_2/x_1 = C_2/C_1$ and

$$\Delta\mu_s = k_B T\,\ln(C_2/C_1),$$

which agrees with Eq. 3.48.

The change in chemical potential of the *water* can be written in terms of the *solute* concentration. Since $x_w + x_s = 1$, $\mu_w = k_B T \ln(1 - x_s)$. For small values of x_s the logarithm can be expanded in a Taylor's series (Appendix D):

$$\ln(1 - x_s) = -x_s - \frac{1}{2}x_s^2 - \cdots.$$

The final result is

$$\mu_w = -k_B T\,x_s = -k_B T\,N_s/(N_s + N_w) \approx -k_B T\,(N_s/V)/(N_w/V),$$

or

$$\mu_w \approx -k_B T\,\frac{C_s}{C_w}. \tag{3.73a}$$

To reiterate: this is the chemical potential of the water for small solute concentrations. The zero of chemical potential is pure water. The term is negative because the addition of solute decreases the chemical potential of the water, due to the entropy of mixing term. For a change of solute concentration, the chemical potential of the water changes by

$$\Delta\mu_w = -\frac{k_B T\,\Delta C_s}{C_w}. \tag{3.73b}$$

We now know the concentration dependence of the chemical potential. In Chap. 5 we will be concerned with the movement of solute and water, and we will also need to know

the dependence of the chemical potentials on pressure. To find this, we write

$$\Delta\mu_w = \left(\frac{\partial\mu_w}{\partial p}\right)_{T,N_w,C_s} \Delta p + \left(\frac{\partial\mu_w}{\partial C_s}\right)_{T,p,N_w} \Delta C_s.$$

The second term is just Eq. 3.73b. To obtain the derivative in the first term, we use the fact that when the partial derivative of a function is taken with respect to two variables, the result is independent of the order of differentiation (Appendix N):

$$\left[\frac{\partial}{\partial p}\left(\frac{\partial G}{\partial N_w}\right)_{T,p,N_s}\right]_{T,N_w} = \left[\frac{\partial}{\partial N_w}\left(\frac{\partial G}{\partial p}\right)_{T,N_w,N_s}\right]_{T,p}$$

From Eqs. 3.70a and 3.70c, we get

$$\left(\frac{\partial\mu_w}{\partial p}\right)_{T,N_w} = \left(\frac{\partial V}{\partial N_w}\right)_{T,p}. \tag{3.74}$$

For a process at constant temperature, the rate of change of μ_w with p for constant solute concentration is the same as the rate of change of V with N_w when p is fixed.

The quantity $(\partial V/\partial N_w)_{T,p}$ is the rate at which the volume changes when more molecules are added at constant temperature and pressure. For an ideal incompressible liquid it is the molecular volume, $\overline{V}_w$. We can repeat this argument for the solute to obtain

$$\left(\frac{\partial\mu_w}{\partial p}\right)_{T,N_w} = \overline{V}_w, \qquad \left(\frac{\partial\mu_s}{\partial p}\right)_{T,N_s} = \overline{V}_s. \tag{3.75}$$

In a solution, the total volume is $V = N_w\overline{V}_w + N_s\overline{V}_s$ where $\overline{V}_w$ and $\overline{V}_s$ are the average volumes occupied by one molecule of water and solute. Dividing by V gives $1 = C_w\overline{V}_w + C_s\overline{V}_s$. If the solution is dilute,

$$\overline{V}_w \approx \frac{1}{C_w}. \tag{3.76}$$

In an ideal solution $\overline{V}_w = \overline{V}_s$. For an ideal dilute solution, we then have

$$\Delta\mu_w = \overline{V}_w(\Delta p - k_BT\ \Delta C_s) \approx \frac{\Delta p - k_BT\ \Delta C_s}{C_w}. \tag{3.77}$$

$$\begin{aligned}\Delta\mu_s &= k_BT\ \ln(C_{s2}/C_{s1}) + \overline{V}_s\ \Delta p \\ &\approx k_BT\ \ln(C_{s2}/C_{s1}) + \overline{V}_w\ \Delta p.\end{aligned} \tag{3.78}$$

We saw this concentration dependence earlier, in Sect. 3.13. If the concentration difference is small, we can write $C_{s2} = C_{s1} + \Delta C_s$ and use the expansion $\ln(1+x) \approx x$ to obtain

$$\Delta\mu_s \approx \frac{k_BT\ \Delta C_s}{C_s} + \frac{\Delta p}{C_w}. \tag{3.79}$$

3.19 Transformation of Randomness to Order

When two systems are in equilibrium, the total entropy is a maximum. Yet a living creature is a low-entropy, highly ordered system. Are these two observations in conflict? The answer is no; the living system is not in equilibrium, and it is this lack of equilibrium that allows the entropy to be low. The conditions under which order can be brought to a system—its entropy can be reduced—are discussed briefly in this section.

A car travels with velocity $\mathbf{v}$ and has kinetic energy $\frac{1}{2}mv^2$. In addition to the random thermal motions of the atoms making up the car, all the atoms have velocity $\mathbf{v}$ in the same direction (except for those in rotating parts, which have an ordered velocity that is more complicated to describe). If the brake shoes are brought into contact with the brake drums, the car loses kinetic energy, and the shoes and drums become hot. Ordered energy has been converted into disordered, thermal energy; the entropy has increased. Is it possible to heat the drums and shoes with a torch, apply the brakes, and have the car move as the drums and shoes cool off? Energetically, this is possible, but there are only a few microstates in which all the molecules are moving in a manner that constitutes movement of the car. Their number is vanishingly small compared to the number of microstates in which the brake drums are hot. The probability that the car will begin to move is vanishingly small.

An animal is placed in an insulated, isolated container. The animal soon dies and decomposes. Energetically, the animal could form again, but the number of microstates corresponding to a live animal is extremely small compared to all microstates corresponding to the same total energy for all the atoms in the animal.

In some cases, thermal energy can be converted into work. When gas in a cylinder is heated, it expands against a piston that does work. Energy can be supplied to an organism and it lives. To what extent can these processes, which apparently contradict the normal increase of entropy, be made to take place? These questions can be stated in a more basic form.

1. To what extent is it possible to convert internal energy distributed randomly over many molecules into energy that involves a change of a macroscopic parameter of the system? (How much work can be captured from the gas as it expands the piston?)
2. To what extent is it possible to convert a random mixture of simple molecules into complex and highly organized macromolecules?

Both these questions can be reformulated: under what conditions can the entropy of a system be made to decrease?

The answer is that the entropy of a system can be made to decrease if, and only if, it is in contact with one or more auxiliary systems that experience at least a compensating increase in entropy. Then the total entropy remains the same or

increases. This is one form of the *second law of thermodynamics*. For a fascinating discussion of the second law, see Atkins (1994).

One device that can accomplish this process is a *heat engine*. It operates between two thermal reservoirs at different temperatures, removing heat from the hotter one and injecting heat into the cooler one. Even though less heat goes into the cooler reservoir than was removed from the hotter one (the difference being the mechanical work done by the engine), the overall entropy of the two reservoirs increases. The entropy change of the hot reservoir is a decrease, $-\Delta Q/T$, while the entropy change of the cooler reservoir is an increase, $+\Delta Q'/T'$. Since $T' < T$, the entropy increase more than balances the decrease, even though $\Delta Q' < \Delta Q$. The increase in the number of accessible microstates of the cooler reservoir is greater than the decrease in the number of accessible microstates of the hotter reservoir. The coupled chemical reactions that we saw in Sect. 3.17 are analogous.

Symbols Used in Chap. 3

Symbol	Use	Units	First used page
$\mathbf{a}$	Acceleration	m s^{-2}	53
a	Number of atoms in a molecule		57
a, b, c, d	Number of atoms of species A, B, C, and D		71
a	Area	m^2	69
c_j	Concentration (molar)	mol m^{-3}, mol l^{-1}	67
c	Specific heat capacity	J K^{-1} kg^{-1}	65
e	Elementary charge	C	63
f	Number of degrees of freedom		57
g	Gravitational acceleration	m s^{-2}	64
k_B	Boltzmann's constant	J K^{-1}	61
m	Mass	kg	53
n	Number of particles in a volume		54
p	Probability of "success"		55
p	Pressure	Pa	60
p_x, p_y, p_z	Momentum	kg m s^{-1}	65
q	Probability of "failure"		55
q	Electric charge	C	69
t	Time	s	53
u_i	Energy of the ith energy level	J	58
v, v'	Volume	m^3	55
v	Electrical potential	V	63
v, v_x, v_y, v_z	Velocity	m s^{-1}	53
x, y, z	Position coordinate	m	53
x	General variable		59
x	Extensive variable		69
x_s, x_w	Mole fractions of solute and water		73
y	General variable		62
y	Height	m	64
z	Valence		63
A, A', A^*	Thermodynamic systems		60
A, B, C, D	Chemically reacting species		71
C_i, C	Concentration (particles per volume)	m^{-3}, l^{-1}	63
C	Heat capacity	J K^{-1}	65
E_k	Kinetic energy	J	63
E_p	Potential energy	J	63
$\mathbf{F}, F$	Force	N	53
F	Faraday constant	C mol^{-1}	64
G	Ratio of accessible microstates in a small system		62
G	Gibbs free energy	J	70
H	Enthalpy	J	73
K_{eq}	Equilibrium constant in a chemical reaction		71
M	Number of molecules in a system		57
M	Number of repeated measurements		60
N, N', N^*	Number of particles		54
N_w, N_s	Number of solvent (water) or solute molecules		72
N_A	Avogadro's number	mol^{-1}	64
N_A, N_B, N_C, N_D	Number of molecues of species A, B, C, and D consumed or produced in a chemical reaction		71
P	Probability		54
Q	Flow of heat to a system	J	58
R	Ratio of accessible states in a reservoir (Boltzmann factor)		63
R	Gas constant	J mol^{-1} K^{-1}	64
S	Area	m^2	64
S, S', S^*	Entropy	J K^{-1}	62
T	Absolute temperature	K	61
U, U', U^*	Total energy of a system	J	58
V	Volume	m^3	55
$\overline{V}_w, \overline{V}_s$	Volume of water or solute molecule	m^3	74
W	Work done by a system on the surroundings	J	58
W_{conc}	Work done on a system to increase the concentration	J	67
X	Generalized force		69
α	General variable		63
ρ	Density	kg m^{-3}	65
σ	Surface tension	N m^{-1}	69
τ	$k_B T$	J	61
μ	Chemical potential	J molecule^{-1}	66
μ_w, μ_s	Chemical potential of water or solute	J molecule^{-1}	73
$\Omega, \Omega', \Omega^*$	Number of accessible microstates		59
–	An overscript bar means an average over an ensemble		56
$\langle\rangle$	Angular brackets mean an average over time		56

Problems

Section 3.1

Problem 1. Some systems are so small that only a few molecules of a particular type are present, and statistical arguments begin to break down. Estimate the number of hydrogen ions inside an *E. coli* bacterium with $pH = 7$. (When $pH = 7$ the concentration of hydrogen ions is 10^{-7} mol l^{-1}.)

Problem 2. Use the last column of Table 3.2 to calculate the average value of n, which is defined by $\overline{n} = \sum nP(n)$. Verify that $\overline{n} = Np$ in this case.

Problem 3. A loose statement is made that "if we throw a coin 1 million times, the number of heads will be very close to half a million." What is the mean number of occurrences of heads in 1 million tries? What is the standard deviation? What does "very close" mean? (You may need to consult Appendices G and H.)

Problem 4. Evaluate $P(n; 4, 0.5)$ using Eq. 3.4. Check your results against the histogram of Fig. 3.2 and by listing all the possible arrangements of four particles in the left or right sides of the box.

Problem 5. Write a computer program to simulate measurements of which half of a box a gas molecule is in. Make several measurements with different sets of random numbers, and plot a histogram of the number of times n molecules are found in the left half. Try this for $N = 1$, 10, and 100. Many computer languages have a built-in routiine to generate random numbers. For a discussion of how to construct and use random number generators, see Press et al. (1992).

Problem 6. Color blindness is a sex-linked defect. The defective gene is located in the X chromosome. Females carry an XX chromosome pair, while males have an XY pair. The trait is recessive, which means that the patient exhibits color blindness only if there is no normal X gene present. Let X_d be a defective gene. Then for a female, the possible gene combinations are

$$XX, XX_d, X_dX_d.$$

For a male, they are

$$XY, X_dY.$$

In a large population about 8 % of the males are color-blind. What percentage of the females would you expect to be color-blind?

Problem 7. A patient with heart disease will sometimes go into *ventricular fibrillation*, in which different parts of the heart do not beat together, and the heart cannot pump. This is cardiac arrest. The following data show the fraction of patients failing to regain normal heart rhythm after attempts at ventricular defibrillation by electric shock (Weaver 1982).

Number of attempts	Fraction persisting in fibrillation
0	1.00
1	0.37
2	0.15
3	0.07
4	0.02

Assume that the probability p of defibrillation on one attempt is independent of other attempts. Obtain an equation for the probability that the patient remains in fibrillation after N attempts. Compare it to the data and estimate p.

Problem 8. There are N people in a class ($N = 25$). What is the probability that no one in the class has a birthday on a particular day? Ignore seasonal variations in birth rate and ignore leap years.

Problem 9. The death rate for 75-year-old people is 0.089 per year (Commissioners 1941 Standard Ordinary Mortality Table).

(a) What is the probability that an individual aged 75 will die during a 12-h period? Neglect the fact that some are sick, some are terminally ill, and so on, and assume that the probability is the same for everyone.

(b) Suppose that 10 000 people, all aged 75, are given the flu vaccine at $t = 0$. What is the probability that none will die during the next 12 h? (This underestimates the probability, since sick people would not be given the vaccine, but they are included in the death rate.)

Problem 10. This problem is intended to help you understand some of the nuances of the binomial probability distribution.

(a) In a macabre "game" of "roulette" the victim places one bullet in the cylinder of a revolver. (A less hazardous game could be done with dice.) There is room for six bullets in the cylinder. The victim spins the cylinder, so there is a probability of 1/6 that the bullet is in firing position. The victim then places the gun to the head and fires. If the victim survives, the cylinder is spun again and the process is repeated. We can look either at the cumulative probability of "success" (being killed), or the cumulative probability of "failure" (surviving). Make a table for 1000 victims who keep playing the game over and over. Plot the number surviving, the number killed on each try, and the cumulative number killed.

(b) Show that the number surviving can be expressed as $1000e^{-bN}$, where N is the number of tries, and find b.

(c) The data in the following table are from Schwartz and Mayaux (1982). They show the cumulative success rates in different age groups for patients being treated for infertility by artificial insemination from a donor. That is, each month at the time of ovulation each patient who has not yet become pregnant is inseminated artificially.

The table shows the fraction of patients who have become pregnant at the end of each cycle. Plot these data. What do they suggest? Make whatever plots can confirm or rule out what you suspect.

Cycle	Age ≤ 25	Age $\geqslant 35$
0	0	0
1	0.11	0.03
2	0.23	0.14
3	0.30	0.20
4	0.39	0.27
5	0.44	0.35
6	0.51	0.35
7	0.55	0.36
8	0.63	0.39
9	0.65	0.43
10	0.67	0.43
11	0.70	0.46
12	0.74	0.54

Problem 11. Use Eq. 3.4 to verify that the probability of all 80 particles being in the left half of the box is approximately 10^{-24}.

Problem 12. Appendix H describes how to calculate the magnitude of fluctuations for N particles in a box (the standard deviation). Calculate $\langle n \rangle$ and its standard deviation for $N = 80$ and $p = q = 0.5$. Estimate the value of N in Fig. 3.4.

Section 3.3

Problem 13. A thermally insulated ideal gas of particles is confined within a container of volume V. The gas is initially at absolute temperature T. The volume of the container is very slowly reduced by moving a piston that constitutes one wall of the container. Give qualitative answers to the following questions.

(a) What happens to the energy levels of each particle?
(b) Is the work done on the gas as its volume decreases positive or negative?
(c) What happens to the energy of the gas?

Section 3.5

Problem 14. Suppose you have a system with 10 particles and three energy levels. The particles are distributed among the levels as follows: 5 particles are in the level with energy 0, three particles are in the level with energy $2E$, and two in the level with energy $4E$. An interaction with the surroundings occurs in which work is done on the system and heat flows out of the system in such a way that $\Delta U = 0$. The work causes the energy of each level to rise by an amount E.

(a) Draw a picture like Fig. 3.7 showing the new levels and the distribution of particles among the levels before and after the interaction.
(b) Calculate the average energy of the particles before and after the interaction.
(c) Draw a picture like that in Fig. 3.8 appropriate for this system.

Problem 15. System A has 10^{20} microstates, and system A' has 10^{19} microstates. How many microstates does the combined system have?

Problem 16. Calculate the Celsius and absolute temperatures corresponding to a room temperature of 68 °F, a normal body temperature of 98.6 °F, and a febrile body temperature of 104 °F.

Problem 17. Calculate and plot Ω, Ω', and Ω^* for Fig. 3.10, thus reproducing the figure. Write down an analytic expression for Ω^* and differentiate to find the value of U for which Ω^* is a maximum.

Problem 18. Let $\Omega(U) = 5U^2 + 1$, $\Omega'(U') = U' + 1$, and $U + U' = 100$. Make plots like those in Fig. 3.10 for this system and determine the most probable value of U.

Problem 19. Systems A and A' each consist of 3 particles, whose energy levels are u, $2u$, $3u$, etc. The total energy available to the combined system is $U^* = 12u$.

(a) Make a table similar to Table 3.3. (If you have difficulty, see part (d) of this problem.)
(b) Find the most probable state. To what values of U and U' does it correspond?
(c) Plot Ω^* vs U. What is the probability that all three particles in system A have energy u?
(d) Consider system A. If it has energy U, the maximum energy the first particle can have is $U - 2u$. How many microstates are there for which the first particle has energy $U - 2u$? $U - 3u$? Show that the total number of microstates for system A is given by

$$\sum_{i=1}^{U/u-2} \left(\frac{U}{u} - i - 1 \right) = \frac{1}{2} \left[\left(\frac{U}{u} \right)^2 - 3 \left(\frac{U}{u} \right) + 2 \right].$$

This proves the assertion in the text that for 3 particles, Ω increases as U^2.

Problem 20. We have seen that in general with volume, number of particles, and other parameters that determine the positions of the energy levels held fixed,

$$\frac{1}{\Omega} \frac{d\Omega}{dU} = \frac{1}{k_B T}.$$

Suppose that $U = CT$, where C is the heat capacity of the system. Find $\Omega(U)$.

Problem 21. Systems A and A' are in thermal contact. Show that if $T < T'$, energy flows from A' to A to increase Ω^*, while if $T > T'$, energy flows from A to A'.

Problem 22. A simple system has *only two* energy levels for each single entity in the system. (The system could, for example, be a collection of "gates" in a cell membrane, each

with two states, open and closed.) One level has energy u_1, the other has energy u_2. There are N entities in the system. You can answer the following questions without doing any calculations.

(a) What is the minimum energy of the system? How many microstates are there for the minimum energy?
(b) What is the maximum energy of the system? How many microstates are there for which the system has maximum energy?
(c) Sketch what $\Omega(U)$ must look like.
(d) Recall the definition of T, Eqs. 3.14 and 3.15. Are there any values of U for which the temperature is negative? Where?

Section 3.6

Problem 23. Calculate the temperature (in K) and entropy (in eV K^{-1}) of system A' in Fig. 3.10 at equilibrium. Assume U and U' are given in electron volts. Your values may seem odd because this example is not biologically realistic.

Problem 24. Consider the following arrangements of the 26 capital letters of the English alphabet: (a) TWO, (b) any three letters, in any order, that are all different, and (c) any three letters, in any order, which may repeat themselves. For (b) and (c), consider the same letters in a different order to be a different arrangement. If each arrangement is a "microstate," find Ω and S in each case.

Problem 25. Ice and water coexist at 273 K. To melt 1 mol of ice at this temperature, 6 000 J are needed. Calculate the entropy difference and the ratio of the number of microstates for 1 mol of ice and 1 mol of water at this temperature. Do not worry about any volume changes of the ice and water.

Problem 26. If a system is maintained at constant volume, no work is done on it as the energy changes. In that case $dU = C(T)\,dT$, where U is the internal energy, C is the heat capacity, and T is the temperature. The heat capacity in general depends on the temperature. Suppose that in some temperature region the heat capacity varies linearly with temperature: $C(T) = C_0 + DT$.

(a) What is the entropy change of the system when it is heated from temperature T_1 to temperature T_2, both of which are in the region where $C(T) = C_0 + DT$?
(b) What is the ratio of the number of microstates at T_2 to the number at T_1?

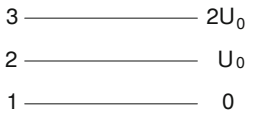
Problem 27. A substance melts at constant temperature. There are 7 times as many microstates accessible to each molecule of the liquid as there were to each molecule of the solid. Ignore volume changes.

(a) What is the change in entropy of each molecule?
(b) How much heat is required to melt a mole of the substance if the melting temperature is 50 °C?

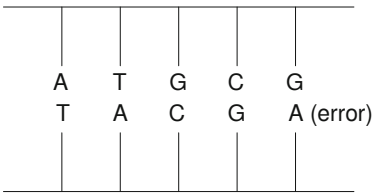
Problem 28. The entropy of a monatomic ideal gas at constant energy depends on the volume as $S = Nk_B \ln V + \text{const.}$ A gas of N molecules undergoes a process known as a free expansion. Initially it is confined to a volume V by a partition. The partition is ruptured and the gas expands to occupy a volume $2V$. No work is done and no heat flows, so the total energy is unchanged. Calculate the change of entropy and the ratio of the number of microstates after the volume change to the number before.

Section 3.7

Problem 29. A pore has three configurations with the energy levels shown. The pore is in thermal equilibrium with the surroundings at temperature T. Find the probabilities p_1, p_2, and p_3 of being in each level. Each level has only one microstate associated with it.



Problem 30. The DNA molecule consists of two intertwined linear chains. Sticking out from each monomer (link in the chain) is one of four bases: adenine (A), guanine (G), thymine (T), or cytosine (C). In the double helix, each base from one strand bonds to a base in the other strand. The correct matches, A–T and G–C, are more tightly bound than are the improper matches. The chain looks something like this, where the last bond shown is an "error."



The probability of an error at 300 K is about 10^{-9} per base pair. Assume that this probability is determined by a Boltzmann factor e^{-U/k_BT}, where U is the additional energy required for a mismatch.

(a) Estimate this excess energy.
(b) If such mismatches are the sole cause of mutations in an organism, what would the mutation rate be if the temperature were raised 20 °C?

Problem 31. In Chap. 18 we will study how the "spin" magnetic moment of an atomic nucleus interacts with a magnetic field **B**, leading to magnetic resonance imaging. Assume a nucleus has a *magnetic dipole moment* μ, which can point in only one of two directions: parallel to B ("spin up") or antiparallel ("spin down"). The energy of a nucleus with spin up is $-\mu B$; with spin down it is $+\mu B$. Use the Boltzmann factor to determine an expression for the ratio of the number of particles with spin up to the number with spin down.

Evaluate this ratio for $\mu = 1.4 \times 10^{-26}\,\mathrm{J\,T^{-1}}$, $B = 2\,\mathrm{T}$, and $T = 300\,\mathrm{K}$.

Problem 32. The data of Problem 2.10 were used to obtain an empirical relationship between the charge integration time τ and the temperature T. It might be that τ is determined by a chemical reaction whose rate is given by a Boltzmann factor. Make a new plot based on that assumption and determine the appropriate constants.

Problem 33. Oxygen and carbon monoxide compete for binding to hemoglobin. If enough CO binds to hemoglobin, the ability of the blood to deliver oxygen is impaired, and carbon monoxide poisoning ensues. Consider the hemoglobin molecule to be a two-state system: the heme group is bound either to O_2 or to CO. Calculate the probability of binding to CO. Let the G factor of Eq. 3.25 be equal to the ratio of the concentrations of CO and O_2. Assume CO is 100 times less abundant than O_2. CO is more tightly bound than O_2 to the heme group by about 0.15 eV. Let $T = 300$ K.

Problem 34. The function of many enzymes is to act as a *catalyst*: they increase the speed of a chemical reaction. To get an idea of how a catalyst works, consider the reaction

$$\text{enzyme} + \text{substrate} \rightarrow \text{enzyme} + \text{product}.$$

In order for the reaction to proceed, some energy barrier ΔE must be overcome. The probability of the substrate having an energy ΔE or greater depends primarily on a Boltzmann factor, $e^{-\Delta E/k_B T}$. Determine by what factor this probability increases if the enzyme decreases the activation energy by (a) 0.1 eV, (b) 1 eV. Assume $T = 310\,\mathrm{K}$.

Problem 35. Chemists use Q_{10} to characterize a chemical reaction. It is defined by

$$Q_{10} = \frac{(\text{reaction rate at } T + 10)}{(\text{reaction rate at } T)},$$

where T is the absolute temperature. If the reaction rate is proportional to the fraction of reacting atoms that have an energy exceeding some threshold ΔU, then to a first approximation

$$R \propto \int_{\Delta U}^{\infty} e^{-U/k_B T}\, dU.$$

(This neglects more slowly varying factors such as a $U^{1/2}$ which are introduced in more accurate analyses.)

(a) Show that $R \propto k_B T e^{-\Delta U/k_B T}$.

(b) Show that

$$\frac{Q_{10} T}{T + 10} = \exp\left[\frac{\Delta U}{k_B}\frac{10}{T(T + 10)}\right].$$

(c) Estimate ΔU if $Q_{10} = 2$ at $T = 300$ K.

Problem 36. The vapor pressure of a substance can be calculated using the following model. All molecules in the vapor that strike the surface of the liquid stick. (This number is proportional to the pressure.) Those molecules in the liquid that reach the surface and have enough energy escape. Equilibrium is established when the number sticking per unit area per unit time is equal to the number escaping.

(a) The number of molecules with energy U is proportional to $e^{-U/k_B T}$. What will be the number with energy greater than the escape energy, U_0?

(b) Use the result of part (a) and look up values for the vapor pressure of water as a function of temperature, to make a plot on semilog paper. From this plot, estimate the escape energy U_0.

(c) The "heat of vaporization" of water is 540 cal per g. Convert the energy per molecule you found in part (b) to calories per gram and compare it with this figure.

Problem 37. A macromolecule of density ρ and mass m is immersed in an incompressible fluid of density ρ_w at temperature T. The volume v occupied by one macromolecule is known. A dilute solution of the macromolecules is placed in an ultracentrifuge rotating with high angular velocity ω. In the frame of reference rotating with the centrifuge, a particle at rest is acted on by an outward force $m\omega^2 r$, where r is the distance of the particle from the axis.

(a) What is the net force acting on the particle in this frame? Include the effect of buoyancy of the surrounding fluid, of density ρ_w.

(b) Suppose that equilibrium has been reached. Use the Boltzmann factor to find the number of particles per unit volume at distance r.

Problem 38. Suppose that particles in water are subjected to an external force $F(y)$ that acts in the y direction. The force is related to the potential energy $E_p(y)$ by $F = -dE_p/dy$. Neglect gravity and buoyancy effects.

(a) Apply Newton's first law to a slice of the fluid in equilibrium to obtain an expression for $p(y)$.

(b) If the particles have a Boltzmann distribution, show that $p(y) - p(0) = k_B T\,\big[C(y) - C(0)\big]$.

Section 3.8

Problem 39. The concentrations of various ions are measured on the inside and outside of a nerve cell. The following values are obtained when the potential inside the cell is −70 mV with respect to the outside.

Ion	Inside (mmol $\mathrm{l^{-1}}$)	Outside (mmol $\mathrm{l^{-1}}$)
Na^+	15	145
K^+	150	5
Cl^-	9	125

Comment on which species have concentrations that are consistent with being able to pass freely through the cell wall. Assume $T = 300$ K.

Problem 40. Calculate the volume of 1 mole of water in liters. Pour yourself a mole of water and drink it. Calculate the concentration of water in moles per liter.

Section 3.9

Problem 41. A virus has a mass of 1.7×10^{-14} g. If the virus particles are in thermal equilibrium in the atmosphere, their concentration will vary with height as $C(y) = C(0)e^{-y/\lambda}$. Evaluate λ. Do you think this answer is reasonable?

Problem 42. Calculate the length constant λ for the exponential decay ($e^{-y/\lambda}$) of atmospheric pressure. Assume the atmosphere is made up entirely of nitrogen, N_2. Nitrogen has an atomic weight of 14. Use your result to compare air pressure at sea level to air pressure at the top of Mt. Everest (8.8 km). Assume the atmosphere is all at the same temperature; it is not.

Section 3.10

Problem 43. Use Appendix K to verify the expressions given for the integrals in the numerator and denominator of Eq. 3.37.

Problem 44. Calculate the average kinetic energy (in J and eV) of a particle moving in three dimensions at body temperature, 37 °C .

Problem 45. This is our first model for the important problem of detecting a "signal" in the presence of "noise." We will discuss this in detail in Chapters 9 and 11. A sensitive balance consists of a weak spring hanging vertically in the earth's gravitational field. The equilibrium position of the end of the spring is $x = 0$. When a mass m is added to the spring, it elongates to an average position x_0, around which it vibrates because of thermal energy. In terms of $\Delta x = x - x_0$, the momentum of the mass p_x and the spring constant K, the force which the spring exerts on the mass is Kx_0, and the total energy is $U = p_x^2/2m + \frac{1}{2}K(\Delta x)^2$.

(a) What is x_0 in terms of m, g, and K?
(b) Find $\overline{\Delta x^2} = \overline{(x - x_0)^2}$.
(c) What is the smallest mass that can be measured taking a single "snapshot" of the system to find the position of the mass?

Section 3.11

Problem 46. The specific heat capacity of water is $4\,184\,\mathrm{J\,K^{-1}\,kg^{-1}}$ (Denny 1993). Convert this to cal $\mathrm{g^{-1}\,°C^{-1}}$. Historically, the calorie was defined in terms of the specific heat capacity of water.

Problem 47. The "Calorie" we see listed on food labels is actually 1 000 cal or 1 kcal. How many kcal do you expend each day if your average metabolic rate is 100 W?

Problem 48. Your body must dissipate energy from metabolism at a rate of about 100 W by various mechanisms to keep the body from overheating. Suppose these mechanisms stopped working (perhaps you are wrapped in a very good thermal blanket, so no heat can flow from or to your body). At what rate will your body temperature increase? How long will it take for your body temperature to increase by 5 °C? Assume you have a mass of 70 kg, and the specific heat of your body tissue is the same as of water, $4\,200\,\mathrm{J\,K^{-1}\,kg^{-1}}$.

Problem 49. A person of mass 70 kg and body temperature 37 °C breathes in 0.5 l of air at a temperature of 20 °C. Assume that there are no other sources of heat (turn off metabolism for a moment), and the body as a whole is insulated so no heat is lost to the environment. Find the equilibrium temperature that the air and body will ultimately attain. Useful data: $\rho_{\mathrm{air}} = 1.3\ \mathrm{kg\,m^{-3}}$, $\rho_{\mathrm{water}} = 1000\ \mathrm{kg\,m^{-3}}$, $c_{\mathrm{air}} = 1000\ \mathrm{J\ K^{-1}\ kg^{-1}}$, $c_{\mathrm{water}} = 4200\ \mathrm{J\,K^{-1}\ kg^{-1}}$. Assume that the person's body tissue has the same heat capacity and density as water.

Problem 50. Fish are cold blooded, and "breathe" water (in other words, they extract dissolved oxygen from the water around them using gills). Could a fish be warm blooded and still breathe water? Assume a warm-blooded fish maintains a body temperature that is 20 °C higher than the surrounding water. Furthermore, assume that the blood in the gills is cooled to the temperature of the surrounding water as the fish breathes water. Calculate the energy required to reheat 1 l of blood to the fish's body temperature. One liter of blood carries sufficient oxygen to produce about 4000 J of metabolic energy. Is the energy needed to reheat 1 l of blood to body temperature greater than or less than the metabolic energy produced by 1 l of blood? What does this imply about warm-blooded fish? Why must a warm-blooded aquatic mammal such as a dolphin breathe air, not water? Use $c = 4\,200\,\mathrm{J\ K^{-1}\,kg^{-1}}$ and $\rho = 10^3\,\mathrm{kg\ m^{-3}}$ for both the body and the surrounding water. For more on this topic, see Denny (1993).

Problem 51. Forensic scientists sometimes use *Newton's law of cooling* to determine how long ago a victim died. Assume that at the time of death (t_{death}) the body had a temperature T_{body}, and after death it cools to the temperature of the surroundings, T_{surr}. Assume that the rate of heat loss by the body is proportional to the surface area of the body, S, and the temperature difference $T - T_{\mathrm{surr}}$. The constant of proportionality is called the *convection coefficient*. As the corpse cools, the decrease in temperature is determined by the heat capacity.

(a) Relate the rate of heat loss to the rate of temperature change, and derive a differential equation for the body temperature T.

(b) Solve this differential equation (if you are having trouble, see Sect. 2.8). The solution is Newton's law of cooling.
(c) Write an expression for the time constant of cooling in terms of the specific heat capacity, density, volume, area, and the convection coefficient.
(d) For two bodies with the same shape but different sizes, which will cool faster: the large body or the small one?

Problem 52. Determine whether the specific heat capacity of air, 1000 J K^{-1} kg^{-1} is the same as the molar specific heat capacity of a monatomic ideal gas, $3R/2$. If not, why not? Assume air is all nitrogen, N_2.

Section 3.12

Problem 53. Modify the system shown in Fig. 3.10 so that $\Omega(U, N) = 5U^2N^3$, $\Omega'(U') = 4(U')^2(N')^3$, $U + U' = 6$, and $N + N' = 10$.
(a) Show that this change does not affect the calculation of the temperature.
(b) Plot $\Omega(N)$, $\Omega'(N')$ and $\Omega^*(N)$ over $0 < N, N' < 10$ using the equilibrium value $U = 3\,\text{eV}$.
(c) Find the average value $\langle N \rangle$.
(d) Calculate the chemical potential (in eV) in equilibrium.

Problem 54. A small system A is in contact with a reservoir A' and can exchange both heat and particles with the reservoir. The number of microstates available to system A does not change. Show that the difference in total entropy when A is in two distinct states is

$$\Delta S^* = -(N_1 - N_2)\left(\frac{\partial S}{\partial N}\right)_U - (U_1 - U_2)\left(\frac{\partial S}{\partial U}\right)_N,$$

so that

$$\frac{P(N_1, U_1)}{P(N_2, U_2)} = \frac{e^{(N_1\mu - U_1)/k_BT}}{e^{(N_2\mu - U_2)/k_BT}}.$$

where T and μ are the temperature and chemical potential of the reservoir. This is called the *Gibbs factor*, and it reduces to the Boltzmann factor if $N_1 = N_2$. Chemists use the notation $\lambda = e^{\mu/k_BT}$, where λ is the absolute activity. Then

$$\frac{P(N_1, U_1)}{P(N_2, U_2)} = \frac{\lambda^{N_1}}{\lambda^{N_2}} \frac{e^{-U_1/k_BT}}{e^{-U_2/k_BT}}.$$

Problem 55. Specialize the results of the previous problem to a series of binding sites on a surface, such as a myoglobin molecule. The two states are

No particle bound at the site	$N_1 = 0$,	$U_1 = 0$
One particle bound at the site	$N_2 = 1$,	$U_2 = U_0$

(a) Show that the fraction of sites occupied is

$$f = \frac{\lambda e^{-U_0/k_BT}}{1 + \lambda e^{-U_0/k_BT}}.$$

(b) If the sites are in equilibrium with a gas, then $\mu_{\text{gas}} = \mu_{\text{sites}}$ or $\lambda_{\text{gas}} = \lambda_{\text{sites}}$. From the definition $\mu = -T(\partial S/\partial N)_{U,V}$ and the expression for the entropy of a monatomic ideal gas,

$$S(U, V, N) = Nk_B\left(\ln V + \frac{3}{2}\ln U - \frac{5}{2}\ln N + \frac{5}{2} + c\right),$$

where $c = \frac{3}{2}\ln(m/3\pi\hbar^2)$, show that $f = p/(p_0 + p)$, where p is the gas pressure and

$$p_0 = \frac{(k_BT)^{5/2}\, m^{3/2}\, e^{U_0/k_BT}}{(2\pi\hbar^2)^{3/2}}.$$

This expression fits the data very well. See Rossi-Fanelli and Antonini (1958).

Section 3.13

Problem 56. The entropy of a monatomic ideal gas is

$$S(U, V, N) = Nk_B\left(\ln V + \frac{3}{2}\ln U - \frac{5}{2}\ln N + \frac{5}{2} + c\right),$$

where $c = \frac{3}{2}\ln(m/3\pi\hbar^2)$ depends only on the mass of the molecule. Consider two containers of gas at the same temperature and pressure that can exchange particles. Expand the total entropy in a Taylor's series, keep terms to second order, and use the result to find the variance in the fluctuating number of particles in one system. Assume $N \ll N'$. You should obtain the same result obtained from the binomial distribution ($\sigma^2 = N$) if you take into account that it is the *temperature* of the gas in the container, and not its energy, that should be held fixed. (For a monatomic ideal gas $U = 3Nk_BT/2$. Use this result to rewrite the entropy in terms of T, V, and N.)

Problem 57. Show that the chemical potential of an ideal gas is proportional to the logarithm of the concentration, a result that we have now seen several times for dilute ideal systems. To do so, use the expression for the entropy of a monatomic ideal gas given in the previous problems. Rewrite the thermodynamic identity as $dU = TdS + \mu\, dN - p\, dV$, from which we can identify the partial derivative

$$\mu = \left(\frac{\partial U}{\partial N}\right)_{S,V}.$$

The chemical potential is the increase in energy of the system if one particle is added while keeping the entropy and volume fixed. Use the expression for the entropy of the monatomic ideal gas, for the case of N particles with total energy U and $N + 1$ particles with total energy $U + \mu$, to show that the chemical potential of the ideal gas is

$$\mu = k_BT\left[\ln\left(\frac{N}{V}\right) - \frac{3}{2}\ln(3k_BT/2) - \text{const}\right]$$

or

$$\mu = -k_B T \ln\left[\frac{V}{N}\left(\frac{mk_B T}{2\pi\hbar^2}\right)^{3/2}\right].$$

A more extensive discussion for other simple systems is given by Cook and Dickerson (1995).

Problem 58. Derive the Nernst equation (Eq. 3.34) by making the chemical potential the same on each side of a charged membrane. Use Eq. 3.48, with the potential energy per particle given as zev.

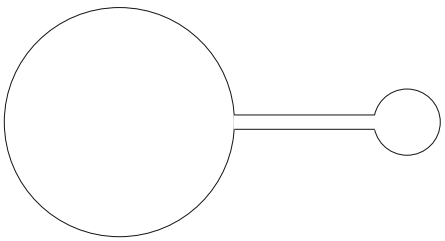
Section 3.15

Problem 59. Consider two systems that can exchange energy U and surface area a, but not volume V or number of particles N. The total energy is $U^* = U + U'$ and the total surface area is $a^* = a + a'$. Repeat the analysis of Sect. 3.5 and show that in equilibrium $T = T'$ and $\sigma = \sigma'$, where the surface tension is defined as

$$\sigma = -T\left(\frac{\partial S}{\partial a}\right)_{U,V,N}.$$

Problem 60. Consider a spherical air bubble in water.

(a) Equate the pressure-volume work to the surface work, and find a relationship between the pressure and the radius. This relationship is analogous to the Law of Laplace (Problem 1.19).



(b) Consider a small bubble attached to a large one. Use the relationship derived in (a) to determine which bubble has the larger internal pressure. Which bubble tends to shrink and which tends to expand?

(c) The bubbles in (b) are a model for two alveoli connected by a bronchiole in our lungs. Explain why a special fluid called a surfactant is needed to reduce the surface tension in the water on the surface of the alveolus. For more on the biological implications of surface tension, see Denny (1993).

Section 3.16

Problem 61. Use the analysis presented in Sect. 3.16 to show that the surface tension is

$$\sigma = \left(\frac{\partial U}{\partial a}\right)_{S,V,N}.$$

Therefore, increasing the surface area when the entropy, volume and number of particles are fixed requires energy. For water, the surface tension is approximately $0.07\,\mathrm{J\ m^{-2}}$, which is a large value Denny (1993).

Section 3.17

Problem 62. The reaction $1\ \mathrm{glucose} + 6O_2 \leftrightarrow 6CO_2 + 6H_2O$ must conserve the number of each type of atom. Determine the chemical formula of glucose.

Section 3.18

Problem 63. System A consists of N particles that move from a region where the concentration is C_1 to another where the concentration is C_2, each experiencing a change in chemical potential $\Delta\mu = k_B T\ \ln(C_2/C_1)$. The process occurs at constant temperature and pressure. What is the ratio of the total number of microstates of system and surroundings after the move to the number before the move? Assume the concentrations do not change.

Problem 64. In pure water, some of the molecules dissociate into $H_2O \rightarrow H^+ + OH^-$. The standard Gibbs free energies are $G^0_{\mathrm{H2O}} = -237.2\,\mathrm{kJ\ mol^{-1}}$, $G^0_{\mathrm{OH}} = -157.3\,\mathrm{kJ\ mol^{-1}}$ and $G^0_{\mathrm{H}} = 0$.

(a) Determine ΔG^0 for this reaction.

(b) Calculate K_{eq} assuming $T = 25\,^\circ\mathrm{C}$.

(c) Derive an expression that relates K_{eq}, $[H^+]$ and $[OH^-]$. Note: By convention the reference concentration for water is taken to be the concentration of pure water instead of 1 mole per liter. The small amount of dissociation does not change $[H_2O]$ significantly, so the logarithmic term for water is zero: $K_{eq} = [H^+][OH^-]$.

(d) H^+ and OH^- are produced as a pair, so their concentrations are equal. Calculate $[H^+]$.

(e) The pH of water is defined as $-\log_{10}([H^+])$. What is the pH of pure water?

Problem 65. If one increases the volume of a liquid at constant p and T, a portion of the liquid evaporates. The amount of liquid decreases as V increases until all the liquid is vaporized. The pressure at which the two phases coexist is called the *vapor pressure*. The vapor pressure depends on the temperature, as shown. When two phases are in equilibrium, they are in mechanical, thermal, and diffusive equilibrium:

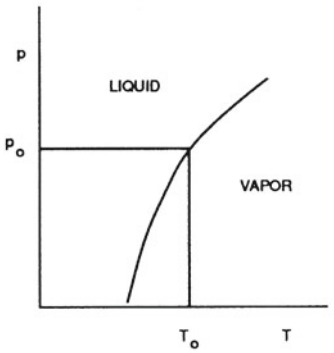
$T_l = T_g$, $p_l = p_g$, $\mu_l = \mu_g$. Thus, at any arbitrary point on the vapor-pressure curve, $\mu_g(T_0, p_0) = \mu_l(T_0, p_0)$. Consider some nearby point in the vapor-pressure curve, and expand both chemical potentials in a Taylor's series to show that

$$\frac{dp}{dT} = \frac{(\partial\mu_g/\partial T)_p - (\partial\mu_l/\partial T)_p}{(\partial\mu_l/\partial p)_T - (\partial\mu_g/\partial p)_T},$$

where dp/dT is the slope of the vapor-pressure curve. Use the fact that $G = N\mu(p, T)$, that $(\partial G/\partial T)_{N,p} = -S$, and that $(\partial G/\partial p)_{N,T} = V$, to show that

$$\frac{dp}{dT} = \frac{L}{T\Delta V},$$

where L is the latent heat of vaporization and ΔV is the volume change on vaporization. (Since L and V are both extensive parameters, they can be expressed per mole or per molecule.) This is called the *Clausius–Clapeyron* equation.



Problem 66. Use the Clausius–Clapeyron equation for the vapor pressure as a function of temperature (see Problem 65), $dp/dT = L/T\Delta V$, and assume an ideal gas so that $\Delta V \approx V_g = Nk_BT/p$ to find the vapor pressure p as a function of temperature.

Problem 67. Use the definition of Gibbs free energy $G = U - TS + pV$ and the thermodynamic identity $T\,dS = dU - \mu\,dN + p\,dV$ to find the partial derivatives of G when N, T, and p are the independent variables. Note that U, S, and V are all extensive variables so that G is proportional to N: $G = N\Phi$. Thereby relate Φ to the chemical potential.

Problem 68. (a) Find the change in Gibbs free energy $G = U - TS + pV$ for an ideal gas which changes pressure reversibly from p_1 to p_2 at a constant temperature.

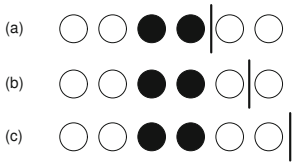
(b) Since $\Delta G = N\Delta\mu$, find $\Delta\mu$.

Problem 69. The argument leading to the change in G in a chemical reaction can be applied to a single particle moving from a region where the chemical potential is μ_A to a region where the chemical potential is μ_B by letting $dN = -dN_A = dN_B$, in which case $dG = (\mu_B - \mu_A)\,dN$. We saw in Sect. 3.13 that the chemical potential of a solute in an ideal solution had the form $\Delta\mu = k_BT\,\ln(C/C_0) + \Delta$(potential energy per particle). Sodium ions of charge $+e$ ($e = 1.6 \times 10^{-19}$ C) are found on one side of a membrane at concentration 145 mmol l^{-1}. The electrical potential is zero. On the other side of the membrane the concentration is 15 mmol l^{-1} and the potential is -90 mV. The change in electrical potential energy is $e\Delta v$. What is the change in Gibbs free energy if a single sodium ion goes from one side to the other? The temperature is 310 K and the pressure is atmospheric.

Section 3.18

Problem 70. Suppose that a potential energy term as well as a pressure must be added to the chemical potential, as was argued in Sect. 3.13. Consider a column of pure water. What is the difference in chemical potential between the top of the column and the bottom?

Problem 71. The open circles in the drawing represent water molecules. The solid circles are solute molecules. The vertical line represents a membrane that is permeable to water but not solute. In case (a) there are two water molecules to the right of the membrane. In (b) there is one, and in (c) none. What is the total number of microstates of the combined system in each case?



Problem 72. If we want to apply Eq. 3.79 when there is an appreciable difference in concentration, we can define an average concentration by

$$\Delta\mu_s = k_BT\,\ln(C_{s2}/C_{s1}) \equiv k_BT\,(\Delta C_s/\overline{C}_s),$$

$$\overline{C}_s \equiv \frac{\Delta C_s}{\ln(C_{s2}/C_{s1})} = \frac{\Delta C_s}{\ln(1 + \Delta C_s/C_{s1})}.$$

Use the Taylor's-series expansion $y = x/ln(1 + x) \approx 1 + x/2 - x^2/12 + \cdots$ to find an approximate expression for $\overline{C}_s$.

Problem 73. Verify that differentiation of Eq. 3.69 with respect to N_w and N_s gives Eq. 3.71.

References

Atkins PW (1994) The 2nd law: energy, chaos and form. Scientific American, New York

Cook G, Dickerson RH (1995) Understanding the chemical potential. Am J Phys 63(8):737–742

Denny MW (1993) Air and water: the biology and physics of life's media. Princeton University Press, Princeton

Haynie DT (2008) Biological thermodynamics, 2nd edn. Cambridge Univ Press, Cambridge

Hildebrand JH, Prausnitz JM, Scott RL (1970) Regular and related solutions: the solubility of gases, liquids, and solids. Van Nostrand Reinhold, New York

Hildebrand JH, Scott R L (1964) The solubility of nonelectrolytes, 3rd edn. Dover, New York

Huang S, Wikswo J (2006) Dimensions of systems biology. Rev Physiol Biochem Pharmacol 157:81–104

Press WH, Teukolsky SA, Vetterling WT Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd edn., reprinted with corrections. Cambridge University Press, New York

Reif F (1964) Statistical physics. Berkeley physics course, vol 5. McGraw-Hill, New York

Reif F (1965) Fundamentals of statistical and thermal physics. McGraw-Hill, New York

Rossi-Fanelli A, Antonini E (1958) Studies on the oxygen and carbon monoxide equillibria of human myoglobin. Arch Biochem Biophys 77(478):478–492

Schroeder DV (2000) An introduction to thermal physics Addison Wesley Longman, San Francisco

Schwartz D, Mayaux MJ (1982) Female fecundity as a function of age: results of artificial insemination in 2193 nulliparous women with azoospermic husbands. N Engl J Med 306(7):404–406

Weaver W (1963) Lady luck: the theory of probability Anchor, Garden City

Weaver WD (1982) Ventricular defibrillation—a comparative trial using 175-J and 320-J shocks. N Engl J Med 307:1101–1106

Transport in an Infinite Medium

4

In this book, Chaps. 4 and 5 are devoted to one of the most fundamental problems in physiology: the transport of solvent (water) and uncharged solute particles. Chapter 4 develops some general ideas about the movement of solutes in solution. Chapter 5 applies these ideas to movement of water and solute through a membrane.

Section 4.1 defines flux and fluence rate and derives the continuity equation. Section 4.2 shows how to calculate the solute fluence rate when the solute particles are drifting with a constant velocity, as when they are being dragged along by flowing solvent.

The next several sections are devoted to diffusion, the random motion of solute particles. Sections 4.3–4.5 describe random motion in a gas and a liquid. Section 4.6 states Fick's first law, which relates the fluence rate of diffusing particles to the gradient of their concentration. Section 4.7 relates the proportionality constant in Fick's first law to the viscous drag coefficient of the particle in the solution. Section 4.8 combines Fick's first law and the equation of continuity to give Fick's second law, the diffusion equation, that tells how the concentration $C(x, y, z, t)$ evolves with time. Section 4.9 discusses various time-independent (steady-state) solutions to the diffusion equation. Section 4.10 analyzes steady-state diffusion to or from a cell, including both diffusion through the membrane and in the surrounding medium. Section 4.11 discusses a model of steady-state diffusion of a substance that is being produced at a constant rate inside a spherical cell. Section 4.12 develops a steady-state solution when both drift and diffusion are taking place in one dimension. One technique for solving the time-dependent diffusion equation is introduced in Sect. 4.13. Section 4.14 describes a simple random-walk model for diffusion.

This chapter discusses how molecules and other objects can diffuse or drift. These physical processes occur in both living and nonliving material. However, much motion in the body arises from truly biological mechanisms (Fletcher and Theriot 2004; Hoffmann 2012). A simple example is the flagella that power the swimming of *Escherichia coli* bacteria (Berg 2003). Perhaps the best known example is the contractions caused by *myosin* and *actin* in skeletal muscle. Actin proteins form a "track," and myosin proteins "step" along the track, using energy stored in ATP. A similar molecular motor, *kinesin*, causes motion along microtubules, and is responsible for many intracellular types of motion such as chromosome migration during cell division (*mitosis*). The details about how these motors work is beyond the scope of our book, but understanding them requires knowledge from Chap. 1 (viscosity), Chap. 3 (bioenergetics), and this chapter (Brownian motion).

4.1 Flux, Fluence, and Continuity

Flow was introduced in Sect. 1.17 of Chap. 1. The *flow rate*, *volume flux*, or *volume current* i is the total volume of material transported per unit time and has units of $m^3\ s^{-1}$. One can also define the *mass flux* as the total mass transported per unit time or the *particle flux* as the total number of particles, and so on.

The *particle fluence* is the number of particles transported per unit area across an imaginary surface (m^{-2}). The *volume fluence* is the total volume transported across the surface per unit area and has units m^3 times m^{-2}, or m.

The *fluence rate* or *flux density* is the amount of "something" transported across an imaginary surface per unit area per unit time. It can be represented by a vector pointing in the direction the "something" moves and is denoted by $\mathbf{j}$. It has units of "something" $m^{-2}\ s^{-1}$. It is traditional to use a subscript to tell what is being transported: $\mathbf{j}_s$ is the solute particle fluence rate ($m^{-2}\ s^{-1}$), $\mathbf{j}_m$ is the mass fluence rate (kg $m^{-2}\ s^{-1}$), and $\mathbf{j}_v$ is volume flux density ($m^3\ m^{-2}\ s^{-1}$ or $m\ s^{-1}$). In a flowing fluid, $\mathbf{j}_v$ is the velocity with which the fluid moves.

Slightly different nomenclature is used in different fields. The words *flux* and *flux density* are often used interchangeably. Table 4.1 shows some of the names that are encountered. Do not spend much time memorizing it; it is provided

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_4, © Springer International Publishing Switzerland 2015

Table 4.1 Units and names for j and jS in various fields

	j		jS	
	Units	Names	Units	Names
Particles	$m^{-2}\ s^{-1}$	Particle fluence rate Particle current density Particle flux density Particle flux	s^{-1}	Particle flux Particle current Particle flux
Electric charge	$C\ m^{-2}\ s^{-1}$ or $A\ m^{-2}$	Current density	$C\ s^{-1}$ or A	Current
Mass	$kg\ m^{-2}\ s^{-1}$	Mass fluence rate Mass flux density Mass flux	$kg\ s^{-1}$	Mass flux Mass flow
Energy	$J\ m^{-2}\ s^{-1}$ or $W\ m^{-2}$	Energy fluence rate Intensity Energy flux	$J\ s^{-1}$ or W	Energy flux Power

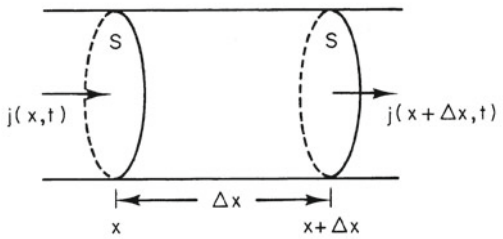


Fig. 4.1 The fluence rates used to derive the continuity equation in one dimension

to help you when you must deal with the notation in other books.

4.1.1 The Continuity Equation in One Dimension

As long as we are dealing with a substance that does not appear or disappear (as in a chemical reaction, radioactive decay, or the like), the number of particles or the mass, or in the case of an incompressible liquid, the volume, remains constant or is *conserved*. This conservation leads to a very useful equation called the *equation of continuity*. It will be derived here in terms of the number of particles.

We will first derive it in one dimension. Let the fluence rate of some species be j particles per unit area per unit time, passing a point. All motion takes place in the x direction along a tube of constant cross-sectional area S. The value of j may depend on the position in the tube and on the time: $j = j(x, t)$. The number of particles in the volume shown in Fig. 4.1 between x and $x + \Delta x$ is $N(x, t)$. At x, there may be particles moving both to the right and to the left; the net number to the right in Δt is $j(x, t)$ times the area S times the time Δt. A flux density in the $+x$ direction is called positive. The net number of particles in at x is $j(x, t)S\Delta t$. Similarly, the net number out at $x+\Delta x$ is $j(x+\Delta x, t)S\Delta t$. Combining these gives the net increase in the number of particles in the volume $S\Delta x$:

$$\Delta N = \left[j(x, t) - j(x + \Delta x, t)\right] S\, \Delta t. \tag{4.1}$$

As $\Delta x \to 0$, the quantity involving j is, by definition, related to the partial derivative of j with respect to x (Appendix N):

$$j(x, t) - j(x + \Delta x, t) = -\frac{\partial j(x, t)}{\partial x}\, \Delta x.$$

Similarly, the increase in $N(x, t)$ is

$$\Delta N(x, t) = N(x, t + \Delta t) - N(x, t) = \frac{\partial N}{\partial t}\, \Delta t.$$

These two expressions can be substituted in Eq. 4.1 to give

$$\frac{\partial}{\partial t} N(x, t) = -(S\Delta x)\, \frac{\partial}{\partial x} j(x, t).$$

This equation can be written in terms of the concentration $C(x, t)$ by dividing both sides by the volume $S\Delta x$:

$$\frac{\partial C}{\partial t} = -\frac{\partial j}{\partial x}. \tag{4.2}$$

This is the *continuity equation in one dimension*.

4.1.2 The Continuity Equation in Three Dimensions

In three dimensions $\mathbf{j}$ is a vector with components j_x, j_y, and j_z. The flux across a surface dS oriented at some arbitrary direction with the x, y, z axes is equal to the component of $\mathbf{j}$ perpendicular to the surface times dS. To see this, imagine that $\mathbf{j}$ lies in the xy plane with components j_x and j_y. If $\mathbf{j}$ makes an angle ϕ with the vertical, then $j_x = j \sin\phi$, $j_y = j \cos\phi$.

Consider the small volume shown in Fig. 4.2. If there is no buildup of particles within the volume, the flux in across the two faces parallel to the axes is equal to the flux across

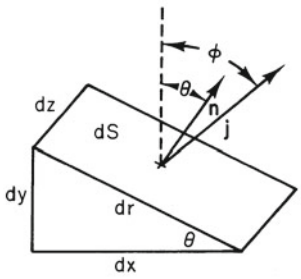


Fig. 4.2 Volume element used to relate the fluence rate across the slant face to the components of the fluence rate parallel to the x and y axes

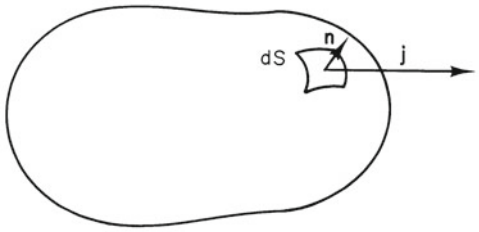


Fig. 4.3 The total number of particles per second passing through the closed surface (flux) is the sum of the contributions $j_n dS$ from all elements of the surface

dS. The area dS of the slant surface is $drdz$, where dz is the thickness of the volume perpendicular to the paper. The number of particles per second across the face $dydz$ is $j_x\, dydz = (j \sin\phi)(dydz)$. Since $dy = dr \sin\theta$, this may be written as $j \sin\phi \sin\theta dzdr$. Similarly, the number of particles per second in across the bottom face is $j_y\, dxdz = j \cos\phi \cos\theta dzdr$. The sum of these must be equal to the number leaving across the slant face: $j\, dzdr(\sin\phi \sin\theta + \cos\phi \cos\theta) = j\, dzdr \cos(\phi - \theta) = j\, dS \cos(\phi - \theta)$. The number of particles per unit area per second across the slant face is, therefore, $j\ \cos(\phi - \theta)$. Now $\phi - \theta$ is the angle between $\mathbf{j}$ and the unit vector $\hat{\mathbf{n}}$ perpendicular to the surface. We can write the flux density across dS as j_n (the component of $\mathbf{j}$ parallel to $\hat{\mathbf{n}}$), or $\mathbf{j} \cdot \hat{\mathbf{n}}$ (the dot product of $\mathbf{j}$ and the normal). The flux (flow per second) is sometimes written as

$$(\mathbf{j} \cdot \hat{\mathbf{n}})dS, \quad j_n dS, \quad \text{or} \quad (\mathbf{j} \cdot d\mathbf{S}). \tag{4.3}$$

These are all equivalent: vector $d\mathbf{S}$ is defined to have magnitude dS and to point along the normal to the surface that points outward from the enclosed volume. The same result is obtained (with more algebra) when $\mathbf{j}$ is not in the xy plane.

4.1.3 The Integral Form of the Continuity Equation

If we consider a closed volume as shown in Fig. 4.3, the total number of particles flowing out of the volume can be obtained by adding up the contribution from each element dS. It is

$$\text{(total number of particles out in time } \Delta t) = \left(\iint_{\text{closed surface}} j_n dS \right) \Delta t.$$

Since the total number of particles in the volume enclosed by the surface is

$$\iiint_{\text{enclosed volume}} C(x, y, z, t)\, dxdydz,$$

we can write[1]

$$\frac{\partial}{\partial t} \iiint_{\text{enclosed volume}} C\, dV = - \iint_{\substack{\text{surface enclosing} \\ \text{the volume}}} j_n\, dS. \tag{4.4}$$

The outward flux density or fluence rate of the substance integrated over a closed surface (the net flux through the surface) is equal to the rate of decrease of the amount of substance within the volume enclosed by the surface.

How to evaluate the surface integral is best shown by two examples. First consider a volume defined by a sphere of radius r. A lamp at the center of the sphere radiates light uniformly in all directions. The light leaves through the surface of the sphere. The amount of light energy in the volume defined by the sphere is not changing, so the rate of energy production by the lamp P is equal to the energy flux through the surface of the outer sphere:

$$P = \iint j_n\, dS. \tag{4.5}$$

Because of the spherical symmetry, $\mathbf{j}$ is perpendicular to the surface and is the same at all points on the sphere. Therefore,

$$P = j_n \iint dS.$$

Since the integral of dS over the surface of a sphere of radius r is $4\pi r^2$,

$$j = j_n = \frac{P}{4\pi r^2}. \tag{4.6}$$

The amount of energy per unit area per unit time crossing the surface of the sphere is the energy fluence rate or the intensity.

The second example is slightly more complicated. Suppose that $\mathbf{j}$ is parallel to the z-axis and has the same value everywhere. The net flux through any closed surface will be zero in that case, and we will verify it to show how to evaluate a surface integral. Consider the situation shown in

[1] We can write dV as $d^3\mathbf{r}$ or $dxdydz$.

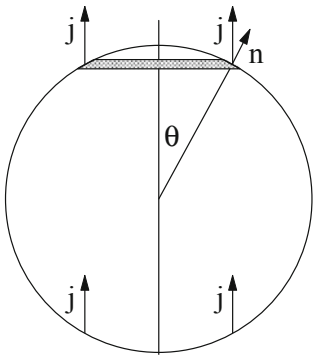


Fig. 4.4 The fluence rate is the same everywhere. The flux is $\int j_n dS$ over the entire sphere. When the normal component of the fluence rate is outward, the contribution is positive. When it is inward, the contribution is negative

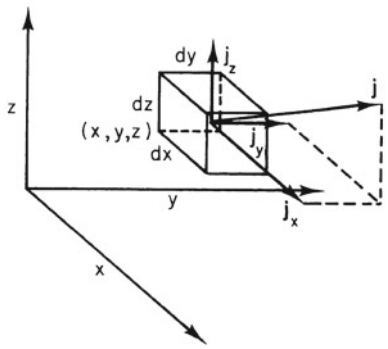


Fig. 4.5 The small volume used to derive the differential form of the continuity equation

Fig. 4.4, where j_n is integrated over the surface of the sphere. At every point in the shaded strip, $j_n = j\cos\theta$. The strip has width $r\,d\theta$ and circumference $2\pi r\sin\theta$, so its area is $2\pi r^2 \sin\theta\, d\theta$. Thus

$$\int j_n\, dS = \int_0^\pi j\cos\theta\, 2\pi r^2 \sin\theta\, d\theta$$
$$= 2\pi r^2 j \int_0^\pi \cos\theta\, \sin\theta\, d\theta = 0.$$

4.1.4 The Differential Form of the Continuity Equation

The continuity equation can be expressed in terms of derivatives instead of integrals. To derive this form, consider the increase in the number of particles in a small rectangular volume located at (x, y, z) and having sides (dx, dy, dz) as shown in Fig. 4.5. Apply Eq. 4.4 to each face of the volume. The rate at which the substance flows in through the face at x is $j_x(x)(dydz)$. At face $x+dx$, it flows out at a rate $j_x(x+dx)dydz$. There is no contribution to the flow through this face from j_y or j_z, since they are parallel to the face. The net increase in the number of particles in the volume due to the two terms is

$$-\left[j_x(x+dx) - j_x(x)\right] dydz = -\frac{\partial j_x}{\partial x}\, dxdydz.$$

Similar terms can be written for the faces perpendicular to the y and z axes. The total amount of the substance entering the volume per unit time is the rate of change of the amount within the volume, which is the rate of change of concentration times the volume $dxdydz$. Therefore,

$$\frac{\partial C}{\partial t}(dxdydz) = -\left(\frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y} + \frac{\partial j_z}{\partial z}\right)(dxdydz)$$

or

$$-\frac{\partial C}{\partial t} = \frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y} + \frac{\partial j_z}{\partial z}. \tag{4.7}$$

This is the differential form of the continuity equation. Equation 4.2 was a special case of this when **j** was parallel to the x axis.

The combination of derivatives on the right-hand side of Eq. 4.7 occurs frequently enough to warrant a special name. It is called the *divergence* of the vector **j**:[2]

$$\text{div}\,\mathbf{j} = \nabla\cdot\mathbf{j} = \frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y} + \frac{\partial j_z}{\partial z}.$$

The continuity equation is therefore

$$\frac{\partial C}{\partial t} = -\text{div}\,\mathbf{j}. \tag{4.8}$$

This differential form of the continuity equation is completely equivalent to the integral form, Eq. 4.4. It is sometimes more convenient to use Eq. 4.4 and at other times more convenient to use Eq. 4.8.

The continuity equation says that the rate of decrease of the amount of a conserved substance in a certain region expressed as $-\partial C/\partial t$ is equal to the rate at which it leaves the region expressed as the flow through the surface surrounding the region. The substance may be a certain kind of molecule, electric charge, heat, or mass. If it is electric charge, **j** is the electric current per unit area and C is the charge per unit volume. If it is mass, C is the mass per unit volume or density ρ. The continuity equation is found in many contexts; in each, it expresses the conservation of some quantity.

In the flow of a liquid, the density of the liquid ρ, the mass M, and volume V are related by $M = \rho V$. If the liquid is

[2] The divergence is one of the concepts of vector calculus. A good review of vector calculus is Schey (2004).

incompressible, a given mass always occupies the same volume, and the density does not change. Therefore, $\partial\rho/\partial t = 0$, and the equation of continuity gives

$$\operatorname{div} \mathbf{j}_m = 0. \tag{4.9}$$

4.1.5 The Continuity Equation with a Chemical Reaction

Our derivation of the continuity equation assumed that the substance was conserved—neither created nor destroyed. If a chemical reaction is creating the substance at a rate Q particles $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$ (which may depend on position) then the continuity equation becomes

$$\frac{\partial C}{\partial t} = Q - \operatorname{div} \mathbf{j}, \tag{4.10a}$$

$$\begin{aligned} &\frac{\partial}{\partial t} \iiint_{\text{volume}} C(x, y, z)\, dV \\ &= \iiint_{\text{volume}} Q(x, y, z)\, dV - \iint_{\substack{\text{surface} \\ \text{enclosing} \\ \text{the volume}}} j_n\, dS. \end{aligned} \tag{4.10b}$$

If particles are being consumed in the chemical reaction, then Q is negative.

4.2 Drift or Solvent Drag

One simple way that solute particles can move is to drift with constant velocity. They can do this in a uniform electric or gravitational field if they are also subject to viscous drag, or they can be carried along by the solvent, a process called *drift* or *solvent drag*. (The solute particles are dragged by the solvent.) The solute fluence rate is $\mathbf{j}_s$, with units of particles $\mathrm{m}^{-2}\ \mathrm{s}^{-1}$ or just $\mathrm{m}^{-2}\ \mathrm{s}^{-1}$. The number of solute particles passing through a surface is the volume of solution that moves through the surface times the concentration of solute particles. Therefore,

$$\mathbf{j}_s = C\, \mathbf{j}_v. \tag{4.11}$$

This effect will be explored in greater detail in Sect. 4.12.

4.3 Brownian Motion

There is also movement of solute molecules when the water is at rest. If the solution is dilute, the solute particles are far apart and hit each other only occasionally. They are struck by water molecules much more often. The result is that they are in continual helter–skelter motion. Each solute molecule is influenced by the water molecules around it, but not by other solute molecules.

Table 4.2 Values of the rms velocity for various particles at body temperature

Particle	Molecular weight	Mass (kg)	v_{rms} ($\mathrm{m\ s}^{-1}$)
H_2	2	3.4×10^{-27}	1940
H_2O	18	3×10^{-26}	652
O_2	32	5.4×10^{-26}	487
Glucose	180	3×10^{-25}	200
Hemoglobin	65,000	1×10^{-22}	11
Bacteriophage	6.2×10^6	1×10^{-20}	1.1
Tobacco mosaic virus	40×10^6	6.7×10^{-20}	0.4
E. coli		2×10^{-15}	0.0025

In Chap. 3, it was shown that the relative probability for a particle to have energy u when it is in thermal equilibrium with a reservoir at temperature T is given by a Boltzmann factor:[3] $P \propto e^{-u/k_B T}$. In Chap. 3, the Boltzmann factor was used to show that if any energy term depends on the square of some variable, then the average value of that term is $k_B T/2$. A particle with kinetic energy of translation $m(v_x^2 + v_y^2 + v_z^2)/2$ has an average energy $k_B T/2$ for each of the three terms, or a total translational kinetic energy of $3k_B T/2$. This is true regardless of the mass of the particle. Any particle in thermal equilibrium with a reservoir (which can be the surrounding fluid) will move with a mean square velocity given by[4]

$$\overline{v^2} = \frac{3k_B T}{m}. \tag{4.12}$$

The square root of $\overline{v^2}$ is called the *root-mean-square* or *rms* velocity. It decreases with increasing mass of the particle. Table 4.2 shows values of $v_{\text{rms}} = \left(\overline{v^2}\right)^{1/2}$ for different particles at body temperature.

This movement of microscopic-sized particles, resulting from bombardment by much smaller invisible atoms, was first observed by the English botanist Robert Brown in 1827 and is called *Brownian motion*. Solute particles are also subject to this random motion. If the concentration of particles is not uniform, there will be more particles wandering from a region of high concentration to one of low concentration than vice versa. This motion is called *diffusion*.

[3] The Boltzmann factor provided Jean Perrin with the first means to determine Avogadro's number. The density of particles in the atmosphere is proportional to $\exp(-mgy/k_B T)$, where mgy is the gravitational potential energy of the particles. Using particles for which m was known, Perrin was able to determine k_B for the first time. Since the gas constant R was already known, Avogadro's number was determined from the relationship $R = N_A k_B$. See Problem 12.

[4] The average velocity is $\bar{v}_x = 0$, since a particle with a given speed moves with equal probability to the left or right.

In the next several sections, we study random motion and diffusion, first for a gas and then for a liquid.

4.4 Motion in a Gas: Mean Free Path and Collision Time

It is possible to define a *mean free path*, which is the average distance a particle travels between successive collisions, and a *collision time*, the average length of time between collisions. Consider a collection of N_0 molecules. The number that have moved distance x without suffering a collision is $N(x)$. For short distances dx, the probability that a molecule collides with another molecule is proportional to dx: call it $(1/\lambda)dx$. Then, on the average, the number of molecules having their first collision between x and $x + dx$ is $dN = -N(x)(1/\lambda)dx$. This is the familiar equation for exponential decay. The number of molecules surviving without any collision is $N(x) = N_0 e^{-x/\lambda}$.

To compute the average distance traveled by a molecule between collisions, we multiply each possible value of x by the number of molecules that suffer their first collision between x and $x + dx$. Since $N(x)$ is the number surviving at distance x, and dx/λ is the probability that one of those will have a collision between x and $x + dx$, the mean value of x is

$$\overline{x} = \frac{1}{N_0}\int_0^\infty xN(x)\frac{1}{\lambda}\,dx.$$

With the substitutions $s = x/\lambda$ and $N(x) = N_0 e^{-s}$, this can be written as

$$\begin{aligned}\overline{x} &= \lambda\int_0^\infty e^{-s}s\,ds \qquad (4.13)\\ &= -\lambda\left[e^{-s}(s+1)\right]_0^\infty = \lambda.\end{aligned}$$

Thus λ is the mean free path.

A similar argument can be made for the length of time that each molecule survives before being hit. The probability that a molecule is hit during a short time dt is proportional to dt: call it $(1/t_c)dt$. The number of molecules surviving a time t is given by $N = N_0 e^{-t/t_c}$, and the mean time between collisions can be calculated as above. It is t_c, which is called the *collision time*. The number of collisions per second is the *collision frequency*, $1/t_c$.

It is possible to estimate the mean-free path and the collision frequency. Consider a particle of radius a_1 moving through a dilute gas of other particles of radius a_2. For convenience, imagine that particle 1 is moving and that all the other particles are fixed in position. The path of the first particle is shown in Fig. 4.6. If the center of one of these other molecules lies within a distance $a_1 + a_2$ of the moving molecule, there will be a collision. The effect is the same as if the moving particle had radius $a_1 + a_2$ and all the other particles were points. In moving a distance x, the particle sweeps out a volume $V(x) = \pi(a_1 + a_2)^2 x$. On average, when the particle has traveled a mean-free path there is one collision. The average number of gas particles in the volume $V(\lambda) = \pi(a_1 + a_2)^2\lambda$ is therefore 1. The average number of particles per unit volume is C. Thus $1 = C\pi(a_1 + a_2)^2\lambda$, or

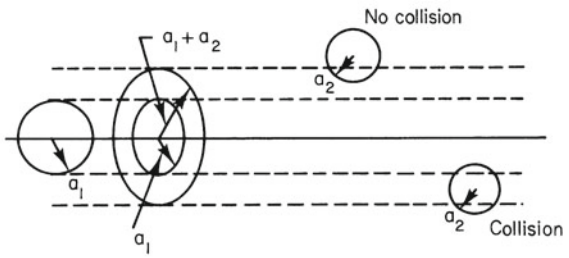


Fig. 4.6 A particle of radius a_1 moves through a gas of particles of radius a_2. A collision will occur if the center of another particle lies within a distance $a_1 + a_2$ of the trajectory of the particle under consideration

$$\lambda = \frac{1}{\pi(a_1 + a_2)^2 C}. \qquad (4.14)$$

The quantity $\pi(a_1+a_2)^2$ is the area of a circle. It is called the *cross-section* for the collision of these particles. The concept of cross-section is used extensively in Chap. 15.

This estimation is somewhat crude in its assumption that only one molecule is moving. If all the molecules are of the same kind then the factor 1 in the numerator is replaced by $2^{-1/2} = 0.707$ (Reif 1965, p. 471).

For a gas at standard temperature and pressure, the volume of 1 mol is $22.4\ \mathrm{l} = 22.4 \times 10^{-3}\ \mathrm{m}^3$, so $C = 2.7 \times 10^{25}\ \mathrm{m}^{-3}$. If $a_1 = a_2 = 0.15$ nm, then Eq. 4.14 can be used to calculate the mean free path:

$$\begin{aligned}\lambda &= \frac{1}{(3.14)(.3\times10^{-9})^2\,\mathrm{m}^2(2.7\times10^{25}\,\mathrm{m}^{-3})}\\ &= 0.13\ \mu\mathrm{m}.\end{aligned}$$

For a gas at standard temperature and pressure, the mean-free path is about 1000 times the molecular diameter, and the assumption of infrequent collisions is justified.

The collision time can be estimated by saying that

$$t_c = \frac{\lambda}{\overline{v}},$$

where $\overline{v}$ is the average speed of the molecules. Using the rms velocity for $\overline{v}$, we can use Eq. 4.12 to write

$$t_c \approx \lambda\left(\frac{m}{3k_B T}\right)^{1/2}. \qquad (4.15)$$

The important feature of this is the dependence on $m^{1/2}$ and on λ. For air at room temperature, $t_c = 2 \times 10^{-10}$ s.

4.5 Motion in a Liquid

The assumptions of the previous section do not hold in a liquid, in which the particle is being continually bombarded by neighbors. Blindly applying Eq. 4.14 to water, we can use the fact that 1 mol is 18 g and occupies 18 cm^3, to obtain $\lambda = 0.1$ nm, so that $a/\lambda \approx 1$, and the assumptions behind the derivation break down. Estimating the collision time with Eq. 4.15 gives a value that is a factor of 1000 less than for the gas, or 10^{-13} s.

Although these estimates of the mean-free path and the collision time are undoubtedly wrong, the concepts appear to be valid. Computer simulations of molecular collisions show that the distribution of free paths is exponential even though the mean-free path is only a fraction of a molecular diameter. In Sect. 4.12, we will regard diffusion as a random walk of the diffusing particles and relate the diffusion constant to the mean-free path and collision time. Equations 4.14 and 4.15 can then be used to show that the diffusion constant should be inversely proportional to the square of the particle radius. This has been verified experimentally for the diffusion of certain liquids. Evidence for the validity of this random-walk model for diffusion in liquids has been summarized by Hildebrand et al. (1970, pp. 36–39).

A particle in a liquid is subject to a fluctuating force $\mathbf{F}(t)$, which is random in magnitude and direction. The particle begins to move in response to this force. However, once it has begun to move, it suffers more collisions in front than behind, so the force slows it down. As the particle can neither stay at rest nor continue to move in the same direction, it undergoes a random, zig-zag motion with average translational kinetic energy $3k_BT/2$. The mean square velocity is not zero, but the mean vector velocity is zero.

For each particle, Newton's second law is $m(d\mathbf{v}/dt) = \mathbf{F}(t)$. This is not very useful as it stands. To make it more tractable, consider a particle with average velocity $\bar{\mathbf{v}}$. (The average means that an ensemble of identically prepared particles is examined.) The particle has more collisions on the front that slow it down. We therefore break up $\mathbf{F}(t)$ into two parts: an average drag force, which will be the same for all the particles in the ensemble, and a rapidly fluctuating part $\mathbf{g}(t)$, which will vary with time and from particle to particle. Newton's second law is then $m(d\mathbf{v}/dt) = (\text{drag force}) + \mathbf{g}(t)$, where $\mathbf{g}(t)$ is random in direction. The drag force will be zero when $\bar{\mathbf{v}}$ is zero. For average velocities that are not too large, it can be approximated by a linear term:

$$(\text{drag force}) = -\beta\,\bar{\mathbf{v}}.$$

With this approximation, Newton's second law is known as the *Langevin equation:*

$$m\frac{d\mathbf{v}}{dt} = -\beta\,\bar{\mathbf{v}} + \mathbf{g}(t). \tag{4.16}$$

(If the liquid is moving, the drag force will be zero when the particle has the same average velocity as the liquid. So $\bar{\mathbf{v}}$ can be interpreted as the relative velocity of the particle with respect to the liquid.) This equation often has another term in it, which does not average to zero and which represents some external force such as gravity that acts on all the particles. This approximate equation can be solved in some cases, though with difficulty, and has formed the basis for some treatments of the motion of large particles in fluids. With suitable interpretation, it can describe motion of the fluid molecules themselves.[5] In particular, when dealing with molecular motion it is necessary to consider the fact that the molecules do not move independently of one another.

For a Newtonian fluid (Eq. 1.33) with viscosity η, one can show (although it requires some detailed calculation,[6] see Problem 46 in Chap. 1) that the drag force on a spherical particle of radius a is given by

$$\mathbf{F}_{\text{drag}} = -\beta\bar{\mathbf{v}} = -6\pi\eta a\bar{\mathbf{v}}. \tag{4.17}$$

This equation is valid when the sphere is so large that there are many collisions of fluid molecules with it and when the velocity is low enough so that Reynolds number is small. This result is called *Stokes' law.*

If the sphere is not moving in an infinite medium but is confined within a cylinder, then a correction must be applied.[7] In that case, the viscous drag depends on the velocity of the spherical particle through the fluid, the average velocity of the fluid through the cylinder, and the distance of the particle from the axis of the cylinder.[8]

[5] See, for example, Pryde (1966, p. 161).

[6] This is an approximate equation. See Barr (1931, p. 171).

[7] An early correction for particles on the axis of a cylinder is found in Barr (1931, p. 183). More recent work is by Levitt (1975), by Bean (1972), and by Paine and Scherr (1975).

[8] Stokes' law is valid for a particle in a gas if the mean free path is much less than the particle radius a, so that many collisions with neighboring molecules occur. At the other extreme, a mean free path much greater than the particle radius, the drag force turns out to be $F_{\text{drag}} = \alpha\eta a(a/\lambda)\overline{v}$. Although this will not be directly useful to us in considering biological systems, it is mentioned here to show how important it is to understand the conditions under which an equation is valid. Although the dimensions of this new equation are unchanged (we have introduced a factor a/λ, which is dimensionless), the drag force depends on a^2 instead of on a. The reason for the difference is that collisions are now infrequent and that the probability of a collision that imparts some average momentum change is proportional to the projected cross-sectional area of the sphere, πa^2. In the regime of interest to us, in which there are many collisions, we would not expect the force

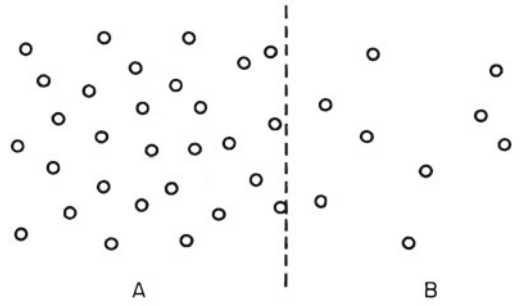


Fig. 4.7 An example of diffusion. Each molecule at A or B can wander with equal probability to the left or right. There are more molecules at A to wander to the right than there are at B to wander to the left. There is a net flow of molecules from A to B

4.6 Diffusion: Fick's First Law

Diffusion is the random movement of particles from a region of higher concentration to a region of lower concentration. The diffusing particles move independently of one another; they may collide frequently with the molecules of the fluid in which they are immersed, but they rarely collide with one another. The surrounding fluid may be at rest, in which case diffusion is the only mechanism for transport of the solute, or it may be flowing, in which case it carries the solute along with it (solvent drag). Both effects can occur together.

We first consider diffusion from a macroscopic point of view and write down an approximate differential equation to describe it. We then obtain a second equation describing diffusion by combining this with the continuity equation. After discussing some solutions to these equations, we look at the problem from a microscopic point of view, considering the random motion of the particles, and show that we get the same results.

Suppose that the surrounding solvent does not move. If the solute concentration is completely uniform, there is no net flow. As many particles wander to the left as to the right, and the concentration remains the same. There will be local fluctuations in concentration, analogous to those we have seen in the preceding chapter for fluctuations in the concentration of a gas, but that is all.

However, if the concentration is higher in region A than in region B to the right of it, there are more particles to wander to the right from A to B than there are to wander to the left from B to A (Fig. 4.7). If the problem is one-dimensional, there is no net flow if $\partial C/\partial x = 0$, but there is flow if $\partial C/\partial x \neq 0$. If the concentration difference is small, then the flux density j is linearly proportional to the concentration gradient $\partial C/\partial x$. The equation is

$$j_x = -D\,\frac{\partial C}{\partial x}. \tag{4.18a}$$

Constant D is called the *diffusion constant*. The units of D are $\mathrm{m}^2\ \mathrm{s}^{-1}$, as may be seen by noting that the units of j are (something) $\mathrm{m}^{-2}\ \mathrm{s}^{-1}$ and the units of $\partial C/\partial x$ are (something) m^{-4}. This relationship is called *Fick's first law of diffusion*, after Adolf Fick, a German physiologist in the last half of the nineteenth century. The minus sign shows that the flow is in the direction from higher concentration to lower concentration: if $\partial C/\partial x$ is positive, the flow is in the $-x$ direction.

If the actual process is not linear, this can be thought of as the first term of a Taylor's series expansion (Appendix D).

Fick's first law is one of many forms of the transport equation. Other forms are shown in Table 4.3. The units of the constant are different for the last three entries in the table because the quantity that appears on the right has different units than the quantity on the left. In each case, however, a fluence rate or flux density (of particles, mass, energy, electric charge, or momentum) is related to a rate of change of some other quantity with position. This rate of change is called the *gradient* of the quantity. The gradient is often called the *driving force*. The concentration gradient or driving force causes the diffusion of particles; the temperature gradient "causes" the heat flow; the electric voltage gradient "causes" the current flow; the velocity gradient "causes" the momentum flow.

The diffusive fluence rate can be related to the gradient of the chemical potential of the solute. With the notation $C_1 = C_s$ and $C_2 - C_1 = \Delta C_s$, Eq. 3.48 can be rewritten as

$$\begin{aligned}\Delta\mu_s &= k_B T \ln(C_2/C_1) = k_B T \ln(1 + \Delta C_s/C_s)\\ &\approx k_B T\,\Delta C_s/C_s,\end{aligned}$$

from which $\Delta C_s \approx C_s \Delta\mu_s/k_B T$, so

$$\frac{\partial C_s}{\partial x} = \frac{C_s}{k_B T}\frac{\partial \mu_s}{\partial x}$$

and

$$j_{sx} = -\frac{DC_s}{k_B T}\frac{\partial \mu_s}{\partial x}. \tag{4.18b}$$

The solute flux density is proportional to the diffusion constant, the solute concentration, and the gradient in the chemical potential per solute particle.

In three dimensions, the flow of particles can point in any direction and have components j_x, j_y, and j_z. An equation can be written for each component that is analogous to Eq. 4.18a or 4.18b. We can write one vector equation instead

to depend on λ. We hope that this will convince you of the danger in using someone else's equation without understanding it.

Table 4.3 Various forms of the transport equation

Substance flowing	Equation	Units of j	Units of the constant
Particles	$j_s = -D\dfrac{\partial C}{\partial x}$	$\mathrm{m^{-2}\ s^{-1}}$	$\mathrm{m^2\ s^{-1}}$
Mass	$j_m = -D\dfrac{\partial \rho}{\partial x}$	$\mathrm{kg\ m^{-2}\ s^{-1}}$	$\mathrm{m^2\ s^{-1}}$
Heat	$j_H = -\kappa\dfrac{\partial T}{\partial x}$	$\mathrm{J\ m^{-2}\ s^{-1}}$ or $\mathrm{kg\ s^{-3}}$	$\mathrm{J\ K^{-1}\ m^{-1}\ s^{-1}}$
Electric charge	$j_e = -\sigma\dfrac{\partial V}{\partial x}$	$\mathrm{C\ m^{-2}\ s^{-1}}$	$\mathrm{C\ m^{-1}\ s^{-1}\ V^{-1}}$ or $\Omega^{-1}\ \mathrm{m^{-1}}$
Viscosity (y component of momentum transported in the x direction)	$j_p = -\eta\dfrac{\partial v_y}{\partial x}$	$\mathrm{N\ m^{-2}}$ or $\mathrm{kg\ m^{-1}\ s^{-2}}$	$\mathrm{kg\ m^{-1}\ s^{-1}}$ or Pa s

of three equations for the three components by defining $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ to be unit vectors along the axes. Then

$$j_x\hat{\mathbf{x}} + j_y\hat{\mathbf{y}} + j_z\hat{\mathbf{z}} = -D\left(\frac{\partial C}{\partial x}\hat{\mathbf{x}} + \frac{\partial C}{\partial y}\hat{\mathbf{y}} + \frac{\partial C}{\partial z}\hat{\mathbf{z}}\right).$$

We have created a vector that depends on $C(x, y, z, t)$ by performing the indicated differentiations on C and multiplying the results by the appropriate unit vectors. This vector function is the gradient of C in three dimensions:

$$\text{grad}\, C = \nabla C = \frac{\partial C}{\partial x}\hat{\mathbf{x}} + \frac{\partial C}{\partial y}\hat{\mathbf{y}} + \frac{\partial C}{\partial z}\hat{\mathbf{z}}. \tag{4.19}$$

Fick's first law with this notation is

$$\mathbf{j} = -D\ \text{grad}\ C = -D\,\nabla C. \tag{4.20}$$

Remember that this is simply shorthand for three equations like Eq. 4.18a. If you feel a need to review vector calculus, which deals with the divergence and gradient, an excellent text is the one by Schey (2004).

4.7 The Einstein Relationship Between Diffusion and Viscosity

Before we can apply Fick's first law to real problems, we must determine the value of the diffusion constant D. The experimental determination of D is often based on Fick's second law of diffusion, which combines the first law with the equation of continuity and is discussed in the next section. It is closely related to the viscosity, as was first pointed out by Albert Einstein. This is not surprising, since diffusion is caused by the random motion of the particles under the bombardment of neighboring atoms, and viscous drag is also caused by the bombardment by neighboring atoms. What is remarkable is that a general relationship between them can be deduced quite easily by imagining just the right sort of experiment.

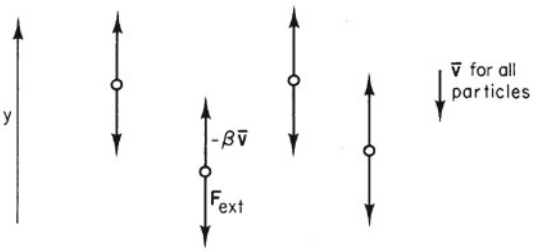


Fig. 4.8 Particles drifting under the influence of a downward force $\mathbf{F}_{\text{ext}}$

Consider a collection of particles uniformly suspended in a fluid at rest. Imagine that each particle is suddenly subjected to an external force $\mathbf{F}_{\text{ext}}$ (such as gravity) that acts in the $-y$ direction, as shown in Fig. 4.8. The particles will all begin to drift downward, speeding up until the upward viscous force on them balances the external force: $\mathbf{F}_{\text{ext}} - \beta\,\bar{\mathbf{v}} = 0$. In terms of magnitudes, $F_{\text{ext}} = \beta v$.

Because these particles are all moving downward, there is a downward flux density. With reference to Fig. 4.9, the number of particles crossing area S in time Δt will be those within the cylinder of height $\bar{v}\Delta t$. That number is the concentration times the volume ($S\bar{v}\Delta t$). Dividing by S and Δt gives

$$\mathbf{j}_{\text{drift}} = -\bar{v}C(y)\hat{\mathbf{y}}.$$

As the particles move down, they deplete the upper region of the fluid and cause a concentration gradient. This concentration gradient causes an upward diffusion of particles, with a flux density given by

$$\mathbf{j}_{\text{diff}} = -D\frac{\partial C}{\partial y}\hat{\mathbf{y}}.$$

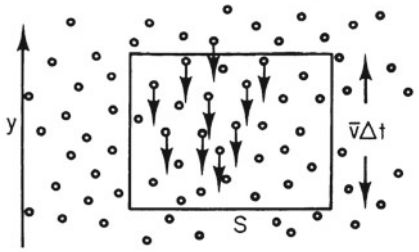


Fig. 4.9 Calculating the fluence rate of particles drifting downward

Equilibrium will be established when these two flux densities are equal in magnitude: $|\mathbf{j}_{\text{drift}}| = |\mathbf{j}_{\text{diff}}|$,

$$|\overline{v}C(y)| = \left| D\frac{\partial C}{\partial y} \right|. \tag{4.21}$$

But equilibrium means that the particles have a Boltzmann distribution in y, because their potential energy increases with y (work is required to lift them in opposition to $\mathbf{F}_{\text{ext}}$). For a constant $\mathbf{F}_{\text{ext}}$ independent of y, the energy is $u(y) = F_{\text{ext}}y$, where F_{ext} is the magnitude of the force. The concentration is

$$C(y) = C(0)e^{-F_{\text{ext}}y/k_BT}.$$

Therefore

$$\frac{\partial C}{\partial y} = -\frac{F_{\text{ext}}}{k_BT}C(y).$$

Inserting this in Eq. 4.21 gives $\overline{v} = DF_{\text{ext}}/k_BT$ or $D = \overline{v}k_BT/F_{\text{ext}}$. In equilibrium, the magnitude of F_{ext} is equal to the magnitude of the viscous force $\mathbf{f}$. Therefore $D = k_BT\overline{v}/f$. Since the viscous force is proportional to the velocity, $|f| = |\beta\overline{v}|$,

$$D = \frac{k_BT}{\beta}. \tag{4.22}$$

The derivation of this equation required only that the velocities be small enough so that the linear approximations for Fick's first law and the viscous force are valid. It is independent of the nature of the particle or its size. If in addition the diffusing particles are large enough so that Stokes' law is valid, then $\beta = 6\pi\eta a$ and

$$D = \frac{k_BT}{6\pi\eta a}. \tag{4.23}$$

The diffusion constant is inversely proportional to the fluid viscosity and the radius of the particle.

Combining Eqs. 4.18b and 4.22 shows that in terms of the chemical potential,

$$j_{sx} = -\frac{C_s}{\beta}\frac{\partial \mu_s}{\partial x}.$$

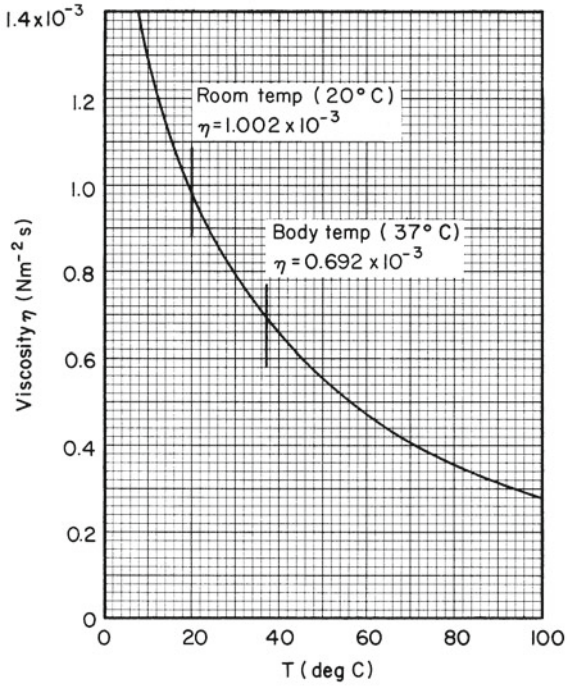


Fig. 4.10 Viscosity of water at various temperatures. (Data are from Weast 1972, p. F-36)

Sometimes minus the gradient of the chemical potential is called the driving force. To see why, note that for solvent drag, $j_s = C_s\overline{v}$, so $\beta\overline{v} = -\partial\mu_s/\partial x$ is the driving force.

The viscosity of water varies rapidly with temperature, as shown in Fig. 4.10. These values of viscosity and Eq. 4.23 have been used to calculate the solid lines for D vs a shown in Fig. 4.11. Various experimental values are also shown. The diffusion constant increases rapidly with temperature, so that care must be taken to specify the temperature at which the data are obtained. Since not all the molecules are spherical, there is some uncertainty in the value of the particle radius a.

Figure 4.12 is a plot of D for particles diffusing in water at 20 °C (293 K) vs. molecular weight M. Although the solid line provides a rough estimate of D if M is known, scatter is considerable because of varying particle shape. DNA lies a factor of 10 below the curve, presumably because it is partially uncoiled and presents a larger size than other molecules of comparable molecular weight.

It is possible to measure the *self-diffusion* of water in water by using a few water molecules in which one hydrogen atom is radioactive and measuring how they diffuse. Water has an unusually large self-diffusion constant.[9]

[9] For self-diffusion (such as radioactively tagged water in water), a hydrodynamic calculation shows that $\beta = 4\pi\eta a$ instead of $6\pi\eta a$ (Bird et al. 1960, p. 514ff.).

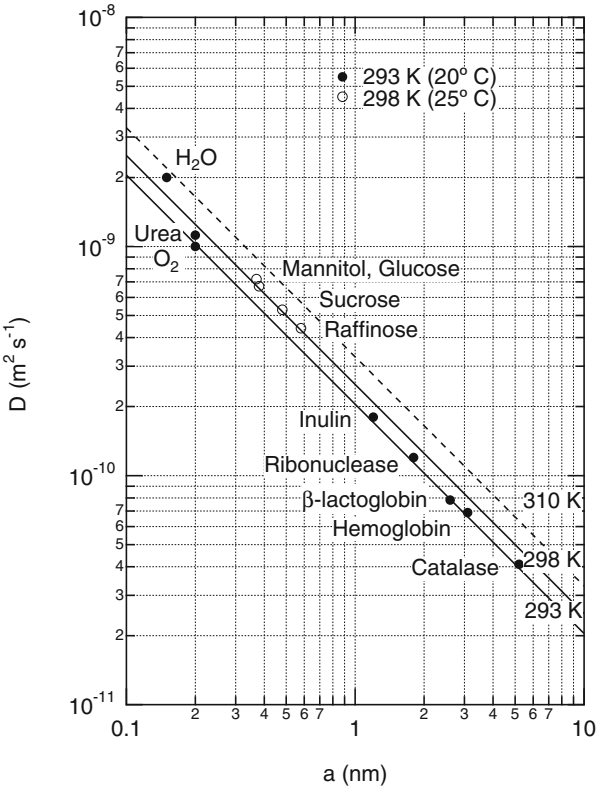


Fig. 4.11 Diffusion constant versus sphere radius a for diffusion in water at three different temperatures. Experimental data at 20 °C (293 K) are from Benedek and Villars (2000, Vol. 2, p. 122). Data at 25 °C (298 K) are from Weast (1972, p. F-47)

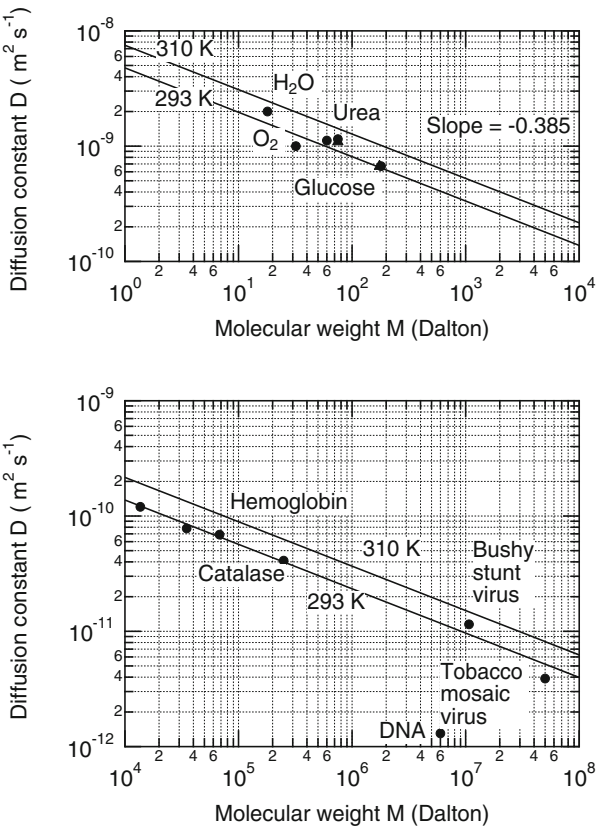


Fig. 4.12 Diffusion constant versus molecular weight in daltons. (One dalton is the mass of one hydrogen atom.) Data at 293 K are from Benedek and Villars (2000, Vol. 2, p. 122). The 293-K solid line was drawn by eye through the data; the line at 310 K was drawn parallel to it using the temperature change in Eq. 4.23. Data scatter around the line by about 30%, with occasional larger departures

If all of the molecules shown had the same density, then their radius would depend on $M^{1/3}$ and the line would have a slope of $-\frac{1}{3}$. The slope is steeper than this, suggesting that the molecules are larger for large M than constant density would predict. This increase in size may be partially attributable to water of hydration. The precise values of diffusion constants depend on many details of the particle structure; however, the lines in Fig. 4.12 provide an order-of-magnitude estimate.

The assumption that the flux depends linearly on the concentration gradient was an approximation. The diffusion constant is found, as a result, to be somewhat concentration dependent.

4.8 Fick's Second Law of Diffusion

Fick's first law of diffusion, Eq. 4.18a, is the observation that for small concentration gradients, the diffusive flux density is proportional to the concentration gradient: $j_x = -D\,\partial C/\partial x$. If this is differentiated, one obtains $\partial j_x/\partial x = -D\,\partial^2 C/\partial x^2$. Similar equations hold for the y and z directions. The equation of continuity, Eq. 4.2, is

$$-\frac{\partial C}{\partial t} = \frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y} + \frac{\partial j_z}{\partial z}.$$

If we combine these two equations, we get *Fick's second law of diffusion,* also known as the *diffusion equation*:

$$\frac{\partial C}{\partial t} = D\left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2}\right). \qquad (4.24)$$

The first law relates the flux of particles to the concentration gradient. The second law tells how the concentration at a point changes with time. It combines the first law and the equation of continuity. The function on the right-hand side of Eq. 4.24,

$$\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2},$$

is called the *Laplacian* of C. It is often abbreviated as $\nabla^2 C$ (read "del squared C") in American textbooks or ΔC in

European books. It is given in other coordinate systems in Appendix L.

In principle, if $C(x, y, z)$ is known at $t = 0$, Eq. 4.24 can be solved for $C(x, y, z, t)$ at all later times. (We develop a general, and sometimes useful, equation for doing this below.) We may also look at this equation as a local equation, telling how C changes with time at some point if we know how the concentration changes with position in the neighborhood of that point. The change of concentration with position determines the flux $\mathbf{j}$. The changes in flux with position determine how the concentration changes with time.

There is extensive literature on how to solve the diffusion equation (or the heat-flow equation, which is the same thing).[10] Instead of discussing a large number of techniques, we show by substitution that a Gaussian or normal distribution function, spreading in a certain way with time, is one solution to Eq. 4.24. In Sect. 4.14, we independently derive the same solution from a random-walk model of diffusion. An important feature of the Gaussian solution is that the center of the distribution of concentration does not move.

For simplicity, consider the one-dimensional case. Take the distribution to be centered at the origin and find those conditions under which[11]

$$C(x,t) = \frac{N}{\sqrt{2\pi}\sigma(t)} e^{-x^2/2\sigma^2(t)}. \tag{4.25}$$

We can view the one-dimensional case in either of two ways. If it represents diffusion along a pipe, then $C(x,t)$ is the number of particles per unit length in a slice between x and $x+dx$, and N is the total number of particles. If it represents a three-dimensional problem with concentration changing only in the x direction, then $C(x,t)$ is the number of particles per unit volume and N is the number of particles per unit area.

Eq. 4.25 is a solution to the one-dimensional version of Eq. 4.24:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}. \tag{4.26}$$

To check this, we will need various derivatives of Eq. 4.25. They can be evaluated using the chain rule:

$$\frac{\partial C}{\partial t} = \frac{N}{\sqrt{2\pi}}\left(-\frac{1}{\sigma^2}e^{-x^2/2\sigma^2} + \frac{x^2}{\sigma^4}e^{-x^2/2\sigma^2}\right)\frac{d\sigma}{dt},$$

$$\frac{\partial C}{\partial x} = -\frac{N}{\sqrt{2\pi}}e^{-x^2/2\sigma^2}\frac{x}{\sigma^3},$$

$$\frac{\partial^2 C}{\partial x^2} = \frac{N}{\sqrt{2\pi}}\left(-\frac{1}{\sigma^3}e^{-x^2/2\sigma^2} + \frac{x}{\sigma^3}e^{-x^2/2\sigma^2}\frac{x}{\sigma^2}\right).$$

When these are substituted in Eq. 4.26, the result is

$$\frac{N}{\sqrt{2\pi}\sigma^2}e^{-x^2/2\sigma^2}\left(-1+\frac{x^2}{\sigma^2}\right)\frac{d\sigma}{dt} = D\frac{N}{\sqrt{2\pi}\sigma^3}e^{-x^2/2\sigma^2}\left(-1+\frac{x^2}{\sigma^2}\right).$$

We can divide both sides of this equation by

$$\frac{N}{\sqrt{2\pi}\sigma^2}e^{-x^2/2\sigma^2}$$

because this factor is never zero. The result is

$$\left(\frac{x^2}{\sigma^2}-1\right)\frac{d\sigma}{dt} = \frac{D}{\sigma}\left(\frac{x^2}{\sigma^2}-1\right).$$

We can divide by $\left(x^2/\sigma^2 - 1\right)$ for all values of x except $x = \pm\sigma$. These values of x are where the second derivative of C vanishes; at these points, $\partial C/\partial t = 0$ for any value of σ. At all other points, the solution will satisfy the equation only if

$$\sigma\frac{d\sigma}{dt} = D.$$

This can be integrated to give

$$\int \sigma\, d\sigma = \int D\, dt$$

or

$$\frac{1}{2}\sigma^2(t) = Dt + \text{const.}$$

Multiply through by 2 and observe that $\sigma^2(0) = 2\text{const}$, so that

$$\sigma^2(t) = 2Dt + \sigma^2(0). \tag{4.27}$$

If the concentration is initially Gaussian with variance $\sigma^2(0)$, after time t it will still be Gaussian, centered on the same point, with a larger variance given by Eq. 4.27. Figure 4.13 shows this spreading in a typical case. At still earlier times, the concentration would have been even more narrowly peaked. In the limit when $\sigma(t)$ is zero, all the particles are at the origin, giving an infinite concentration. This is, of course, impossible. However, all the particles could be very close to the origin, giving a very tall, narrow curve for $C(x)$.

The width of the curve, determined by σ, increases as the square root of the time. A square-root increase is less rapid than a linear increase, reflecting the fact that as the particles spread out, the concentration does not change as rapidly with distance, so that the flux and the rate of spread decrease.

Note that the rate of change of concentration with time depends on the second derivative of the concentration with distance. This is because the rate of buildup is the flux into a region at some surface minus the flux out through a nearby

[10] See, for example, Crank (1975) or Carslaw and Jaeger (1959).

[11] The properties of the Gaussian function, Eq. 4.25, are discussed in Appendix I.

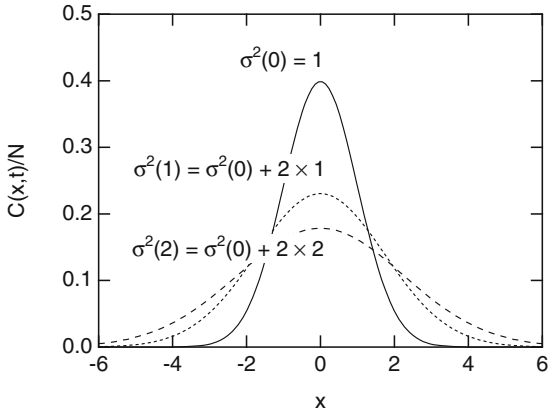


Fig. 4.13 Spreading of particles by diffusion assuming $D = 1$

surface; each flux is proportional to the gradient of the concentration, so the buildup is proportional to the difference in gradients or the second derivative.

In the problems at the end of this chapter, you will discover that diffusion of small particles through water for a distance of 1 μm takes about 1 ms, and diffusion through 100 μm takes 100^2 times as long, or 10 s. The times are even longer for larger particles. Thus, diffusion is an effective mode of transport for distances comparable to the size of a cell, but it is too slow for larger distances. This is why multicelled organisms evolve circulatory systems.

4.9 Time-Independent Solutions

In this section, we develop general solutions for diffusion and solvent drag when particles are conserved and the concentration and fluence rate are not changing with time. The system is in the steady state. The continuity equation, Eq. 4.8, then becomes div $\mathbf{j} = 0$. We consider the solutions for C and $\mathbf{j}$ in one, two, and three dimensions when the symmetry is such that $\mathbf{j}$ depends on only one position coordinate, x or r. These solutions are sometimes appropriate models for limited regions of space. There is always some other region of space, serving as a source or sink for the particles that are diffusing, where the model does not apply.

The behavior of $\mathbf{j}$ can be deduced from the continuity equation. In one dimension, such as flow in a pipe or between two infinite planes, the continuity equation is

$$\frac{dj_x}{dx} = 0, \tag{4.28}$$

which has a solution $j_x = b_1$ where b_1 is a constant. (The subscript denotes the constant for the one-dimensional case.)

The total flux or current i is constant, so

$$j_x = \frac{i}{S}, \tag{4.29}$$

where S is the area perpendicular to the flow.

In two dimensions, we consider a problem with cylindrical symmetry and consider only flow radially away from or toward the z-axis. In that case, the equation in Table L.1 for the divergence becomes

$$\frac{1}{r}\frac{d}{dr}(rj_r) = 0, \tag{4.30}$$

from which

$$\frac{d}{dr}(rj_r) = 0. \tag{4.31}$$

This means that (rj_r) is constant, or

$$j_r = \frac{b_2}{r}. \tag{4.32}$$

This is valid everywhere except along the z-axis, where there is a source of particles and the divergence is not zero. The total current i leaving a region of length L parallel to the z axis is also constant,

$$j_r = \frac{i}{2\pi Lr}. \tag{4.33}$$

In three dimensions with spherical symmetry, the radial component of the divergence is

$$\frac{1}{r^2}\frac{d}{dr}(r^2 j_r) = 0,$$

from which

$$\frac{d}{dr}(r^2 j_r) = 0, \tag{4.34}$$

so that

$$j_r = \frac{b_3}{r^2} \tag{4.35}$$

or

$$j_r = \frac{i}{4\pi r^2}. \tag{4.36}$$

This is valid everywhere except at the origin, where there is a source of particles.

These results depend only on continuity, time independence, and the assumed symmetry. They are true for diffusion, solvent drag, or any other process. Note the progression in going to higher dimensions: in n dimensions $r^{n-1}j_r$ is constant.

Now consider how the concentration varies in the two limiting cases of pure solvent drag and pure diffusion. (Sect. 4.12 discusses what happens when both transport modes are important.)

For solvent drag, the velocity of the solvent is the volume flux density $\mathbf{j}_v$, which also satisfies the continuity equation. In one dimension $j_v = i_v/S$. In two dimensions $j_v = i_v/2\pi Lr$, and in three dimensions $j_v = i_v/4\pi r^2$. In each case

$$C_s = \frac{j_s}{j_v} = \frac{i_s}{i_v}. \tag{4.37}$$

Since C_s is constant, there is no diffusion.

For the case of diffusion, $\mathbf{j} = -D\nabla C$. In one dimension this becomes

$$\frac{dC}{dx} = -\frac{i}{SD},$$

which is integrated to give

$$C = -\frac{i}{SD}x + b_1,$$

where b_1 is the constant of integration. The concentration varies linearly in the one-dimensional case. If i is positive (flow in the $+x$ direction), C decreases as x increases. Often the concentration is known at x_1 and x_2, and one wants to know the current. We can write

$$C_1 = -\frac{i}{SD}x_1 + b_1,$$
$$C_2 = -\frac{i}{SD}x_2 + b_1,$$

and solve for i:

$$i = \frac{(C_1 - C_2)}{(x_2 - x_1)}SD. \tag{4.38a}$$

In two dimensions

$$\frac{dC}{dr} = -\frac{i}{2\pi LD}\frac{1}{r},$$

and the solution is

$$C(r) = -\frac{i}{2\pi LD}\ln r + b_2.$$

We can again solve for the current when the concentrations are known at two different radii:

$$i = \frac{2\pi LD(C_1 - C_2)}{\ln(r_2/r_1)} = \frac{2\pi LD(C_2 - C_1)}{\ln(r_1/r_2)}. \tag{4.38b}$$

Diffusion in two dimensions with cylindrical symmetry has been used to model the concentration of substances in the region between two capillaries.

In three dimensions, the diffusion equation is

$$\frac{dC}{dr} = -\frac{i}{4\pi Dr^2},$$

which has a solution

$$C(r) = \frac{i}{4\pi Dr} + b_3.$$

The current in terms of the concentration is

$$i = \frac{4\pi D\,[C(r_1) - C(r_2)]}{1/r_1 - 1/r_2}. \tag{4.38c}$$

The three-dimensional case is worth further discussion, because it can help us to understand the diffusion of nutrients to a single spherical cell or the diffusion of metabolic waste products away from the cell. Consider the case in which the cell has radius $r_1 = R$, the concentration at the cell surface is C_0, and the concentration at infinity is zero. Then

$$i = 4\pi D\,C_0 R, \tag{4.39a}$$
$$C(r) = \frac{C_0 R}{r}, \tag{4.39b}$$
$$j_r = \frac{C_0 DR}{r^2}. \tag{4.39c}$$

The particle current depends on the radius of the cell, R, not on R^2. This very important result is not what we might naively expect. Diffusion-limited flow of solute in or out of the cell is proportional not to the cell surface area, but to the cell radius. The reason is that the particle movement is limited by diffusion in the region around the cell, and as the cell radius increases, the concentration gradient decreases. (It is possible for the rate of particle migration into the cell to be proportional to the surface area of the cell if some other process, such as transport through the cell membrane, is the rate-limiting step.)

If diffusion is toward the cell, the concentration is C_0 infinitely far away. At the cell surface, every diffusing molecule that arrives is assumed to be captured, and the concentration is zero. The solutions are then

$$i = -4\pi DC_0 R, \tag{4.40a}$$
$$C(r) = C_0\,(1 - R/r)\,, \tag{4.40b}$$
$$j_r(r) = -\frac{C_0 DR}{r^2}. \tag{4.40c}$$

4.10 Example: Steady-State Diffusion to a Spherical Cell and End Effects

In the preceding section, we considered diffusion from infinitely far away to the surface of a spherical cell where the concentration was zero. We now add the effect of steady-state diffusion through a series of pores or channels in the cell membrane. This will lead to a very important result: it does not require very many pores per unit area in the cell membrane to "keep up with" the rate of diffusion of chemicals toward or away from the cell. The result is important for understanding how cells acquire nutrients, how bacteria move in response to chemical stimulation (chemotaxis), and how the leaves of plants function.

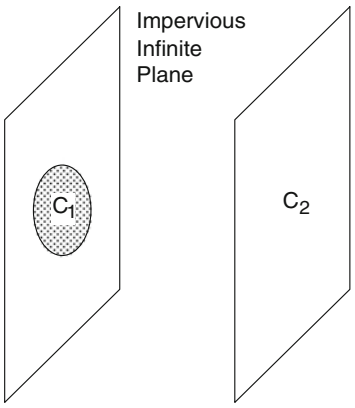


Fig. 4.14 The diffusion flux from the disk of radius a and concentration C_1 to the infinite sheet where the concentration is C_2 is given by $i = 4Da(C_1 - C_2)$

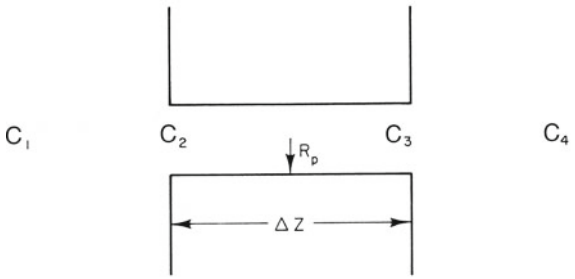


Fig. 4.15 End effects in diffusion through a pore

To develop the model we need one more result: the current due to diffusion from a disk of radius a where the concentration is C_1 to a plane far away where the concentration is C_2. The disk is embedded in the surface of an impervious plane as shown in Fig. 4.14, so particles cannot cross to the region behind the disk. The current is (Eq. 6.97)

$$i = 4Da\,(C_1 - C_2). \tag{4.41}$$

It is proportional to the radius of the disk, not its surface area. (Obtaining this result requires solving the diffusion equation in three dimensions. See Carslaw and Jaeger (1959), p. 215.)

Consider diffusion through a pore of radius R_p which pierces a membrane of thickness ΔZ, including diffusion in the medium on either side of the membrane (Fig. 4.15). If the material on either side were well stirred, there would be a uniform concentration C_1 on the left and C_4 on the right. Because it is not stirred, there is diffusion in the exterior fluid. Let C_1 and C_4 be measured far away, and call the concentrations at the ends of the pore C_2 on the left and C_3 on the right.

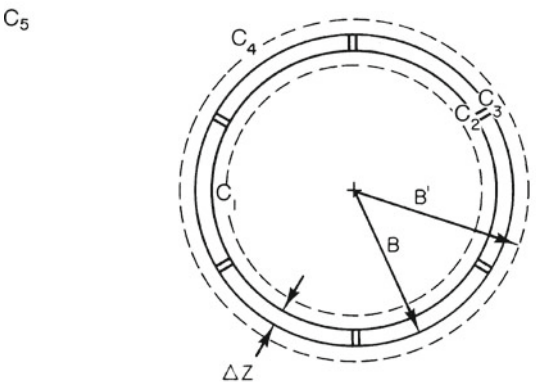


Fig. 4.16 Diffusive end effects for a spherical cell pierced by pores

Equation 4.38a gives the diffusion flux within the pore

$$i = \frac{\pi R_p^2 D\,(C_2 - C_3)}{\Delta Z}. \tag{4.42}$$

Diffusion from C_1 to C_2 is given by Eq. 4.41. It is

$$i = 4D\,R_p(C_1 - C_2), \tag{4.43}$$

while from C_3 to C_4, it is

$$i = 4D\,R_p(C_3 - C_4). \tag{4.44}$$

In the steady state, there is no buildup of particles and i is the same in each region. We can solve Eqs. 4.42–4.44 simultaneously to relate i to concentrations C_1 and C_4:

$$i = \frac{\pi R_p^2\,D}{\Delta Z + 2\pi R_p/4}(C_1 - C_4). \tag{4.45}$$

This has the same form as Eq. 4.42, except that the membrane thickness has been replaced by an effective thickness

$$\Delta Z' = \Delta Z + 2\frac{\pi R_p}{4}. \tag{4.46}$$

An extra length $\pi R_p/4$ has been added at each end to correct for diffusion in the unstirred layer on each side of the pore. This correction is important when the pore length is less than two or three times the pore radius.

Now consider diffusion in or out of the spherical cell shown in Fig. 4.16. The radius of the cell is B. The membrane has thickness ΔZ and is pierced by a total of N pores, each of radius R_p. Within the cell, we do not know the details of the concentration distribution, since they depend on what sort of chemical reactions are taking place and where. But we will assume that at the radius where diffusion to the pores becomes important, the concentration is C_1. At the inner face of each pore it is C_2, at the outer face it is C_3, and

over an approximately spherical surface of radius B' it is C_4. Far away, the concentration is C_5. As a result, there are four separate regions in which we must consider diffusion. The first is from C_1 to the opening of each pore; the second is through the pore; third, there is diffusion from the outer face of each pore to C_4; and, finally, there is diffusion from the spherical object of radius B' to the surrounding medium.

4.10.1 Diffusion Through a Collection of Pores, Corrected

The first three processes are taken into account by applying the end correction to each end of the pores. The flow through one pore is, using Eq. 4.45,

$$i_{\text{pore}} = \frac{\pi R_p^2 D}{\Delta Z'}(C_1 - C_4), \tag{4.47}$$

where $\Delta Z'$ is given by Eq. 4.46. Since there are N pores in all, the total flow through the cell membrane is

$$i_{\text{cell}} = N i_{\text{pore}} = \frac{N\pi R_p^2 D}{\Delta Z'}(C_1 - C_4). \tag{4.48}$$

The diffusion from C_4 to infinity is given by Eq. 4.38c.

$$i_{\text{cell}} = 4\pi D\, B'(C_4 - C_5), \tag{4.49}$$

where B' is the effective radius for diffusion to the surrounding medium. It is slightly larger than B. If we equate Eqs. 4.48 and 4.49, solve for C_4 and substitute this result back in Eq. 4.49, we get

$$i_{\text{cell}} = \frac{4\pi D B' N R_p^2}{N R_p^2 + 4B'\Delta Z'}(C_1 - C_5). \tag{4.50}$$

This can be rewritten as

$$i_{\text{cell}} = \frac{N\pi R_p^2 D}{\Delta Z_{\text{eff}}}(C_1 - C_5), \tag{4.51}$$

where

$$\Delta Z_{\text{eff}} = \Delta Z + 2\frac{\pi R_p}{4} + N\frac{R_p^2}{4B'}. \tag{4.52}$$

The first term in ΔZ_{eff} is the membrane thickness. The second term corrects for diffusion from the end of each pore to the surrounding fluid; the last corrects for diffusion away from the cell into the surrounding medium. The third term can be expressed as

$$\frac{N R_p^2}{4B'} = \frac{B}{B'} B f,$$

where

$$f = \frac{N\pi R_p^2}{4\pi B^2} \tag{4.53}$$

is the fraction of the cell surface occupied by pores.

We now assume that $B = B'$. (Problem 33 shows that the difference is usually very small.) The effective pore length is then

$$\Delta Z_{\text{eff}} = \Delta Z + 2\left(\frac{\pi R_p}{4}\right) + Bf. \tag{4.54}$$

Equations 4.51–4.54 treat the problem as diffusion through a collection of N pores, corrected for diffusion outside the pore by increasing the length of the pore.

4.10.2 Diffusion from a Sphere, Corrected

It is also useful to write these results as the equation for diffusion to or from a sphere, Eqs. 4.39, corrected for the diffusion through the cell membrane. Writing it in this form gives us insight into how much of the cell membrane must be occupied by pores for efficient particle transfer. Solve Eq. 4.53 for $N R_p^2$ and substitute the result in Eq. 4.50. The result is

$$\begin{aligned} i_{\text{cell}} &= \frac{4\pi D\, B' B^2 f\,(C_1 C_5)}{B^2 f + B'\Delta Z'} \\ &= 4\pi B D\,(C_1 - C_5)\left(\frac{B'}{B}\right)\frac{f}{f + (B'/B)(\Delta Z'/B)}. \end{aligned} \tag{4.55}$$

This has the form of diffusion to the sphere multiplied by a correction factor. With B'/B again approximated by unity, the correction factor is

$$\frac{f}{f + \Delta Z'/B}.$$

The correction factor is zero when f is zero and becomes nearly unity when the entire cell surface is covered by pores.

4.10.3 How Many Pores Are Needed?

We now ask what fraction of the cell's surface area must be occupied by pores. The cell will receive half the maximum possible diffusive flow when the fraction $f = \Delta Z'/B$. For a typical cell with $B = 5$ μm and $\Delta Z = 5$ nm, $f = 0.001$. This is a surprisingly small number, but it means that there is plenty of room on the cell surface for different kinds of pores. There are two ways to understand why this number is so small. First, we can regard the ratio of concentration difference to flow as a resistance, analogous to electrical resistance. The total resistance from the inside of the cell to infinity is made up of the resistance from the outside of the cell to infinity plus the resistance of the parallel combination of N pores. Once the resistance of this parallel combination is equal to the resistance from the cell to infinity, adding more pores in parallel does not change the overall resistance

very much. The second way to look at it is in terms of the random walks of the diffusing solute molecules. Once a solute molecule has diffused into the neighborhood of the cell, it undergoes many random walks. When it strikes the cell membrane, it wanders away again, to return shortly and strike the cell membrane someplace else. If the first contact is not at a pore, there are more opportunities to strike a pore on a subsequent contact with the surface.

4.10.4 Other Applications of the Model

The same sort of analysis that we have made here can be applied to a plane surface area, such as the underside of a leaf (Meidner and Mansfield 1968) and to a cylindrical geometry, such as a capillary wall.

The analysis can also be applied to the problem of bacterial chemotaxis—the movement of bacteria along concentration gradients. This problem has been discussed in detail by Berg and Purcell (1977).[12] The cell detects a chemical through some sort of chemical reaction between the chemical and the cell. Suppose that the reaction takes place between the chemical and a binding site of radius R_p on the surface of the cell. We want to know what fraction of the surface area of the cell must be covered by binding sites. This is similar to the diffusion problem of Eq. 4.55, except that if the binding site is on the surface of the cell, there is no diffusion through a pore of length ΔZ. The effective pore length $\Delta Z'$ is just the end correction for one end of the pore, $\pi R_p/4$. Half of the maximum possible flow to the binding site occurs when

$$f = \pi R_p/4B.$$

A typical bacterium might have a radius $B = 1$ μm; the binding site might have a radius of a few atoms or 1 nm. With these values $f = 7.9 \times 10^{-4}$. The number of sites would be $f4\pi B^2/\pi R_p^2 = \pi B/R_p = 3000$. There is plenty of room on the cell surface for many different binding sites, each specific for a particular chemical.

An *E. coli* cell typically travels 10–20 body lengths per second. It detects concentration gradients as changes with time. Because of this, Berg and Purcell concluded that a uniform distribution of chemoreceptors over the surface of the cell would be optimum. It would give the highest probability of capture of a chemical molecule that wandered near the cell. However, studies of *E. coli* have shown that the receptors are located near the poles of the cell [Maddock and Shapiro (1993); see also the comment by Parkinson and Blair (1993), who point out that the reduced efficiency of sensors could make sense if "eating" or transport into the cell is more important than "smelling."]

[12] See also Berg (1975, 1983) and Purcell (1977).

The Berg–Purcell model has been extended to provide a time-dependent solution and allow the receptors not to be perfectly absorbing (Zwanzig and Szabo 1991) and also to have a process in which the molecules attach to the membrane and then diffuse in the two-dimensional membrane surface (Wang et al. 1992; Axelrod and Wang 1994).

4.11 Example: A Spherical Cell Producing a Substance

Here is a simple model that extends the arguments of Sect. 4.9 to develop a steady-state solution for a spherical cell excreting metabolic products. The cell has radius R. The concentration of some substance inside the cell is $C(r)$, independent of time t and the spherical coordinate angles θ and ϕ. (Spherical coordinates are described in Appendix L.) The substance is produced at a constant rate Q particles per unit volume per second throughout the cell and leaves through the surface of the cell at a constant fluence rate $j(R)$, independent of t, θ, and ϕ. Assume that all transport is by pure diffusion and the diffusion constant for this substance is D everywhere inside and outside the cell. The material inside the cell is not well stirred. (For this model we assume that the cell membrane does not affect the transport process. We could make the model more complicated by introducing the features described in Sect. 4.10.) With these assumptions, the cell can be modeled as an infinite homogeneous medium with diffusion constant D that contains a spherical region producing material at rate Q per unit volume per second.

We first find the concentration $C(r)$ inside and outside the cell by using a technique that only works because of the spherical symmetry. We use the continuity equation in the form Eq. 4.10b. Because the concentration is not changing with time, the total amount of material flowing through a spherical surface of radius r is equal to the amount produced within that sphere. For $r < R$

$$4\pi r^2 j(r) = 4\pi r^3 Q/3,$$
$$j(r) = Qr/3.$$

For $r > R$

$$4\pi r^2 j(r) = 4\pi R^3 Q/3,$$
$$j(r) = QR^3/3r^2.$$

Using the fact that $j(r) = -DdC/dr$, we obtain for $r < R$

$$\frac{dC}{dr} = -\frac{Q}{3D}r,$$
$$C(r) = -\frac{Qr^2}{6D} + b_1,$$

where b_1 is the constant of integration. For $r > R$,

$$\frac{dC}{dr} = -\frac{QR^3}{3Dr^2},$$
$$C(r) = \frac{QR^3}{3Dr} + b_2.$$

The fact that the concentration must be zero far from the cell means that $b_2 = 0$. Matching the two expressions at $r = R$ gives

$$-QR^2/6D + b_1 = QR^2/3D,$$
$$b_1 = QR^2/2D,$$

so that

$$C(r) = \begin{cases} \dfrac{Q}{6D}(3R^2 - r^2), & r \le R \\ \dfrac{QR^3}{3Dr}, & r \ge R. \end{cases}$$

The other method is more general and can be extended to problems that do not have spherical symmetry. We find solutions to Fick's second law, modified to include the production term Q and with the concentration not changing with time:

$$0 = \frac{\partial C}{\partial t} = D\nabla^2 C + Q,$$
$$\nabla^2 C = -\frac{Q}{D}.$$

In spherical coordinates (Appendix L; Schey 2004) this is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial C}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial C}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\left(\frac{\partial^2 C}{\partial\phi^2}\right) = -\frac{Q}{D}.$$

Since there is no angular dependence, we have separate equations for each domain:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dC}{dr}\right) = \begin{cases} -\dfrac{Q}{D}, & r < R \\ 0, & r > R. \end{cases}$$

It is necessary to solve each equation in its domain, and then at the boundary require that C be continuous and also that j and therefore dC/dr be continuous. For $r < R$ we get the following (b_1 and b_2 are constants of integration):

$$r^2\frac{dC}{dr} = -\frac{Qr^3}{3D} + b_1,$$
$$\frac{dC}{dr} = -\frac{Qr}{3D} + \frac{b_1}{r^2},$$
$$C(r) = -\frac{Qr^2}{6D} - \frac{b_1}{r} + b_2.$$

Since the concentration is finite at the origin, $b_1 = 0$:

$$C(r) = b_2 - \frac{Qr^2}{6D}, \qquad r < R.$$

For $r > R$, we can use the general solution with $Q = 0$ and different constants:

$$C(r) = -\frac{b_1'}{r} + b_2'.$$

Far away, the concentration is zero, so $b_2' = 0$. Matching dC/dr at the boundary gives

$$-\frac{QR}{3D} = \frac{b_1'}{R^2}, \quad b_1' = -Q\frac{R^3}{3D}.$$

Matching $C(r)$ at the boundary gives

$$-\frac{QR^2}{6D} + b_2 = -\frac{b_1'}{R}.$$

Putting all of this together gives the same expression for the concentration we had earlier. This technique is a bit more cumbersome, but there are many mathematical tools to extend this technique to cases where there is not spherical symmetry and where Q is a function of position. These advanced techniques can also be used when C is changing with time.

4.12 Drift and Diffusion in One Dimension

The particle fluence rate due to diffusion in one dimension is $j_{\text{diff}} = -D\,(\partial C/\partial x)$. That of particles drifting with velocity v is $j_{\text{drift}} = vC$. The total flux density or fluence rate is the sum of both terms:

$$j_s = -D\frac{\partial C}{\partial x} + vC. \tag{4.56}$$

The homogeneous ($j_s = 0$) solution was discussed in Sect. 4.7, where cancellation of these two terms in equilibrium was used to derive the relationship between the diffusion constant and viscosity. Using the techniques of Appendix F, we can write the homogeneous solution as

$$C(x) = Ae^{(v/D)x}. \tag{4.57}$$

This can be used to solve the problem of $j_s = \text{const}$ when the concentration is C_0 at $x = 0$ and C_0' at $x = x_1$. $C(x)$ must vary in such a way that the total flux density, the sum of the diffusive and drift terms, is constant. Suppose both terms give flow from left to right. If the concentration is high, then

the drift flux density is large and the concentration gradient must be small. If the concentration is small, the diffusive flux, and hence the gradient, must be large. To develop a formal solution, write Eq. 4.56 as

$$\frac{dC}{dx} - \frac{1}{\lambda}C = -\frac{j_s}{D}, \tag{4.58}$$

where $\lambda = D/v$ has the dimensions of length and can be interpreted as the distance over which diffusion is important. If the velocity is zero, diffusion is important everywhere and $\lambda = \infty$. If the velocity is very large, $\lambda \to 0$. Since v can be either positive or negative, so can λ. A particular solution to Eq. 4.58 is

$$C(x) = \frac{\lambda j_s}{D} = \frac{j_s}{v}.$$

The general solution is the sum of the particular solution and the homogeneous solution, Eq. 4.57:

$$C(x) = Ae^{x/\lambda} + j_s/v. \tag{4.59}$$

The situation is slightly different than what we encountered in Chap. 2. We must determine two constants, A and j_s, given the two concentrations C_0 and C_0'. Writing Eq. 4.59 for $x = 0$ and for $x = x_1$, we obtain

$$\begin{aligned} C_0 &= A + \frac{j_s}{v}, \\ C_0' &= Ae^{x_1/\lambda} + \frac{j_s}{v}. \end{aligned} \tag{4.60}$$

Subtracting these gives

$$\begin{aligned} C_0' - C_0 &= A(e^{x_1/\lambda} - 1), \\ A &= (C_0' - C_0)/(e^{x_1/\lambda} - 1). \end{aligned} \tag{4.61}$$

This can be combined with either of Eqs. 4.60 to give

$$j_s = \frac{C_0 e^{x_1/\lambda} - C_0'}{e^{x_1/\lambda} - 1} v. \tag{4.62}$$

We can also substitute Eqs. 4.61 and 4.62 in 4.59 to obtain an expression for $C(x)$. The result is

$$C(x) = \frac{C_0(e^{x_1/\lambda} - e^{x/\lambda}) + C_0'(e^{x/\lambda} - 1)}{e^{x_1/\lambda} - 1}. \tag{4.63}$$

We will discuss the implications of this equation below.

Let us first determine the average concentration between $x = 0$ and $x = x_1$. The average concentration is defined by

$$\overline{C} = \frac{1}{x_1}\int_0^{x_1} C(x)\,dx. \tag{4.64}$$

While one could integrate this directly, it is much easier to integrate Eq. 4.56 from 0 to x_1:

$$-D\int_0^{x_1}\left(\frac{dC}{dx}\right)dx + v\int_0^{x_1} C(x)\,dx = +j_s\int_0^{x_1} dx.$$

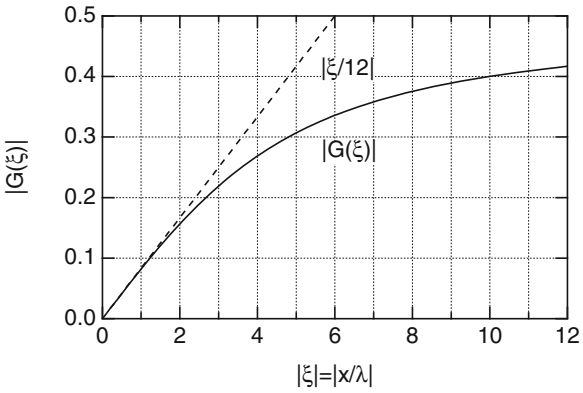


Fig. 4.17 The correction factor $G(\xi)$ used in Eq. 4.68. The dashed line is the approximation $G(\xi) = \xi/12$, which is valid for small ξ and is used in Eq. 4.67

The first term is $-D(C_0' - C_0)$. The second is $vx_1\overline{C}$. The third is $j_s x_1$. The equation can therefore be rewritten as

$$v\overline{C} = \frac{D\,(C_0' - C_0)}{x_1} + j_s. \tag{4.65}$$

Substituting Eq. 4.62 for j_s gives the average concentration

$$\overline{C} = \frac{C_0 e^{x_1/\lambda} - C_0'}{e^{x_1/\lambda} - 1} - \frac{\lambda}{x_1}(C_0 - C_0'). \tag{4.66}$$

The exponentials can be expanded to give an approximate expression for small values of x_1/λ[13]

$$\overline{C} = \frac{(C_0 + C_0')}{2} + \frac{x_1}{\lambda}\frac{1}{12}(C_0 - C_0'). \tag{4.67}$$

For larger values of x_1/λ, the mean can be written

$$\overline{C} = \frac{(C_0 + C_0')}{2} + (C_0 - C_0')\,G\left(\frac{x_1}{\lambda}\right). \tag{4.68}$$

The correction factor $G(x_1/\lambda) = G(\xi)$, given by

$$G(\xi) = \frac{1}{2}\frac{e^{\xi} + 1}{e^{\xi} - 1} - \frac{1}{\xi}, \tag{4.69}$$

is plotted in Fig. 4.17. The function is odd, and only values for $\xi \geq 0$ are shown. For $\xi = 0$ ($\lambda = \infty$, pure diffusion), the average concentration is $(C_0 + C_0')/2$.

Figure 4.18 shows the concentration profile calculated from Eq. 4.63. The concentration is 5 times larger on the left, so diffusion is from left to right. When $x_1/\lambda = x_1 v/D = 0.8$, drift is also from left to right. As the concentration falls, the magnitude of the gradient rises, so that the sum of the diffusive and drift fluxes remains the same. When $x_1/\lambda = -0.8$,

[13] See Levitt (1975, p. 537). For $x_1/\lambda = 1.5$, this approximation is within 1%. For $x_1/\lambda = 2.5$, the error is about 6 %.

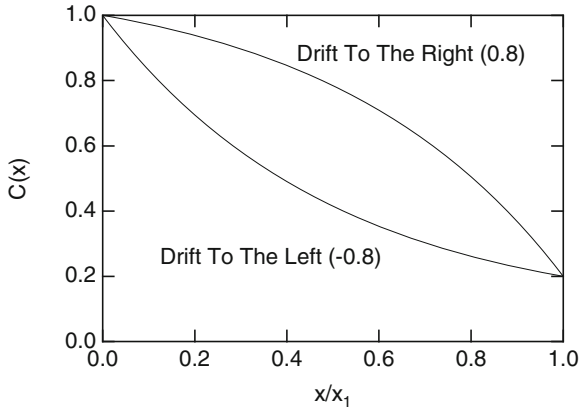


Fig. 4.18 Concentration profile for combined drift and diffusion. The concentration is 1.0 on the left and 0.2 on the right. For $x_1/\lambda = x_1 v/D = 0.8$, drift and diffusion are both to the right. As the concentration falls, the magnitude of the gradient increases. For $x_1/\lambda = x_1 v/D = -0.8$ drift opposes diffusion. As the concentration falls, so does the magnitude of the gradient

drift is opposite to diffusion. Therefore, both the concentration and the magnitude of the gradient must rise and fall together to keep total flux density constant.

Equation 4.65 can be rewritten as

$$j_s = \frac{-D\,(C_0' - C_0)}{x_1} + v\overline{C}. \tag{4.70}$$

This can be interpreted as meaning that the fluence rate is given by the sum of a diffusion term with the average concentration gradient and a drift term with the average concentration. However, the discussion in the preceding paragraph showed that there is actually a continuous change of the relative size of the diffusion and drift terms for different values of x.

4.13 A General Solution for the Particle Concentration as a Function of Time

If $C(x, 0)$ is known for $t = 0$, it is possible to use the result of Sect. 4.8 to determine $C(x, t)$ at any later time. The key to doing this is that if $C(x, t)\,dx$ is the number of particles in the region between x and $x + dx$ at time t, it may be be interpreted as the probability of finding a particle in the interval (x, dx) multiplied by the total number of particles. (Recall the discussion on p. 96 about the interpretation of $C(x, t)$.) The spreading Gaussian then represents the spread of probability that a particle is between x and $x + dx$.

If a particle is definitely at $x = \xi$ at $t = 0$, then $\sigma^2(0) = 0$. The particle cannot remain there because of equipartition of energy: collisions cause it to acquire a mean square velocity

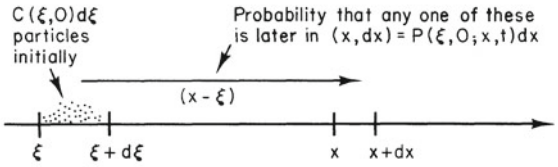


Fig. 4.19 Diffusion from ξ to x

$3k_B T/m$ and move. Some time later

$$\sigma(t) = (2Dt)^{1/2}. \tag{4.71}$$

Define $P(\xi, 0; x, t)\,dx$ to be the probability that a particle has diffused to a location between x and $x + dx$ at time t, if it was at $x = \xi$ when $t = 0$. This probability is given by Eq. 4.25, except that the distance it has diffused is now $x - \xi$ instead of x. The variance $\sigma^2(t)$ is given by Eq. 4.71. The result is

$$P(\xi, 0; x, t)\,dx = \frac{1}{\sqrt{4\pi Dt}} e^{-(x-\xi)^2/4Dt}\,dx. \tag{4.72}$$

The number of particles initially between $x = \xi$ and $x = \xi + d\xi$ is the concentration per unit length times the length of the interval $N = C(\xi, 0)d\xi$, as shown in Fig. 4.19.

The particles can diffuse in either direction. At a later time t, the average number between x and $x + dx$ that came originally from between $x = \xi$ and $x = \xi + d\xi$ is the original number in $(\xi, d\xi)$ times the probability that each one got from there to x. This number is a differential of a differential, $d\,[C(x, t)dx]$, because it is only that portion of the particles in dx that came from the interval $d\xi$:

$$d\,[C(x, t)\,dx] = C(\xi, 0)\,d\xi \frac{1}{\sqrt{4\pi Dt}} e^{-(x-\xi)^2/4Dt}\,dx.$$

To get $C(x, t)dx$, it is necessary to integrate over all possible values of ξ:

$$C(x, t)\,dx = \frac{1}{\sqrt{4\pi Dt}} \left[\int_{-\infty}^{\infty} C(\xi, 0) e^{-(x-\xi)^2/4Dt}\,d\xi\right] dx. \tag{4.73}$$

This equation can be used to find $C(x, t)$ at any time, provided that $C(x, t)$ was known at some earlier time. The factor that multiplies $C(\xi, 0)$ in the integrand is called the *influence function* or *Green's function* for the diffusion problem; it gives the relative weighting of $C(\xi, 0)$ in contributing to the later value $C(x, t)$.

As an example of using this integral, consider a situation in which the initial concentration has a constant value C_0 from $\xi = -\infty$ to $\xi = 0$ and zero for all positive ξ, as shown

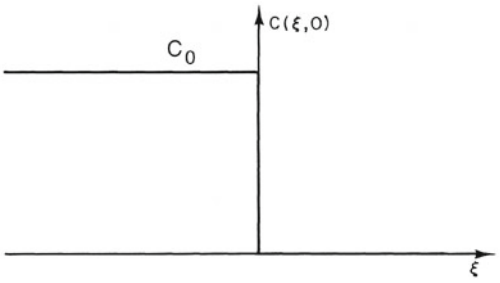


Fig. 4.20 The initial concentration is constant to the left of the origin and zero to the right of the origin

in Fig. 4.20. At $t = 0$ the diffusion starts. The concentration at later times is given by

$$C(x,t) = \frac{C_0}{\sqrt{4\pi Dt}} \int_{-\infty}^{0} e^{-(x-\xi)^2/4Dt}\, d\xi.$$

Such integrals are most easily evaluated by using the *error function*, defined by

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \tag{4.74}$$

The error function is plotted in Fig. 4.21. One must be careful in using tables, which may be for related functions that differ in normalization constants or the limits of integration.

To use the error function in evaluating the integral in Eq. 4.73, make the substitution $s = (x - \xi)/(4Dt)^{1/2}$. The integral becomes

$$C(x,t) = \frac{-C_0}{\sqrt{4\pi Dt}} \int_{\infty}^{x/\sqrt{4Dt}} e^{-s^2} \sqrt{4Dt}\, ds.$$

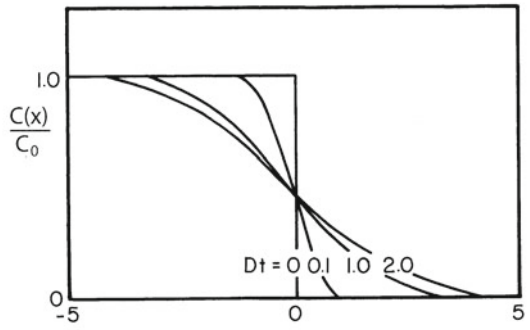


Fig. 4.22 The spread of an initially sharp boundary due to diffusion

Since $\int_A^B f(x)\,dx = \int_0^B f(x)\,dx + \int_A^0 f(x)\,dx = \int_0^B f(x)\,dx - \int_0^A f(x)\,dx$, this can be written as

$$C(x,t) = \frac{-C_0}{\sqrt{\pi}} \left(\int_0^{x/\sqrt{4Dt}} e^{-s^2} ds - \int_0^{\infty} e^{-s^2} ds \right)$$
$$= \frac{C_0}{2} \left[1 - \operatorname{erf}(x/\sqrt{4Dt}) \right]. \tag{4.75}$$

The plot in Fig. 4.22 shows how the initially sharp concentration step becomes more diffuse with passing time. Quantitative measurements of the concentration can be used to determine D. Benedek and Villars (2000, pp. 126–136) discuss some experiments to verify the solution we have obtained above and to determine D.

Many other solutions to the diffusion equation and techniques for solving it are known. See Crank (1975) or Carslaw and Jaeger (1959).

4.14 Diffusion as a Random Walk

The spreading solution to the one-dimensional diffusion equation that we verified can also be obtained by treating the

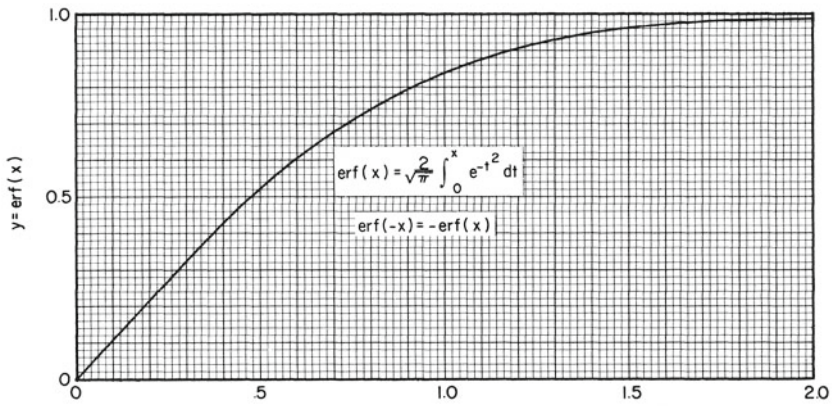


Fig. 4.21 Plot of the error function erf(x)

motion of a molecule as a series of independent steps either to the right or to the left along the x axis. (The same treatment can be extended to three dimensions, but we will not do so.) The derivation gives us a somewhat simplified molecular picture of diffusion. The derivation also provides an opportunity to see how the Gaussian distribution approximates the binomial distribution. This section is not necessary to understand the rest of Chaps. 4 and 5, and you should tackle it only if you are familiar with the binomial and Gaussian probability distributions (Appendices H and I). The model is more restrictive than the diffusion equation derived above, since the latter is the linear approximation to the transport problem.

We use a simplified model in which the diffusing particle always moves in steps of length λ (the mean free path), either in the $+x$ or $-x$ direction. Let the total number of steps taken by the particle be N, of which n are to the right and n' are to the left: $N = n + n'$. Also let $m = n - n'$. The particle's net displacement in the $+x$ direction is then

$$n\lambda - n'\lambda = m\lambda.$$

Since the steps are independent and a step to the left or right is equally likely ($p = 1/2$), the probability of having a displacement $m\lambda$ is given by the binomial probability $P(n; N)$:

$$P(n; N) = \frac{N!}{(n!)(N-n)!}\left(\frac{1}{2}\right)^n\left(\frac{1}{2}\right)^{n'}. \qquad (4.76)$$

Since this problem is analogous to a coin being tossed, and we know that on the average we get the same number of heads (steps to the left) as tails (steps to the right), we know that the distribution is centered at $n = n'$ or $m = 0$. We also know [Eq. G.4] that the variance in n is given by $\overline{n^2} - \overline{n}^2 = Npq = N/4$. Since $\overline{n} = N/2$, this says that $\overline{n^2} = N/4 + N^2/4$. However, we need the variance in m, $\overline{m^2} - \overline{m}^2$. To obtain it, we write $m = 2n - N$ and $m^2 = 4n^2 + N^2 - 4nN$. Therefore,

$$\overline{m^2} = 4\overline{n^2} + N^2 - 4N\overline{n} = N.$$

The variance of the distribution of displacement x is equal to the step length λ times the variance in the number of steps:

$$\sigma^2 = \overline{x^2} = \lambda^2\overline{m^2} = \lambda^2 N.$$

The number of steps is the elapsed time divided by the collision time $N = t/t_c$. Therefore,

$$\sigma^2 = \frac{\lambda^2 t}{t_c}.$$

Comparing this with Eq. 4.71, we identify $D = \lambda^2/2t_c$, so that

$$\sigma^2 = 2Dt. \qquad (4.77)$$

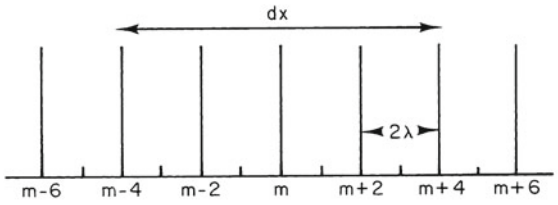


Fig. 4.23 Relationship between the values of x and the allowed values of m. Every other value of m is missing

We have shown that this simple model gives a distribution with fixed mean which spreads with a variance proportional to t. We now must show that the shape is Gaussian. Appendix I shows that the Gaussian is an approximation to the binomial distribution in the limit of large N. Since $\sigma_n^2 = N/4$ and $\overline{n} = N/2$, Eq. G.4 can be used to write

$$P(n) = \left(\frac{2\pi N}{4}\right)^{-1/2} e^{-(n-N/2)^2/(2N/4)}.$$

This can be rewritten in terms of the net number of steps to the right, since $m = n - n' = 2n - N$:

$$P(m) = \left(\frac{2}{\pi N}\right)^{1/2} e^{-m^2/2N}.$$

Note that only every other value of m is allowed. Since $m = 2n - N$, m goes in steps of 2 from $-N$ to N as n goes from 0 to N.

To write the probability distribution in terms of x and t, refer to Fig. 4.23. The spacing between each allowed value of x is 2, so that the number of allowed values of m in interval $(x, x + dx)$ is $dx/2\lambda$. Therefore, $P(x)\,dx = P(m)(dx/2\lambda)$,

$$P(x) = \sqrt{\frac{2}{\pi N 4\lambda^2}}e^{-m^2/2N}.$$

With the substitutions $m = x/\lambda$ and $N = t/t_c$, this becomes

$$P(x,t) = \sqrt{\frac{t_c}{2\pi\lambda^2 t}}e^{-x^2(t_c/2\lambda^2 t)}.$$

With the substitutions $D = \lambda^2/2t_c$ and $C(x,t) = C(0)P(x,t)$, we obtain Eq. 4.25.

The result of Eq. 4.71 is easily extended to two dimensions. Imagine that a total of N steps are taken, half in the x direction and half in the y direction. Then $\sigma_x^2 = \sigma_y^2 = \lambda^2(N/2)$. If $r^2 = x^2 + y^2$, $\sigma_r^2 = \sigma_x^2 + \sigma_y^2 = \lambda^2 N$. We still define D in any direction as $\lambda^2/2t_c$, where t_c is the time between steps in that direction. After a total time t, N steps have been taken, but only half of them were in, say, the x direction. Therefore $t_c = 2t/N$. Therefore

$$\sigma_r^2 = \sigma_x^2 + \sigma_y^2 = 4Dt \quad \text{(two dimensions)}. \qquad (4.78)$$

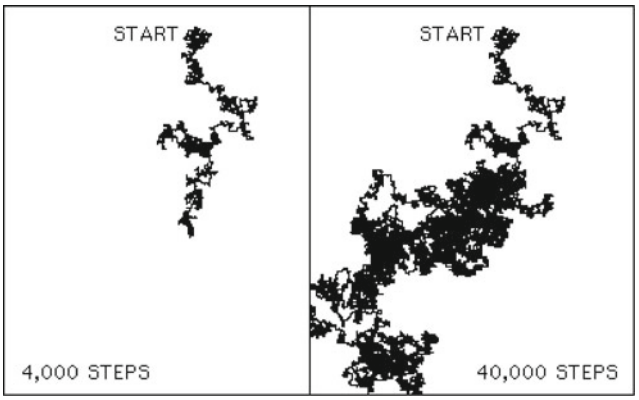


Fig. 4.24 **a** Trail of a particle for 4000 steps. **b** Trail for additional steps to total 40,000

A similar argument in three dimensions gives

$$\sigma_r^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 6Dt \quad \text{(three dimensions)}. \qquad (4.79)$$

Figure 4.24 shows the result of a computer simulation of a two-dimensional random walk. A random number is selected to determine whether to step one pixel to the left, up, right, or down—each with the same probability. The trail for 4000 steps is shown in Fig. 4.24a. The results of continuing for 40,000 steps are shown in Fig. 4.24b. Note how the particle wanders around one region of space and then takes a number of steps in the same direction to move someplace else. The particle trajectory is "thready." It does not cover space uniformly. A uniform coverage would be very nonrandom. It is only when many particles are considered that a Gaussian distribution of particle concentration results.

Both results in Fig. 4.24 were for the same sequence of random numbers. A computer simulation with 328 runs of 10,000 steps each gave $\overline{x} = -3.3$, $\sigma_x^2 = 5142$, $\overline{y} = 8.2$, $\sigma_y^2 = 4773$, and $\overline{x^2} + \overline{y^2} = 10,027$. The expected values are, respectively, 0, 5000, 0, 5000, and 10,000.

Symbols Used in Chap. 4

Symbol	Use	Units	First used page
a, a_1, a_2	Particle radius	m	90
b_1, b_2, b_3	Constants		97
f	Fraction of cell surface area		100
g	Gravitational acceleration	m s^{-2}	89
$\mathbf{g}$	Force	N	91
i	Particle current	s^{-1}	85
$j, \mathbf{j}, j_s$	Solute fluence rate	m^{-2} s^{-1}	85
$j_{\text{drift}}, j_{\text{diff}}$	Solute fluence rate due to drift velocity, diffusion	m^{-2} s^{-1}	93
$\mathbf{j}_m$	Mass fluence rate	kg m^{-2} s^{-1}	85
j_n	Component of $\mathbf{j}$ normal to a surface	m^{-2} s^{-1}	87
j_p	Momentum fluence rate	N m^{-2}	93
$\mathbf{j}_v$	Volume fluence rate	m s^{-1}	85
j_x, j_y, j_z	Components of $\mathbf{j}$	m^{-2} s^{-1}	87
k_B	Boltzmann's constant	J K^{-1}	89
l	Linear separation of pores on cell surface	m	112
m	Mass	kg	89
m	$n - n'$		106
$\hat{\mathbf{n}}$	Unit vector normal to a surface		87
n, n'	Number of steps to right, left		106
p, q	Probabilities		106
r	Distance, radius	m	87
s	Dummy variable		90
t	Time	s	85
t_c	Collision time	s	90
u	Energy of a particle	J	89
$v, \mathbf{v}$	Velocity	m s^{-1}	89
x, y, z	Cartesian coordinates	m	85
A	Constant		103
B, B'	Cell radius	m	99
C, C_s	Concentration	m^{-3}	85
D	Diffusion constant	m^2 s^{-1}	92
$F, \mathbf{F}, \mathbf{F}_{\text{ext}}$	Force	N	91
G	Correction factor for average concentration		103
K	Thermal conductivity	J K^{-1} m^{-1} s^{-1}	93
L	Length	m	97
M	Mass	kg	88
M	Molecular weight		94
N, N_0	Number of molecules		86
N	Number of pores on cell surface		100
N	Number of steps in a random walk		106
P	Rate of energy production (power)	W	87
P	Probability		89
Q	Rate of creating a substance per unit volume	m^{-3} s^{-1}	89
R	Gas constant	J K^{-1} mol^{-1}	103
R	Radius of a sphere	m	98
R_p	Radius of a pore	m	99
S	Surface area	m^2	86
$d\mathbf{S}$	Vector surface element pointing in the direction of the normal	m^2	87
T	Absolute temperature	K	89
V	Volume	m^3	88
ΔZ	Cell membrane thickness	m	99
α	Proportionality constant		91
β	Proportionality constant between force and velocity	N s m^{-1}	91
λ	Mean free path	m	90
λ	Ratio of D/v	m	103
θ, ϕ	Angles		86
η	Coefficient of viscosity	Pa s	93
σ	Standard deviation		96
σ	Electrical conductivity	Ω^{-1} m^{-1}	93
ξ	Position	m	104
ξ	Dimensionless variable		103
ρ	Mass density	kg m^{-3}	88
μ_s	Chemical potential of solute	J molecule^{-1}	92

Problems

Section 4.1

Problem 1. A cylindrical pipe with a cross-sectional area $S = 1\ \text{cm}^2$ and length 0.1 cm has $j_s(0)S = 200\ \text{s}^{-1}$ and $j_s(0.1)S = 150\ \text{s}^{-1}$.

(a) What is the total rate of buildup of particles in the pipe?
(b) What is the average rate of change of concentration in the section of pipe?

Problem 2. Write the continuity equation in cylindrical coordinates if $j_\phi = 0$ but j_r and j_z can be nonzero.

Problem 3. Consider two concentric spheres of radii r and $r+dr$. If the particle fluence rate points radially and depends only on r, and the number of particles between r and $r+dr$ is not changing, show that $d(r^2 j)/dr = 0$.

Problem 4. Integrate Eq. 4.8 over a volume and subtract the result from Eq. 4.4. The resulting relationship is called the *divergence theorem*.

Section 4.2

Problem 5. Suppose that the total blood flow through a region is F ($\text{m}^3\ \text{s}^{-1}$). A chemically inert substance is carried into the region in the blood. The total number of molecules of the substance in the region is N. The amount of blood in the region is not changing. Show that $dN/dt = (C_A - C_V)F$, where C_A and C_V are the concentrations of substance in the arterial and venous blood. This is known as the Fick principle or the Fick tracer method. It is often used with radioactive tracers.

Section 4.3

Problem 6. Allen et al. (1982) report seeing regular movement of particles in the axoplasm of a squid axon. At a temperature of 21 °C, the following mean drift speeds were observed:

Particle size (μm)	Typical speed ($\mu\text{m}\ \text{s}^{-1}$)
0.8 – 5.0	0.8
0.2 – 0.6	2

How do these values compare to thermal speeds? (Make a reasonable assumption about the density of particles and assume that they are spherical.)

Problem 7. This problem looks at the original observations of Robert Brown that established Brownian motion.

(a) Combine Eqs. 4.23 and 4.71 to determine an expression for the average distance a particle of radius a will diffuse through a fluid of viscosity η in time t.
(b) Assume you observe a pollen grain with a radius of 50 microns in water at room temperature, and that your visual perception is particularly sensitive to motions occurring over a time of about one second. What is the average distance you observe the grain to move?
(c) Now assume your eye cannot see movements that occur over angles of less than 1 min of arc, or 3×10^{-4} radians (In Chap. 14, we estimate 3 min of arc, but use 1 min here to be conservative). Most eyes cannot focus on objects closer than 25 cm. Determine the smallest displacement you can observe with the naked eye.
(d) Robert Brown had a microscope that could magnify objects by a factor of about 370. What is the smallest displacement he could observe with his microscope? Is this larger or smaller than the displacement of a pollen grain in one second?

In fact, Brown did not observe the motion of entire pollen grains. He observed fat and starch particles about 2 μm in diameter that are released by pollen. For more on Brown's original observations, see Pearle et al. (2010).

Section 4.4

Problem 8. (a) Use the ideal gas law, $pV = Nk_BT = nRT$ to compute the volume of 1 mole of gas at $T = 30\,^\circ\text{C}$ and $p = 1\,\text{atm}$. Express your answer in liters. Show that this is equivalent to a concentration of 2.4×10^{25} molecule m^{-3}.
(b) Find the concentration of liquid water molecules at room temperature.

Problem 9. Using the information on the mean free path in the atmosphere and assuming that all molecules have a molecular weight of 30, find the height at which the mean free path is 1 cm. Assume the atmosphere has a constant temperature.

Section 4.6

Problem 10. Suppose $C(x,t) = \left(N/\sqrt{4\pi Dt}\right) e^{-x^2/4Dt}$. Find an expression for $j_s(x,t)$.

Problem 11. Show that the momentum flux density, j_p, in Table 4.3 has the same units as force per unit area. Compare the equation to Eq. 1.33 and interpret η physically.

Problem 12. Jean Perrin measured the distribution of gamboge particles in water as a function of height, to determine Avagadro's number (Perrin 1910). The radius of the spherical particles was 0.212 μm, the density of water was 1 g cm^{-3}, the density of the particles was 1.207 g cm^{-3}, and the temperature was 20 °C. He counted 13,000 particles, and found their relative number, N, as a function of height, z, to be (data normalized so N is 100 at $z = 5$ μm)

z (μm)	N
5	100
35	47
65	22.6
95	12

(a) Fit these data to a Boltzmann distribution, and determine a value for Boltzmann's constant. Include the effect of buoyancy in your calculation. Fitting techniques are discussed in Chap. 11.

(b) In Perrin's time, the gas constant was known approximately: $R = 8.32$ J K^{-1} mol^{-1}. Use this value and your result from part (a) to calculate Avogadro's number.

Section 4.7

Problem 13. If all macromolecules have the same density, derive the expression for D versus the molecular weight that was used to draw the line in Fig. 4.12.

Problem 14. For diagnostic studies of the lung, it would be convenient to have radioactive particles that tag the air and that are small enough to penetrate all the way to the alveoli. It is possible to make the isotope ^{99m}Tc into a "pseudogas" by burning a flammable aerosol containing it. The resulting particles have a radius of about 60 nm (Burch et al. 1984). Estimate the mean free path for these particles. If it is small compared to the molecular diameter, then Stokes' law applies, and you can use Eq. 4.23 to obtain the diffusion constant. (The viscosity of air at body temperature is about 1.8×10^{-5} Pa s.)

Problem 15. Figure 4.12 shows that D for O_2 in water at 298 K is 1.2×10^{-9} m^2 s^{-1} and that the molecular radius of O_2 is 0.2 nm. The diffusion constant of a dilute gas (where the mean free path is larger than the molecular diameter) is $D = \lambda^2/2t_c$, where the collision time is given by Eq. 4.15.

(a) Find a numeric value for the diffusion constant for O_2 in O_2 at 1 atm and 298 K and its ratio to D for O_2 in water. The molecular weight of oxygen is 32.

(b) Assuming that this equation for a dilute gas is valid in water, estimate the mean free path of an oxygen molecule in water.

Section 4.8

Problem 16. (a) The three-dimensional normalized analog of Eq. 4.25 is

$$C(x, y, z, t) = \frac{N}{\left[2\pi\sigma^2(t)\right]^{3/2}} \exp\left(-\frac{x^2+y^2+z^2}{2\sigma^2(t)}\right).$$

Find the three-dimensional analog of Eq. 4.27.

(b) Show that $\sigma^2 = \overline{x^2} + \overline{y^2} + \overline{z^2} = 6Dt$.

Problem 17. A crude approximation to the Gaussian probability distribution is a rectangle of height P_0 and width $2L$. It gives a constant probability for a distance L either side of the mean.

(a) Determine the value of P_0 and L so that the distribution has the same value of σ as a Gaussian.

(b) Plot $P(x, t)$ if σ is given by Eq. 4.27 and the mean remains centered at the origin for times of 1, 5, 50, 100, and 500 ms. Use D for oxygen diffusing in water at body temperature.

(c) How long does it take for the oxygen to have a reasonable probability of diffusing a distance of 8 μm, the diameter of a capillary?

(d) For $t = 100$ ms, plot both the accurate Gaussian and the rectangular approximation.

Problem 18. Write an equation for Fick's second law in three-dimensional Cartesian coordinates when the diffusion constant depends on position: $D = D(x, y, z)$.

Problem 19. The heat-flow equation in one dimension is

$$j_H = -\kappa\left(\frac{\partial T}{\partial x}\right),$$

where κ is the thermal conductivity in W m^{-1} K^{-1}. One often finds an equation for the "diffusion" of energy by heat flow:

$$\frac{\partial T}{\partial t} = D_H\left(\frac{\partial^2 T}{\partial x^2}\right).$$

The units of j_H are J m^{-2} s^{-1}. The internal energy per unit volume is given by $u = \rho cT$, where c is the heat capacity per unit mass and ρ is the density of the material. Derive the second equation from the first and show how D_H depends on κ, c, and ρ.

Problem 20. The dimensionless *Lewis number* is defined as the ratio of the diffusion constant for molecules and the diffusion constant for heat flow (see Problem 19). If the Lewis number is large, molecular diffusion occurs much more rapidly than the diffusion of energy by heat flow. If the Lewis number is small, energy diffuses more rapidly than molecules. Use the following parameters:

	Air	Water
D (m^2 s^{-1})	2×10^{-5}	2×10^{-9}
κ (W m^{-1} K^{-1})	0.03	0.6
c (J kg^{-1} K^{-1})	1000	4000
ρ (kg m^{-3})	1.2	1000.

(a) Calculate the Lewis number for oxygen in air and in water.

(b) Is it possible using either air or water to design a system in which oxygen is transported by diffusion with almost no transfer of heat?

Problem 21. A sheet of labeled water molecules starts at the origin in a one-dimensional problem and diffuses in the x direction.

(a) Plot σ vs t for diffusion of water in water.
(b) Deduce a "velocity" versus time.
(c) How long does it take for the water to have a reasonable chance of traveling 1 μm? 10 μm? 100 μm? 1 mm? 1 cm? 10 cm?

Problem 22. In three dimensions the root-mean-square diffusion distance is $\sigma = \sqrt{6Dt}$, where t is the diffusion time. Consider the diffusion of oxygen from air to the blood in the lungs. The terminal air sacs in the lungs, the alveoli, have a radius of about 100 μm. The radius of a capillary is about 4 μm. Estimate the time for an oxygen molecule to diffuse from the center to the edge of an alveolus, and the time to diffuse from the edge to the center of a capillary. Which is greater? From the data in Table 1.4 estimate how long blood remains in a capillary. Is it long enough for diffusion of oxygen to occur? Assume the diffusion constant of oxygen in air is 2×10^{-5} m^2 s^{-1} and in water is 2×10^{-9} m^2 s^{-1}.

Problem 23. Why breathe? Estimate the time required for oxygen to diffuse from our nose to our lungs. Assume the diffusion constant of oxygen in air is 2×10^{-5} m^2 s^{-1}.

Problem 24. At a nerve-muscle junction, the signal from the nerve is transmitted to the muscle by a chemical junction or synapse. Molecules of acetylcholine (ACh) must diffuse from the end of the nerve cell across an extracellular gap about 20 nm wide, to the muscle cell in order to activate the muscle. Assuming one-dimensional diffusion, estimate the signal delay caused by the time needed for ACh to diffuse. The delay of the signal at the nerve-muscle junction is about 0.5 ms. How does this compare to the diffusion time? Use a diffusion constant of 5×10^{-10} m^2 s^{-1}.

Problem 25. A substance has diffusion constant D, and its concentration is distributed in space according to $C(x,t) = A(t)\sin(2\pi x/L)$, where L is the wavelength and $A(t)$ is the amplitude of the distribution. Use the one-dimensional diffusion equation, Eq. 4.26, to show that the concentration decays exponentially with time, $A(t) \propto e^{-t/\tau}$. Determine an expression for the time constant τ in terms of L and D. Which decays faster: a long-wavelength (diffuse) distribution, or a short-wavelength (localized) distribution? This result can be used with the Fourier methods developed in Chap. 11 to derive very general solutions to the diffusion equation.

Problem 26. Some tissues, such as skeletal muscle, are anisotropic: the rate of diffusion depends on direction. In these tissues, Fick's first law in two dimensions has the form

$$\begin{pmatrix} j_x \\ j_y \end{pmatrix} = -\begin{pmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{pmatrix} \begin{pmatrix} \partial C/\partial x \\ \partial C/\partial y \end{pmatrix}.$$

The 2×2 matrix is called the *diffusion tensor*. It is always symmetric, so $D_{xy} = D_{yx}$.

(a) Derive the two-dimensional diffusion equation for anisotropic tissue. Assume the diffusion tensor depends on direction but not on position.
(b) If the coordinate system is rotated from (x, y) to (x', y') by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

the diffusion tensor changes by

$$\begin{pmatrix} D_{x'x'} & D_{x'y'} \\ D_{x'y'} & D_{y'y'} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Find the angle θ such that the tensor is diagonal ($D_{x'y'} = 0$). Typically, this direction is parallel to a special direction in the tissue, such as the direction of fibers in a muscle.
(c) Show that the *trace* of the diffusion tensor (the sum of the diagonal terms) is the same in any coordinate system ($D_{xx} + D_{yy} = D_{x'x'} + D_{y'y'}$ for any θ). Basser et al. (1994) invented a way to measure the diffusion tensor using Magnetic Resonance Imaging (Chap. 18). From the diffusion tensor, they can image the direction of the fiber tracts. When they want images that are independent of the fiber direction, they use the trace.

Problem 27. Calcium ions diffuse inside cells. Their concentration is also controlled by a *buffer*:

$$\mathrm{Ca} + \mathrm{B} \iff \mathrm{CaB}.$$

The concentrations of free calcium, unbound buffer, and bound buffer ([Ca], [B], and [CaB]) are governed, assuming the buffer is immobile, by the differential equations

$$\begin{aligned} \frac{\partial[\mathrm{Ca}]}{\partial t} &= D\nabla^2[\mathrm{Ca}] - k^+[\mathrm{Ca}][\mathrm{B}] + k^-[\mathrm{CaB}], \\ \frac{\partial[\mathrm{B}]}{\partial t} &= -k^+[\mathrm{Ca}][\mathrm{B}] + k^-[\mathrm{CaB}], \\ \frac{\partial[\mathrm{CaB}]}{\partial t} &= k^+[\mathrm{Ca}][\mathrm{B}] - k^-[\mathrm{CaB}]. \end{aligned}$$

(a) What are the dimensions (units) of k^+ and k^- if the concentrations are measured in mol l^{-1} and time in s?
(b) Derive differential equations governing the total calcium and buffer concentrations, $[\mathrm{Ca}]_T = [\mathrm{Ca}] + [\mathrm{CaB}]$ and $[\mathrm{B}]_T = [\mathrm{B}] + [\mathrm{CaB}]$. Show that $[\mathrm{B}]_T$ is independent of time.
(c) Assume calcium and buffer interact so rapidly that they are always in equilibrium:

$$\frac{[\mathrm{Ca}][\mathrm{B}]}{[\mathrm{CaB}]} = K,$$

where $K = k^-/k^+$. Write $[\mathrm{Ca}]_T$ in terms of $[\mathrm{Ca}]$, $[\mathrm{B}]_T$, and K (eliminate [B] and [CaB]).

(d) Differentiate your expression in (c) with respect to time and use it in the differential equation for $[\mathrm{Ca}]_T$ found in (b). Show that [Ca] obeys a diffusion equation with an "effective" diffusion constant that depends on [Ca]:

$$D_{\mathrm{eff}} = \frac{D}{1 + \frac{K[\mathrm{B}]_T}{(K+[\mathrm{Ca}])^2}}.$$

(e) If $[\mathrm{Ca}] \ll K$ and $[\mathrm{B}]_T = 100K$ (typical for the endoplasmic reticulum), determine D_{eff}/D.

For more about diffusion with buffers, see Wagner and Keizer (1994).

Problem 28. Inside cells, calcium is stored in compartments, such as the sarcoplasmic reticulum. In some cells, a rise in calcium concentration, C, triggers the release of this stored calcium. A model of such *calcium-induced calcium release* is

$$\frac{dC}{dt} = -\frac{k}{C_0^2} C\,(4C - C_0)\,(C - C_0) \tag{1}$$

(a) Plot the rate of calcium release (the right-hand side of Eq. 1) vs C. Identify points for which the calcium release is zero (steady-state solutions to Eq. 1). By qualitative reasoning, determine which of these points are stable and which are unstable. (Will a small change in C from the steady-state value cause C to return to the steady-state value or move farther away from it?)

(b) If $C \ll C_0/4$, what does Eq. 1 become, and what is its solution?

(c) Eq. 1 is difficult to solve analytically. To find a numerical solution, approximate it as

$$\frac{C(t+\Delta t) - C(t)}{\Delta t} = -\frac{k}{C_0^2} C(t)\,[4C(t) - C_0]\,[C(t) - C_0]\,. \tag{2}$$

Write a computer program to determine $C(t)$ at times $t = n\Delta t$, $n = 1, 2, 3, \ldots, 100$, using $\Delta t = 0.1\,\mathrm{s}$, $k = 1\,\mathrm{s}^{-1}$, $C_0 = 1\ \mu\mathrm{M}$, and $C(t=0) = C'$. Find the threshold value of C', below which $C(t)$ goes to zero, and above which $C(t)$ goes to C_0.

(d) If we include diffusion of calcium in one dimension, Eq. 1 becomes

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - \frac{k}{C_0^2} C\,(4C - C_0)\,(C - C_0)\,. \tag{3}$$

This is a type of *reaction–diffusion equation*. To solve Eq. 3 numerically, divide the distance along the cell into discrete points, $x = m\Delta x$, $m = 0, 1, 2, \ldots, M$. Approximate Eq. 3 as

$$\frac{C(x, t+\Delta t) - C(x,t)}{\Delta t} \tag{4}$$

$$= D\frac{C(x+\Delta x, t) - 2C(x,t) + C(x-\Delta x, t)}{(\Delta x)^2}$$

$$- \frac{k}{C_0^2} C(x,t)\,(4C(x,t) - C_0)\,(C(x,t) - C_0)$$

Assume the ends of the cell are sealed, so $C(0,t) = C(\Delta x, t)$ at one end and $C(M\Delta x, t) = C((M-1)\Delta x, t)$ at the other. Start with the cell at $C(x,0) = 0$ for all points except at one end, where $C(0,0) = C_0$. Calculate $C(x,t)$ using $\Delta x = 5\,\mu\mathrm{m}$, $\Delta t = 0.1\,\mathrm{s}$, $D = 200\,\mu\mathrm{m}^2\,\mathrm{s}^{-1}$, and $C_0 = 1\ \mu\mathrm{M}$. You should get a wave of calcium propagating down the cell. What is its speed?

Calcium waves play an important role in many cells. This simple model does not include a mechanism to return the calcium concentration to its originally low value after the wave has passed (a process called recovery). For a more realistic model, see Tang and Othmer (1994). For more information about numerical methods, see Press et al. (1992).

Problem 29. The numerical approximation for the diffusion equation, derived as part of Problem 28, has a key limitation: it is unstable if the time step is too large. This problem can be avoided using the Crank–Nicolson method. Replace the first time derivative in the diffusion equation with a finite difference, as was done in Problem 28. Next, replace the second space derivative with the finite difference approximation from Problem 28, but instead of evaluating the second derivative at time t, use the average of the second derivative evaluated at times t and $t + \Delta t$.

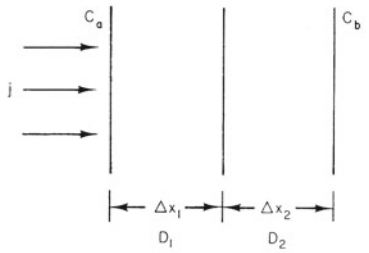
(a) Write down this numerical approximation to the diffusion equation, analogous to Eq. 4 in Problem 28.

(b) Explain why this expression is more difficult to compute than the expression given in the first two lines of Eq. 4. Hint: consider how you determine $C(t+\Delta t)$ once you know $C(t)$. The difficulty you discover in part (b) is offset by the advantage that the Crank–Nicolson method is stable for any time step. For more information about the Crank–Nicolson method, stability, and other numerical issues, see Press et al. (1992).

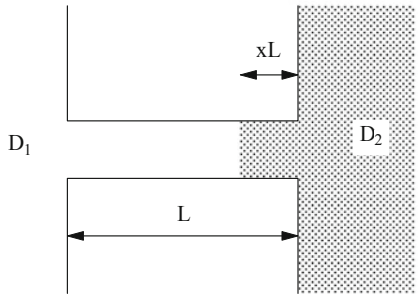
Section 4.9

Problem 30. Consider steady-state diffusion through two plane layers as shown in the figure. Show that the diffusion is the same as through a single plane layer of thickness $\Delta x_1 + \Delta x_2$, with diffusion constant

$$D = \frac{D_1 D_2}{\frac{\Delta x_1}{\Delta x_1 + \Delta x_2} D_2 + \frac{\Delta x_2}{\Delta x_1 + \Delta x_2} D_1}.$$



Problem 31. A fluid on the right of a membrane has different properties than the fluid on the left. Let the diffusion constants on left and right be D_1 and D_2, respectively, and let the pores in the membrane be filled by the fluid on the right a distance xL, where L is the thickness of the membrane.



(a) Use the results of Problem 30 to determine the effective diffusion constant D for a membrane of thickness L when $D_2 = yD_1$, $\Delta x_1 = (1-x)L$, and $\Delta x_2 = xL$. Neglect end effects.
(b) In the case that oxygen is diffusing in air and water at 310 K, the diffusion constants are $D_1 = 2.2 \times 10^{-5}$ m^2 s^{-1}, $D_2 = 1.6 \times 10^{-9}$ m^2 s^{-1}. Plot D/D_1 vs x.

Section 4.10

Problem 32.

(a) Derive Eq. 4.45.
(b) Derive Eqs. 4.51 and 4.52 from Eqs. 4.48 and 4.49.

Problem 33. We can estimate B/B' of Eqs. 4.49–4.55 by noting that B' must be larger than B because of two effects. First, it is larger by $\pi R_p/4$ because of end effects. Second, the concentration varies near the pores and smooths out further away, so B' must also be larger by an amount roughly equal to l, the spacing of the pores. There are $N/4\pi B^2$ pores per m^2, so $l \approx R_p(\pi/f)^{1/2}$. Use the example in the text: $B = 5$ μm, $\Delta Z = 5$ nm, $f = 0.001$, to estimate these two corrections. Assume that the pore radius, R_p, is smaller than ΔZ. Are these corrections important?

Problem 34. Consider an impervious plane at $z = 0$ containing a circular disk of radius a having a concentration C_0. The concentration at large z goes to zero. Carslaw and Jaeger (1959) show that the steady-state solution to the diffusion equation is

$$C(r,z) = \frac{2C_0}{\pi} \sin^{-1}\left[\frac{2a}{\sqrt{(r-a)^2+z^2}+\sqrt{(r+a)^2+z^2}}\right].$$

(a) (optional) Verify that $C(r,z)$ satisfies $\nabla^2 C = 0$. The calculation is quite involved, and you may wish to use a computer algebra program such as Mathematica or Maple.
(b) Show that for $z = 0$, $C = C_0$ if $r < a$.
(c) Show that for $z = 0$, $dC/dz = 0$ if $r > a$.
(d) Integrate j_z over the disk ($z = 0$, $0 < r < a$) and show that $i = 4DaC_0$.

Problem 35. Apply the analysis of Sect. 4.10 to determine how the current i_{vessel} depends on the fraction of surface area covered by pores, for a cylindrical vessel of radius B. Assume that the concentration reaches a value C_5 at some large finite radius R.

Section 4.11

Problem 36. The processes of heat conduction and diffusion are similar: the concentration and temperature both obey the diffusion equation (Problem 19). Consider a spherical cow of radius R having a specific metabolic rate Q W kg^{-1}. Assume the temperature of the outer surface of the cow is the same as the surroundings, T_{sur}. Assume that heat transfer within the cow is by heat conduction.

(a) Calculate the steady state temperature distribution inside the animal and find the core temperature at the center of the sphere.
(b) Consider a smaller (but still spherical) animal such as a rabbit. What is its core temperature?
(c) Calculate the temperature distribution and core temperature in a rabbit covered with fur of thickness d.

Assume the bodies of the cow and rabbit have the thermal properties of water and that the fur has the thermal properties of air. Let $d = 0.03$ m and $T_{sur} = 20\,^\circ$C.

	Water	Air
κ (W m^{-1} K^{-1})	0.6	0.03
c (J kg^{-1} K^{-1})	4000	1000
ρ (kg m^{-3})	1000	1.2
	Cow	Rabbit
R (m)	0.3	0.05
Q (W kg^{-1})	0.6	1.6

Problem 37. The goal of this problem is to estimate how large a cell living in an oxygenated medium can be before it is limited by oxygen transport. Assume the extracellular space is well stirred with uniform oxygen concentration C_0.

The cell is a sphere of radius R. Inside the cell oxygen is consumed at a rate Q molecule $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$.The diffusion constant for oxygen in the cell is D.

(a) Calculate the concentration of oxygen in the cell in the steady state.
(b) Assume that if the cell is to survive the oxygen concentration at the center of the cell cannot become negative. Use this constraint to estimate the maximum size of the cell.
(c) Calculate the maximum size of a cell for $C_0 = 8$ mol m^{-3}, $D = 2 \times 10^{-9}\ \mathrm{m}^2\ \mathrm{s}^{-1}$, $Q = 0.1$ mol $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$. (This value of Q is typical of protozoa; the value of C_0 is for air and is roughly the same as the oxygen concentration in blood.)

Problem 38. A diffusing substance is being consumed by a chemical reaction at a rate Q per unit volume per second. The reaction rate is limited by the concentration of some enzyme, so Q is independent of the concentration of the diffusing substance. For a slab of tissue of thickness b with concentration C_0 at both $x = 0$ and $x = b$, solve the equation to find $C(x)$ in the steady state. This is known as the *Warburg equation* (Warburg 1923). It is a one-dimensional model for the consumption of oxygen in tissue: points $x = 0$ and $x = b$ correspond to the walls of two capillaries side by side.

Problem 39. Suppose that a diffusing substance disappears in a chemical reaction and that the rate at which it disappears is proportional to the concentration $-kC$. Write down the Fick's second law in this case. Show what the equation becomes if one makes the substitution $C(x, y, z, t) = C'(x, y, z, t)e^{-kt}$.

Problem 40. A spherical cell has radius R. The flux density through the surface is given by $j_s = -D \operatorname{grad} C$. Suppose that the substance in question has concentration $C(t)$ inside the cell and zero outside. The material outside is removed fast enough so that the concentration remains zero. Using spherical coordinates, find a differential equation for $C(t)$ inside the cell. The thickness of the cell membrane is $\Delta r \ll R$.

Problem 41. The cornea of the eye must be transparent, so it can contain no blood vessels. (Blood absorbs light.) Oxygen needed by the cornea must diffuse from the surface into the corneal tissue. Model the cornea as a plane sheet of thickness $L = 500\,\mu\mathrm{m}$. The oxygen concentration, C, is governed by a one-dimensional steady-state diffusion equation

$$D\frac{d^2C}{dx^2} = Q.$$

Assume the cornea is consuming oxygen at a rate $Q = 4 \times 10^{22}$ molecule $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$ and has a diffusion constant $D = 3 \times 10^{-9}\,\mathrm{m}^2\,\mathrm{s}^{-1}$. The rear surface of the cornea is in contact with the aqueous humor, which has a uniform oxygen concentration $C_2 = 1.8 \times 10^{24}$ molecule m^{-3}. Consider three cases for the front surface:

(a) Solve the diffusion equation for $C(x)$ when the front surface is in contact with air, which has an oxygen concentration $C_1 = 5 \times 10^{24}\ \mathrm{m}^{-3}$.
(b) The eye is closed, but a layer of tears maintains the concentration at the front surface that is the same as the aqueous humor: $C_1 = 1.8 \times 10^{24}\ \mathrm{m}^{-3}$. Plot $C(x)$.
(c) The eye is covered by an oxygen-impermeable contact lens, so that at the front surface $dC/dx = 0$. Solve the diffusion equation and plot $C(x)$.

Supplying oxygen to the cornea is a major concern for people who wear contact lenses. Often a tear layer between the contact and cornea, replenished by blinking, is sufficient to keep the cornea oxygenated. If you sleep wearing a contact lens, this tear layer may not be replenished, and the cornea will be deprived of oxygen. For a similar but somewhat more realistic model, see Fatt and Bieber (1968).

Problem 42. The distance L that oxygen can diffuse in the steady-state is approximately $L = \sqrt{CD/Q}$,where C is the oxygen concentration, D is the diffusion constant, and Q is the rate per unit volume that oxygen is used for metabolism.

(a) Show that L has dimensions of length.
(b) The diffusion of oxygen in air is about 10,000 times larger than the diffusion of oxygen in water (Denny 1993). By how much will the diffusion distance L change if oxygen diffuses through air instead of water, all other things being equal?

Insects deliver oxygen to their flight muscles by diffusion down air-filled tubes instead of by blood vessels, thereby taking advantage of the large diffusion constant of oxygen in air (Weiss-Fogh 1964).

Section 4.12

Problem 43. Dimensionless numbers, like the Reynolds number of Chap. 1, are often useful for understanding physical phenomena. The *Péclet number* is the ratio of transport by drift to transport by diffusion. When the Péclet number is large, drift dominates. The solute fluence rate from drift is Cv, where C is the concentration and v the solvent speed. The solute fluence rate from diffusion is D times the concentration gradient (roughly C/L, where L is some characteristic distance over which the concentration varies).

(a) Determine an expression for the Péclet number in terms of C, L, v, and D.
(b) Verify that the Péclet number is dimensionless.
(c) Which parameter in Sect. 4.12 is equivalent to the Péclet number?
(d) Estimate the Péclet number for oxygen for a person walking.

(e) Estimate the Péclet number for a swimming bacterium. For more about the Péclet number, see Denny (1993) and Purcell (1977).

The Péclet number is sometimes known as the *Sherwood number*.

Problem 44. Extend Fick's second law in one dimension $\partial C/\partial t = D\,(\partial^2 C/\partial x^2)$ to include solvent drag.

Problem 45. Use Eqs. 4.63 and 4.64 to derive Eq. 4.66.

Problem 46. Expand $e^x = 1 + x + x^2/2! + x^3/3!$ to derive Eq. 4.67 from Eq. 4.66.

Problem 47. Use a Taylor's series expansion to show that $G(\xi)$in Eq. 4.69 is equal to $\xi/12$ for small ξ.

Problem 48. Consider Eq. 4.63 with $C_0 = 0$ and $C_0' = 1$.

(a) If $v > 0$, write an equation for $C(x)$. Plot $C(x)$ for $0 < x/x_1 < 1$ for two cases: $x_1 \ll \lambda$ and $x_1 \gg \lambda$. Interpret these results physically.

(b) Repeat the analysis for $v < 0$.

Section 4.14

Problem 49. We can use the microscopic model of a random walk to derive important information about diffusion without ever using the binomial probability distribution. Let $x_i(n)$ be the position of the ith particle after n steps of a random walk. Then

$$x_i(n) = x_i(n) \pm \lambda,$$

where half the time you take the $+$ sign and half the time the $-$ sign. Then $\overline{x}(n)$, the value of x averaged over N particles, is

$$\overline{x}(n) = \frac{1}{N}\sum_{i=1}^{N} x_i(n).$$

(a) Show that $\overline{x}(n) = \overline{x}(n-1)$ so that on average the particles go nowhere.

(b) Show that $\overline{x^2}(n) = \overline{x^2}(n-1) + \lambda^2$. Use this result to show that $\overline{x^2}(n) = n\lambda^2$.

For a detailed discussion of this approach, see Denny (1993).

Problem 50. We can write the diffusion constant, D, and the thermal speed, v_{rms}, in terms of the step size, λ, and the collision time, t_c, as

$$D = \frac{\lambda^2}{2t_c},$$

$$v_{\text{rms}} = \frac{\lambda}{t_c}.$$

Solve for λ and t_c in terms of D and v_{rms}.

Problem 51. Using the definitions in Prob. 50, write the diffusion constant in terms of λ and v_{rms}. By how much do you expect the diffusion constant for *heavy water* (water in which the two hydrogen atoms are deuterium, ^{2}H) to differ from the diffusion constant for water? Assume the mean free path is independent of mass.

Problem 52. Write a computer program to model a two-dimensional random walk. Make several repetitions of a random walk of 3600 steps and plot histograms of the displacements in the x and y directions and mean square displacement.

Problem 53. Write a program to display the motion of 100 particles in two dimensions.

Problem 54. Particles are released from a point between two perfectly absorbing plates located at $x = 0$ and $x = 1$. The particles random walk in one dimension until they strike a plate. Find the probability of being captured by the right-hand plate as a function of the position of release, x. (Hint: The probability is related to the diffusive fluence rate to the right-hand plate if the concentration is C_0 at x and is 0 at $x = 0$ and $x = 1$.)

Problem 55. The text considered a one-dimensional random-walk problem. Suppose that in two dimensions the walk can occur with equal probability along $+x$, $+y$, $-x$, or $-y$. The total number of steps is $N = N_x + N_y$, where the number of steps along each axis is not always equal to $N/2$.

(a) What is the probability that N_x of the N steps are parallel to the x axis?

(b) What is the probability that the net displacement along the x axis is $m_x\lambda$?

(c) Show that the probability of a particle being at $(m_x\lambda, m_y\lambda)$ after N steps is

$$P'(m_x, m_y) = \sum_{N_x}\left(\frac{N!}{N_x!\,(N-N_x)!}\right)\left(\frac{1}{2}\right)^N P(m_x, N_x)\,P(m_y, N-N_x),$$

where $P(m, N)$ on the right-hand side of this equation is given by Eq. 4.76.

(d) The factor $N!/N_x!(N-N_x)!$ is proportional to a binomial probability. What probability? Where does this factor peak when N is large?

(e) Using the above result, show that $P'(m_x, m_y) = P(m_x, N/2)\,P(m_y, N/2)$.

(f) Write a Gaussian approximation for two-dimensional diffusion.

References

Allen RD, Metuzals J, Tasaki L, Bradt ST, Gilbert SP (1982) Fast axonal transport in squid giant axon. Science 218:1127–1129

Axelrod D, Wang MD (1994) Reduction-of-dimensionality kinetics at reaction-limited cell surface receptors. Biophys J 66(3, Pt. 1):588–600

Barr G (1931) A monograph of viscometry. Oxford University Press, London

Basser PJ, Mattiello J, LeBihan D (1994) MR diffusion tensor spectroscopy and imaging. Biophys J 66:259–267

Bean CP (1972) The physics of neutral membranes—neutral pores. In: Eisenman G (ed) Membranes—a series of advances, vol 1. Dekker, New York, pp. 1–55

Benedek GB, Villars FMH (2000) Physics with illustrative examples from medicine and biology, vol 2. Statistical physics, 2nd edn. Springer, New York

Berg HC (1975) Chemotaxis in bacteria. Ann Rev Biophys Bioeng 4:119–136

Berg HC (1983) Random walks in biology. Princeton University Press, Princeton

Berg HC (2003) *E. coli* in motion. Springer, New York

Berg HC, Purcell EM (1977) Physics of chemoreception. Biophys J 20:193–219

Bird RB, Stewart WE, Lightfoot EN (1960) Transport phenomena. Wiley, New York

Burch WM, Tetley IJ, Gras JL (1984) Technetium 99m "peudogas" for diagnostic studies in the lung. Clin Phys Physiol Meas 5:79–85

Crank J (1975) The mathematics of diffusion, 2nd edn. Oxford University Press, New York

Carslaw HS, Jaeger JC (1959) Conduction of heat in solids. Oxford University Press, New York

Denny MW (1993) Air and water: the biology and physics of life's media. Princeton University Press, Princeton

Fatt I, Bieber MT (1968) The steady-state distribution of oxygen and carbon dioxide in the in vivo cornea. Exp Eye Res 7:103–112

Fletcher DA Theriot JA (2004) An introduction to cell motility for a physical scientist. Phys Biol 1:T1–T10

Hildebrand JH, Prausnitz JM, Scott RL (1970) Regular and related solutions: the solubility of gases, liquids, and solids. Van Nostrand Reinhold, New York

Hoffmann PM (2012) Life's ratchet: how molecular machines extract order from chaos. Basic, New York

Levitt D (1975) General continuum analysis of transport through pores. I. Proof of Onsager's reciprocity postulate for uniform pore. Biophys J 15:533–551

Meidner H, Mansfield TA (1968) Physiology of stomata. McGraw-Hill, New York

Maddock JR, Shapiro L (1993) Polar location of the chemoreceptor complex in the *Escherichia coli* cell. Science 259:1717–1723

Paine PL, Scherr P (1975) Drag coefficients for the movement of rigid spheres through liquid-filled cylinders. Biophys J 15:1087–1091

Parkinson JS, Blair DF (1993) Does *E. coli* have a nose? Science 259:1701–1702

Pearle P, Collett B, Bart K, Bilderback D, Newman D, Samuels S (2010) What Brown saw and you can too. Amer J Phys 78:1278–1289

Perrin J (1910) Brownian movement and molecular reality. Taylor and Francis, London

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, New York (reprinted with corrections, 1995)

Pryde JA (1966) The liquid state. Hutchinson University Library, London

Purcell EM (1977) Life at low Reynolds number. Am J Phys 45:3–11

Reif F (1965) Fundamentals of statistical and thermal physics. McGraw-Hill, New York

Schey HM (2004) Div, grad, curl, and all that: an informal text on vector calculus, 4th edn. Norton, New York

Tang Y-h, Othmer JG (1994) A model of calcium dynamics in cardiac myocytes based on the kinetics of ryanodine-sensitive calcium channels. Biophys J 67:2223–2235

Wagner J, Keizer J (1994) Effects of rapid buffers on Ca^{++} diffusion and Ca^{++}oscillations. Biophys J 67:447–456

Wang D, Gou S-y, Axelrod D (1992) Reaction rate enhancement by surface diffusion of adsorbates. Biophys Chem 43(2):117–137

Warburg O (1923) Versuche an überlebendem Carcinomgewebe. Biochem Z 142:317–350

Weast RC (1972) Handbook of chemistry and physics, 53rd edn. Chemical Rubber Company, Cleveland

Weiss-Fogh T (1964) Diffusion in insect wing muscle, the most active tissue known. J Exp Biol 41:229–256

Zwanzig R, Szabo A (1991) Time dependent rate of diffusion-influenced ligand binding to receptors on cell surfaces. Biophys J 60(3):671–678

Transport Through Neutral Membranes

5

The last chapter discussed some of the general features of solute movement in an infinite medium. Solute particles can be carried along with the flowing solution or they can diffuse. This chapter considers the movement of solute and solvent through membranes, ignoring any electrical forces on the particles.

The movement of electrically neutral particles through aqueous pores in membranes has many applications in physiology. They range from the flow of nutrients through capillary walls, to the regulation of the amount of fluid in the interstitial space between cells, to the initial stages of the operation of the kidney.

Sections 5.1–5.4 are a qualitative introduction to the flow of water through membranes as a result of hydrostatic pressure differences or osmotic pressure differences. The reader who is not interested in the more advanced material can read just this part of the chapter, culminating in the clinical examples of Sect. 5.4.

Sections 5.5 and 5.6 present phenomenological transport equations that are simple linear relationships between the flow of water and solute particles and the pressure and concentration differences that cause the flows. These relationships are valid for any type of membrane as long as a linear relationship adequately describes the flow and the proportionality constants are regarded as experimentally determined quantities. These equations are applied to the artificial kidney in Sect. 5.7.

Section 5.8 presents a simple model for countercurrent transport, which is important in artificial organs, the kidney, and in conserving heat loss from the extremities.

The last section, Sect. 5.9, provides a more advanced treatment of one particular membrane model: a membrane pierced by pores in which electrical forces can be neglected and in which Poiseuille flow takes place. The model leads to expressions for the phenomenological coefficients that can be compared to experimental data, though that is not done here. The last part of the section uses this model to calculate the forces on a membrane when there are osmotic effects.

5.1 Membranes

All cells are surrounded by a membrane 7–10 nm thick. Furthermore, virtually all the physical substructures within the cell are also bounded by membranes. Membranes separate two regions of space; they allow some substances to pass through but not others. The membrane is said to be *permeable* to a substance that can pass through it; it is *semipermeable* when only certain substances can get through. A substance that can pass through is said to be *permeant*.

Simple models for a semipermeable membrane are shown in Fig. 5.1. Figure 5.1a shows a pore that pierces the membrane. A narrower pore, in which the transported particles move single-file, is shown in Fig. 5.1b. Another simple model is shown in Fig. 5.1c: there are no pores, but water and small solute molecules actually "dissolve" in the membrane and diffuse through. The examples in Fig. 5.1 shows water molecules (open circles), solute molecules (small solid circles), and a large protein molecule that cannot pass through the membrane.

In Fig. 5.1a and b the motion of the water molecules is quite different from that of the small solute molecules. Each water molecule is in contact with neighboring water molecules so that when the water molecules move, they flow together. The result is the familiar bulk flow that occurs in a pipe. The solute molecules, on the other hand, are so dilute that they seldom collide with one another. Each solute molecule's motion is *independent of other solute molecules*.

The motion of each solute molecule is *not* independent of the motion of the surrounding water molecules. If the water is at rest, the movement of the solute molecules is diffusion; if the water is moving, this diffusion is superimposed on a flow of the solute molecules with the moving fluid (solvent drag).

In Fig. 5.1c, both the water and solute molecules dissolve into the bilayer lipid membrane. They are very dilute within the membrane, so that both kinds of molecules diffuse. The water molecules are not in contact with each other, but are

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_5, © Springer International Publishing Switzerland 2015

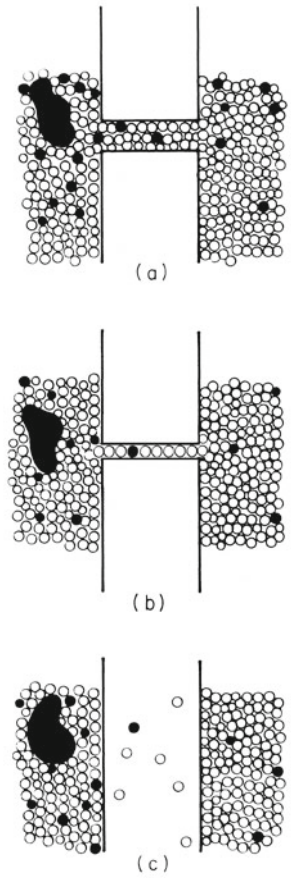


Fig. 5.1 Simple models for a semipermeable membrane. **a** A "large" pore. **b** A single-file pore such as an aquaporin channel. **c** Small molecules dissolve in the membrane and diffuse through

in some sort of interstices within the membrane structure, walking randomly in response to thermal agitation of the membrane.

It has long been known that the rate of water transport through cell membranes was too large to be explained by diffusion as in Fig. 5.1c, although such diffusion does take place. The pores that allow transport are more like those shown in Fig. 5.1a and b. Pores like the one in Fig. 5.1b, called *aquaporins*, were first discovered by Peter Agre in 1993 (Parisi et al. 2007).[1] Other mechanisms for water flow are associated with ion transport and are not discussed here (Zeuthen 2010).

[1] Some aquaporins are permeable only to water, and not to any other small molecules or ions, even hydrogen ions (Preston et al. 1992). Aquaporins are formed by proteins that span the cell membrane. Their structure has been determined by x-ray crystallography (Murata et al. 2000). Their selectivity arises from a narrowing of the channel to about 0.3 nm, about the size of a single water molecule. Aquaporins allow water to cross cell membranes at a much higher rate than it could diffuse through. Genetically defective aquaporins may be responsible for some clinical diseases, such as nephrogenic diabetes insipidus and congenital cataracts (Agre et al. 2002).

5.2 Osmotic Pressure in an Ideal Gas

The selective permeability of a membrane gives rise to some striking effects. The flow of water that occurs because solutes are present that cannot get through the membrane is called *osmosis*. This phenomenon seems strange when it is first encountered, and explanations are often fraught with misconceptions (Kramer and Myers 2012). Osmosis is important in a variety of clinical problems that are described in Sect. 5.4. We begin by finding the conditions under which no flow takes place and the direction of flow when it does occur. Later, in Sect. 5.5, we consider the rate of flow in response to a given pressure difference.

It is easiest to understand osmotic pressure by considering the special case of two ideal gases and a membrane that is permeable to one but not the other. This case is simple because the gas molecules do not interact with one another. Then, in Sect. 5.3, we will examine the phenomenon when the substances are liquids.

Suppose a box with total volume V^* contains N_1^* molecules of gas species 1. If the box is at temperature T, the ideal-gas law relates the pressure, temperature, and the number of molecules:

$$p_1 V^* = N_1^* k_B T. \tag{5.1}$$

This has been written the way physicists like to write it, in terms of the number of molecules N_1^*. Chemists write it in terms of the number of moles n_1^*:

$$p_1 V^* = n_1^* R T.$$

The only difference is that the gas constant R is per mole while the Boltzmann constant k_B is per molecule. Since 1 mole contains N_A molecules, where N_A is Avogadro's number, $N_1^* = N_A n_1^*$ and $R = N_A k_B$. Numerical values are

$$N_A = 6.022 \times 10^{23}\ \text{mol}^{-1},$$
$$k_B = 1.3806 \times 10^{-23}\ \text{J K}^{-1},$$
$$R = 8.3145\ \text{J mol}^{-1}\ \text{K}^{-1},$$
$$R = 0.08206\ \text{atm l mol}^{-1}\ \text{K}^{-1}.$$

The concentration is the number of molecules or moles per unit volume. We denote *molecular* concentration by capital letter C and *molar* concentration by lowercase c:

$$C_1 = \frac{N_1^*}{V^*}\ \text{m}^{-3} \text{ or molecules m}^{-3},$$
$$c_1 = \frac{n_1^*}{V^*}\ \text{m}^{-3} \text{ or mol m}^{-3}.$$

Chemists often express concentrations in moles per liter.

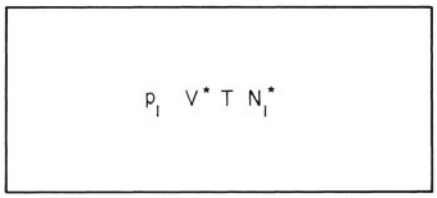


Fig. 5.2 An ideal gas fills a box of volume V^*

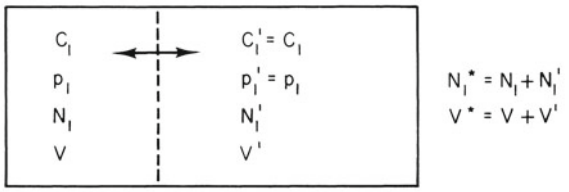


Fig. 5.3 The introduction of a semipermeable membrane does not change the pressure or concentration of the gas

If we were to imagine volume V^* divided into two subvolumes of volume V and V', the average concentration of molecules in each subvolume would remain unchanged. The pressure in each subvolume would still be p_1, and the temperature would be T. We can write

$$p_1 V = N_1 k_B T, \quad p_1 V' = N_1' k_B T.$$

Dividing both sides of each equation by the appropriate volume gives

$$p_1 = C_1 k_B T, \qquad p_1 = p_1' = C_1' k_B T. \tag{5.2}$$

Now place a membrane along the surface separating the subvolumes. The membrane has small holes so that the molecules can pass through, as shown in Fig. 5.3. This does nothing to change the fact that at equilibrium $p_1 = p_1'$. When the pressures are the same on both sides of the membrane, no molecules pass through on average. If the pressure is greater on one side than the other, molecules pass through to bring the pressures into equilibrium, as we saw in Chap. 3. Equations 5.2 say nothing about how frequently a molecule that strikes the membrane passes through. It could take hours or days for equilibrium to be attained if we started away from equilibrium and the molecules do not pass through very often.

Next, keeping V fixed, introduce species 2 on the left as in Fig. 5.4. Suppose that species 2 cannot pass through the membrane. Bombardment of the membrane by the new molecules causes an additional force on the left side of the membrane. The total pressure in volume V is now the sum of the partial pressures p_1 due to species 1 and p_2 due to the second species:

$$p = p_1 + p_2,$$

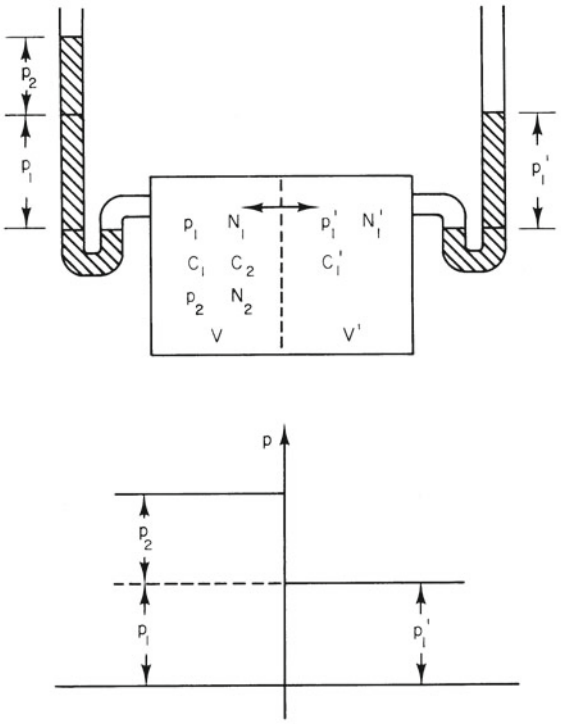


Fig. 5.4 Species 2, which cannot pass through the membrane, has been introduced in V. The pressure in V is higher than in V' by the partial pressure p_2

$$p_1 V = N_1 k_B T, \tag{5.3}$$
$$p_2 V = N_2 k_B T.$$

The ideal-gas law is still obeyed in terms of the total number of molecules in V, $N = N_1 + N_2$: $pV = p_1 V + p_2 V = N_1 k_B T + N_2 k_B T = (N_1 + N_2) k_B T = N k_B T$.

In an ideal gas the presence of the second species does not change the partial pressure p_1. The total pressure on the walls and the membrane is increased by p_2 so the membrane is bowed towards the right, but the total pressure is simply the sum of the two partial pressures. The ratio p_1/p is the fraction of the pressure due to collisions of molecules of the first kind with the membrane.

Suppose now that the pressure in V' is raised, either by compressing the gas or by introducing more molecules of type 1, so that instead of $p_1' = p_1$, we have $p_1' = p$. The partial pressure of species 1 is higher in V' than in V. Since these molecules can pass through the membrane, they will flow from V' to V. An identical flow could have been caused without having species 2, simply by raising the pressure in V'. Not every molecule striking the membrane will pass through, but some fraction of all collisions with the membrane will result in a molecule passing through. The fraction will depend on the details of the membrane structure. The number going through will be proportional to the number

of collisions on one side minus the number of collisions on the other and hence to the difference of partial pressures. If $p_1 > p_1'$, species 1 will flow from V to V'. If $p_1 < p_1'$, the flow will be in the other direction. The details of the membrane will determine how rapid this flow is. *The movement of any species of gas molecule that can pass through the membrane will be from the region of higher partial pressure to lower partial pressure.*

Suppose we start out with only species 1 on each side of the membrane and equal pressures on both sides so that $p = p_1 = p' = p_1'$. There are three ways to make p_1 less than p_1', thereby causing movement from right to left. One is simply to let the gas on the left expand into a larger volume, which lowers $p = p_1$. (Or we could have compressed the gas on the right, raising $p' = p_1'$.) The other two ways involve introducing on the left a species 2 that cannot pass through the membrane. The second way would be to keep the total pressure and volume on the left the same, but remove one molecule of species 1 for every molecule of species 2 that is introduced. The third way would be to increase the volume on the left as each molecule of species 2 is introduced, so that $p = p_1 + p_2$ remains the same.

The total partial pressure of all species that cannot pass through the membrane is called the *osmotic pressure* in region V and is usually denoted by π. If the subscript 2 denotes all impermeant species,

$$\pi_2 = C_2 k_B T. \tag{5.4}$$

The flow through the membrane because of an increase in the osmotic pressure or a decrease in the total pressure is identical. In each case the flow is determined by the difference across the membrane of p_1, the total partial pressure of all the species that can pass through.

The description in the previous paragraphs of partial pressure is easy to visualize, and for the case given it is correct. It is more general, however, to express the condition for equilibrium in terms of the chemical potential, μ. Recall that in Chap. 3 we derived the pressure in terms of volume changes of a system and the chemical potential in terms of the number of particles in the system. Suppose that the membrane separating the two sides is actually a semipermeable piston that is free to move. Equality of the total pressure on both sides of the piston means that the piston will not move and the two systems will not exchange volume. Equality of the chemical potential of a species that can get through the membrane means that the two systems will not exchange particles. It is better, therefore, to say *the flow of any species that can pass through the membrane will be from the region of higher chemical potential to the region of lower chemical potential for that species. If the chemical potentials are the same, there will be no flow.*

The mixture of two ideal gases is a special case of the ideal solution that was described in Sect. 3.18. The chemical potential of species 1 that can pass through the membrane is given by Eq. 3.77:

$$\begin{aligned}
\Delta\mu_1 &= \overline{V}_1(\Delta p - k_B T \Delta C_2), \\
\mu_1 - \mu_1' &= \overline{V}_1\left[p - p' - k_B T(C_2 - 0)\right], \\
\mu_1 - \mu_1' &= \overline{V}_1(p_1 + p_2 - p_1' - k_B T C_2).
\end{aligned}$$

Since $p_2 = k_B T C_2$, the chemical potential is the same on both sides of the membrane when $p_1 = p_1'$.

5.3 Osmotic Pressure in a Liquid

Imagine now that the two volumes are filled with a solvent, such as water. If the pressure of the water is the same in both regions there is no movement of water through the membrane, nor is there exchange of volume if the membrane piston is free to move. Increasing the pressure on one side of the fixed membrane causes water to move through the membrane from the side with higher pressure to the side with lower pressure. There is no flow when $\Delta p = 0$. If there is a solute in the water that can pass freely through the membrane along with the water, the situation is unchanged.

Now let us add some solute on the left that cannot pass through the membrane. We will keep the volume on the left fixed. To add the solute in such a way that the pressure does not change, we must remove some water molecules as we add it.

We saw in Chap. 3 that replacing some water molecules with solute increases the entropy of the solution.[2] This means that the Gibbs free energy and the chemical potential are decreased. Water moves from the region on the right, where the chemical potential is higher, to the region on the left, where it is lower. The chemical potential of the water on the left can be increased by increasing the total pressure on the left.

The chemical potential contains terms proportional to the pressure and the concentration of the impermeant solute. It

[2] This, recall, is because the water molecules are indistinguishable. A simple model (Fig. 3.15) shows why this happens. Suppose that three water molecules occupy three identical energy levels, and that these are the only three levels available. Because the molecules are indistinguishable, there is only one microstate and the entropy is zero. If one molecule of water is replaced by one solute molecule, there are then three separate microstates, corresponding to the solute molecule being in any one of the three. The entropy is $k_B \ln(3)$.

was shown in Sect. 3.18 that for an ideal solution[3]

$$\Delta\mu_w = \frac{\Delta p - k_B T \Delta C_s}{C_w}.$$

The osmotic pressure is the excess pressure that we must apply on the left to prevent the movement of water through the membrane. There is no movement of water when $p = p' + \pi$. It is more convenient to write all the unprimed quantities on the left: $p - \pi = p'$. The quantity $p - \pi$ will occur so often in what follows that it is worth a special name. We will define the *driving pressure*

$$p_d \equiv p - \pi. \tag{5.5}$$

As far as we know, it has not been used by other authors. It is a monotonic function of the chemical potential. In an ideal solution it is $C_w\mu_w$. Except in an ideal gas, it is *not* the same as the partial pressure (a concept that is not normally used in a liquid). On the right there is no solute and $p'_d = p'$. *There is no movement when the driving pressure is the same on both sides*,

$$p_d = p'_d, \tag{5.6a}$$

or the chemical potential of the water is the same on both sides,

$$\mu_w = \mu'_w. \tag{5.6b}$$

The water passes through the membrane in the direction from higher p_d to lower p_d (or from higher chemical potential to lower chemical potential). Either the total pressure or the osmotic pressure can be manipulated to change p_d (and μ_w). An increase of total pressure has the same effect as a decrease of osmotic pressure.

Increasing the concentration of the solute increases the osmotic pressure. The fact that $p_d = p - \pi = C_w\mu_w$ means that for ideal solutions obeying Eq. 3.77,

$$\pi = Ck_BT = cRT. \tag{5.7}$$

In many cases this is confirmed by experiment, particularly in dilute solutions. This is known as the *van't Hoff* law for osmotic pressure.

An *osmole* is the equivalent of a mole of solute particles. The term *osmolality* is used to refer to the number of osmoles per kilogram of *solvent*, while *osmolarity* refers to the number of osmoles per liter of *solution*. The reason for introducing the osmole is that not all impermeant solutes are ideal; their osmotic effects are slightly less than Ck_BT. The osmole takes this correction into account.

[3] An ideal solution can be defined in several equivalent ways. One is that it is a solution that obeys Eq. 3.77. Another is that when the separated components are mixed, there is no change of total volume and no heat is evolved or absorbed. See Hildebrand and Scott (1964, Chap. 2).

5.4 Some Clinical Examples

As blood flows through capillaries, oxygen and nutrients leave the blood and go to the cells. Waste products leave the cells and enter the blood. Diffusion is the main process that accomplishes this transfer. The capillaries are about the diameter of a red cell; the red cells therefore squeeze through the capillary in single file. They move in plasma, which consists of water, electrolytes, small molecules such as glucose and dissolved oxygen or carbon dioxide, and large protein molecules. All but the large protein molecules can pass through the capillary wall.

Outside the capillaries is the interstitial fluid, which bathes the cells. The concentration of protein molecules in the interstitial fluid is much less than it is in the capillaries. Osmosis is an important factor determining the pressure in the interstitial fluid and therefore its volume. The following values (in units[4] of torr) are typical for the osmotic pressure inside and outside the capillary:[5]

Inside capillary	$\pi_i = 28$ torr
Outside capillary, interstitial fluid	$\pi_o = 5$ torr

Measurements of the total pressure in the interstitial fluid are difficult, but the value seems to be about -6 torr. It is maintained below atmospheric pressure (taken here to be 0 torr) by the rigidity of the tissues. The driving pressure of water and small molecules outside is therefore

$$p_{do} = p_o - \pi_o = -6 - 5 = -11 \text{ torr}.$$

The total pressure within the capillary drops from the arterial end to the venous end, causing blood to flow along the capillary. A typical value at the arterial end is 25 torr; at the venous end, it is 10 torr. If the drop is linear along the capillary, the total pressures versus position is as plotted in Fig. 5.5a.[6] Subtracting from this the osmotic pressure of the large molecules gives the curve for the driving pressure inside, p_{di}, which is also plotted in Fig. 5.5a. Figure 5.5b shows the total and driving pressures in the interstitial fluid. Figure 5.5c compares the driving pressure inside and outside. The driving pressure is larger inside in the first half of the capillary and larger outside in the second half of the capillary. The result is an outward flow of plasma through the capillary wall in the first half and an inward flow in the second half. There is a very slight excess of outward flow. This fluid returns to the circulation via the lymphatic system.

[4] 1 torr = 1 mmHg = 133.3 Pa = 0.019 34 lb in.$^{-2}$.

[5] A short account of the pressures used here is found in Hall (2011, Chap. 16). A more detailed discussion is in Guyton et al. (1975).

[6] This simple discussion uses pressures that compensate for the fact that the surface area of the capillary is larger at the venous end than at the arterial end.

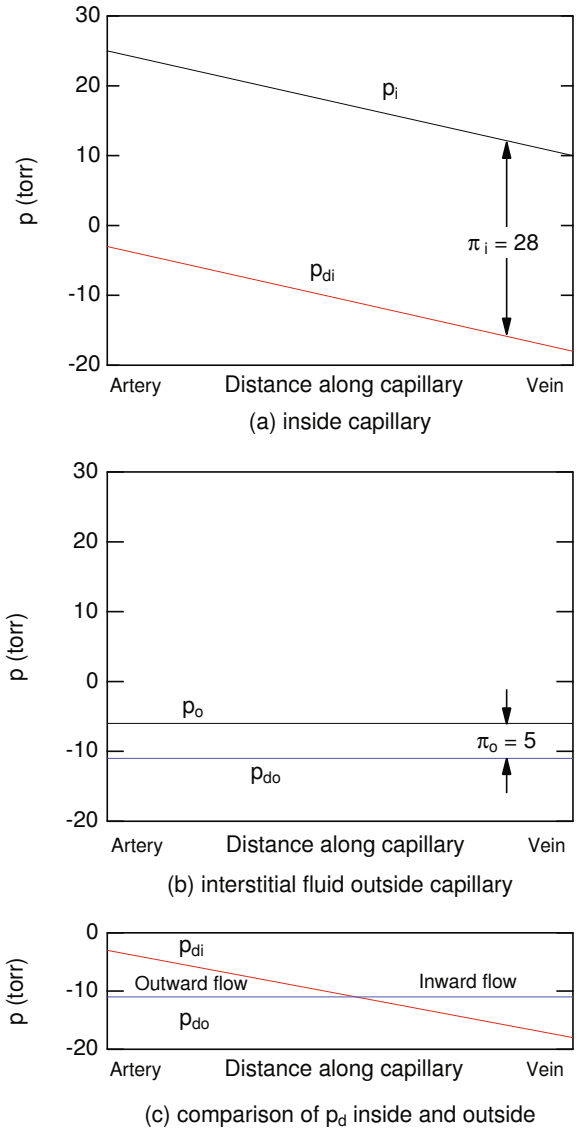


Fig. 5.5 Pressures inside and outside the capillary. **a** Inside. **b** Outside. **c** Comparison of the water driving pressure inside and outside

There are three ways that the balance of Fig. 5.5 can be disturbed, each of which can give rise to *edema*, a collection of fluid in the tissue. The first is a higher average pressure along the capillary. The second is a reduction in osmotic pressure because of a lower protein concentration in the blood (hypoproteinemia). The third is an increased permeability of the capillary wall to large molecules, which effectively reduces the osmotic pressure. Each is discussed below.

5.4.1 Edema Due to Heart Failure

A patient in right heart failure exhibits an abnormal collection of interstitial fluid in the lower part of the body (the legs for a walking patient; the back and buttocks for a patient in bed). This can be understood in terms of the mechanism discussed above. The right heart pumps blood from the veins through the lungs. If it can no longer handle this load, the venous blood is not removed rapidly enough, and the pressure in the veins and the venous end of the capillaries rises. There is a corresponding rise in p_d along the capillary. More fluid flows from the capillary to the interstitial space. The interstitial pressure rises until the net flow is again zero.

The same process can occur in left heart failure in which the pressure in the pulmonary veins builds up. The patient then has pulmonary edema and may literally drown.

5.4.2 Nephrotic Syndrome, Liver Disease, and Ascites

Patients can develop an abnormally low amount of protein in the blood serum, *hypoproteinemia*, which reduces the osmotic pressure of the blood. This can happen, for example, in *nephrotic syndrome*. The nephrons (the basic functioning units in the kidney) become permeable to protein, which is then lost in the urine. The lowering of the osmotic pressure in the blood means that the p_d rises. Therefore, there is a net movement of water into the interstitial fluid. Edema can result from hypoproteinemia from other causes, such as liver disease and malnutrition.

A patient with liver disease may suffer a collection of fluid in the abdomen. The veins of the abdomen flow through the liver before returning to the heart. This allows nutrients absorbed from the gut to be processed immediately and efficiently by the liver. Liver disease may not only decrease the plasma protein concentration, but the vessels going through the liver may become blocked, thereby raising the capillary pressure throughout the abdomen and especially in the liver. A migration of fluid out of the capillaries results. The surface of the liver "weeps" fluid into the abdomen. The excess abdominal fluid is called *ascites*.

5.4.3 Edema of Inflammatory Reaction

Whenever tissue is injured, whether it is a burn, an infection, an insect bite, or a laceration, a common sequence of events initially occurs that cause edema. They include the following:

1. *Vasodilation*. Capillaries and small blood vessels dilate, and the rate of blood flow is increased. This is responsible for the redness and warmth associated with the inflammatory process.
2. *Fluid exudation*. Plasma, including plasma proteins, leaks from the capillaries because of increased permeability of the capillary wall.

3. *Cellular migration*. The capillary walls become porous enough so that white blood cells of the immune system move out of the capillaries at the site of injury.

5.4.4 Headaches in Renal Dialysis

Dialysis is used to remove urea from the plasma of patients whose kidneys do not function. Urea is in the interstitial brain fluid and the cerebrospinal fluid in the same concentration as in the plasma; however, the permeability of the capillary–brain membrane is low, so equilibration takes several hours (Patton et al. 1989, Chap. 64). Water, oxygen, and nutrients cross from the capillary to the brain at a much faster rate than urea. As the plasma urea concentration drops, there is a temporary osmotic pressure difference resulting from the urea within the brain. The driving pressure of water is higher in the plasma, and water flows to the brain interstitial fluid. Cerebral edema results, which can cause severe headaches.

The converse of this effect is to inject into the blood urea or mannitol, another molecule that does not readily cross the blood–brain barrier. This lowers the driving pressure of water within the blood, and water flows from the brain into the blood. Although the effects do not last long, this technique is sometimes used as an emergency treatment for cerebral edema (Fishman 1975; White and Likavec 1992).

5.4.5 Osmotic Diuresis

The functional unit of the kidney is the *nephron*. Water and many solutes pass into the nephron from the blood at the glomerulus. As the urine flows through the rest of the nephron, a series of complicated processes cause a net reabsorption of most of the water and varying amounts of the solutes. Some medium-weight molecules such as mannitol are not reabsorbed at all. If they are present in the nephron—for example, from intravenous administration—the driving pressure of water is lowered and less water is reabsorbed than would be normally. The result is an increase in urine volume and a dehydration of the patient called osmotic diuresis (Gennari and Kassirer 1974; Hall 2011). Similar diuretic action takes place in a diabetic patient who "spills" glucose into the urine.

5.4.6 Osmotic Fragility of Red Cells

Red blood cells (erythrocytes) are normally disk-shaped, with the center thinner than the rim. In the disease called *hereditary spherocytosis* the red cells are more rounded. If a red cell is placed in a solution that has a higher driving pressure than that inside the cell, water moves in and the cell swells until it bursts. Since cell membranes (as distinct from the lining of capillaries) are nearly impermeable to sodium, sodium is osmotically active for this purpose. The osmotic fragility test consists of placing red cells in solutions with different sodium concentrations and determining what fraction of the cells burst. A patient with hereditary spherocytosis has cells that will be destroyed at a lower external p_d (higher sodium concentration) than normal, because the membrane is more permeable to the sodium.

5.5 Volume Transport Through a Membrane

In this section and the next we develop phenomenological equations to describe flow of fluid and flow of solute through a membrane. These are linear approximations to the dependence of the flows on pressure and solute concentration differences. Three parameters are introduced that are widely used in physiology: the filtration coefficient (or hydraulic permeability), the solute permeability, and the solute reflection coefficient.

The volume fluence rate or volume flow per unit area per second through a membrane is J_v.

$$J_v = \frac{\left(\begin{array}{c}\text{total volume per second}\\ \text{through membrane area } S\end{array}\right)}{S} = \frac{i_v}{S}\ \text{m s}^{-1}. \quad (5.8)$$

Consider pure water. The fluence rate depends on the pressure difference across the membrane. When the pressure difference is zero there is no flow. The direction of flow, and therefore the sign of the fluence rate, depends on which side of the membrane has the higher pressure. The simplest relationship that has this property is a linear one:[7]

$$J_v = L_p \Delta p. \quad (5.9)$$

The proportionality constant is called the *filtration coefficient* or *hydraulic permeability*. It depends on the details of the membrane structure, such as the properties of the pores. The SI units for L_p are m s^{-1} Pa^{-1}, m^3 N^{-1} s^{-1}, or m^2 s kg^{-1}. Often in the literature, however, values of L_p are reported in units of cm s^{-1} atm^{-1}. Since 1 atm $= 1.01 \times 10^5$ Pa, the conversion is

$$1\ \text{cm s}^{-1}\ \text{atm}^{-1} = 0.99 \times 10^{-7}\ \text{m s}^{-1}\ \text{Pa}^{-1}. \quad (5.10)$$

[7] The traditional sign convention has been followed here. There would be a minus sign in the equation if Δp were defined to be $p(x + \Delta x) - p(x)$. However, it is usually defined as $p - p'$. The flow is from the region of higher pressure to the region of lower pressure.

If a solute is present to which the membrane is completely impermeable, only water will flow, and the flow will depend on Δp_d:

$$\begin{aligned}\Delta p_d &= p_d - p_d' = p - \pi - (p' - \pi') \\ &= p - p' - (\pi - \pi') \\ &= \Delta p - \Delta\pi\end{aligned}$$

so

$$J_v = L_p(\Delta p - \Delta\pi). \tag{5.11}$$

Figure 5.6 shows the pressure relations on each side of the membrane for no flow and for flow in either direction.

It is important to note that the quantity in parentheses is a property of the solutions on either side of the membrane. The permeability depends on the transport mechanism.

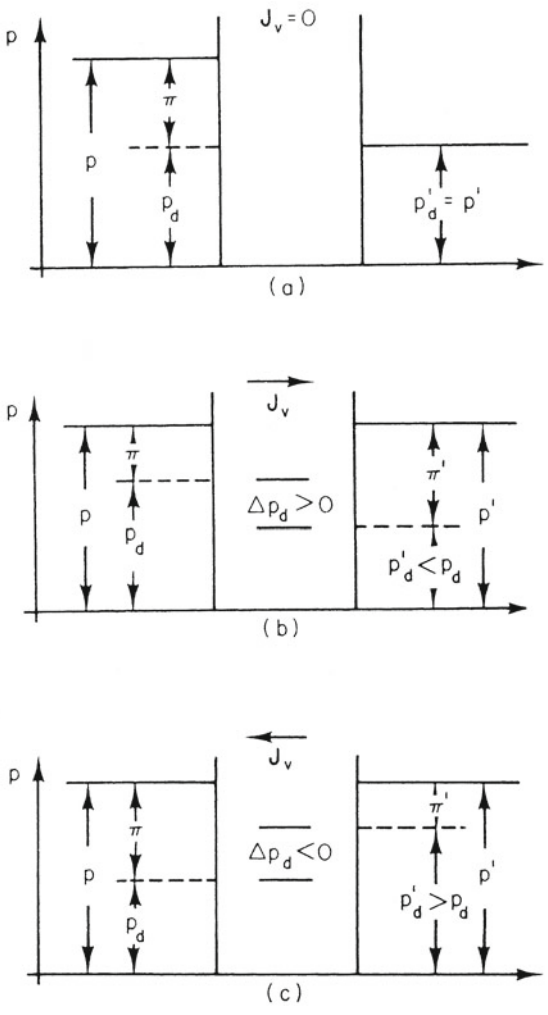


Fig. 5.6 Different flow possibilities for a completely impermeant solute. **a** $\Delta p_d = 0$, so there is no flow even though $p > p'$. **b** Flow to the right even though $p = p'$. **c** Flow to the left even though $p = p'$

When the solute is partially permeant, the volume fluence rate in the linear approximation still depends on both Δp and $\Delta\pi$, but the proportionality constants may be different. Since the solute does not reduce the flow as much as in Eq. 5.11, it is customary to write the two constants as L_p and σL_p:

$$J_v = L_p(\Delta p - \sigma\Delta\pi). \tag{5.12}$$

Parameter L_p is determined by measuring J_v and Δp when $\Delta\pi = 0$, while σ is determined from measurements of Δp and $\Delta\pi$ when $J_v = 0$.

Parameter σ is called the *reflection coefficient*. It has different values for different solutes. When $\sigma = 0$ there is no reflection, and the solute particles pass through like water. When $\sigma = 1$ all the solute particles are reflected and Eq. 5.12 is the same as Eq. 5.11.

We can imagine that part of the solute moves freely with the water and part is reflected. (Later, we will consider a model for partial reflection in which a solute particle of radius $a < R_p$ can enter the pore, but its center cannot be closer to the wall than its radius.) We can write

$$p = p_d + \sigma\pi, \tag{5.13}$$

and we can further break this down to a driving pressure for the water p_{dw} and one for the permeant solute:

$$p = \underbrace{p_{dw}}_{\substack{\text{driving pressure}\\ \text{for permeant}\\ \text{molecules}}} \overbrace{+(1-\sigma)\pi}^{\substack{\text{osmotic pressure}\\ \text{of all solute molecules}}} \underbrace{+\sigma\pi}_{\substack{\text{osmotic pressure}\\ \text{of impermeant}\\ \text{molecules}}}. \tag{5.14}$$

With this substitution the flow equation becomes

$$J_v = L_p\left[\Delta p_{dw} + (1-\sigma)\Delta\pi\right]. \tag{5.15}$$

Figure 5.7 shows the pressure relationships across the membrane.

In the approximation that van't Hoff's law holds, $\pi = k_B T C = RTc$ and Eq. 5.12 can be written as

$$J_v = L_p(\Delta p - \sigma k_B T\,\Delta C), \tag{5.16}$$

$$J_v = L_p(\Delta p - \sigma RT\,\Delta c). \tag{5.17}$$

In Eq. 5.16 the concentration is in molecules m^{-3}; in Eq. 5.17 it is mol m^{-3}. In both cases the units of $k_B T\,\Delta C$ and $RT\,\Delta c$ are pascals.

As an example of volume flow, consider *ultrafiltration*. Ultrafiltration is the process whereby water and small molecules are forced through a membrane by a hydrostatic

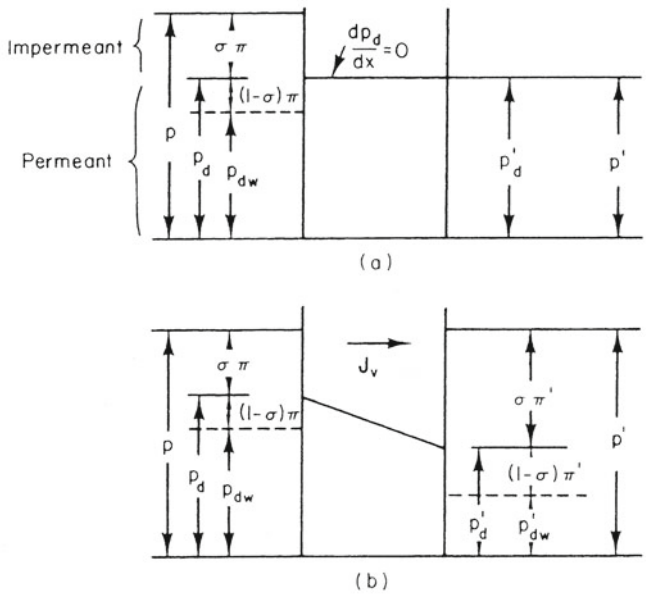


Fig. 5.7 Pressure relationships on each side of the membrane when $\sigma = \frac{2}{3}$. **a** There is no bulk flow. **b** There is flow to the right

pressure difference while larger constituents are left behind. An interesting clinical application of ultrafiltration has been proposed. A severely edematous patient (for any of the reasons mentioned in the previous section) must have the extra water removed from the body. This is usually accomplished with diuretics, drugs that increase the renal excretion of water. Some patients may not respond to these drugs, and in other cases, particularly pulmonary edema, the response may not be fast enough. In the latter case, phlebotomy (bloodletting) is sometimes used to reduce the body water rapidly. This has obvious disadvantages, for example, the removal of blood cells. Silverstein et al. (1974) used ultrafiltration to remove water and sodium from the plasma while leaving the other constituents behind. Ultrafiltration is sometimes called reverse osmosis. The name is unfortunate, because it suggests some mysterious process unrelated to the principles of this section. Ultrafiltration is often used by campers for purifying water and has been suggested for desalinization of sea water.

5.6 Solute Transport Through a Membrane

Solute can pass through the membrane in two ways: it can be carried along with flowing water (solvent drag), and it can diffuse.

If there is no reflection ($\sigma = 0$) and the solute concentration is the same on both sides of the membrane so there is no diffusion, the flux density or fluence rate is caused by solvent drag and is simply the solute concentration (particles per unit volume) times the volume fluence rate (Sect. 4.2):

$$J_s = C_s J_v.$$

If the solute particles are completely reflected ($\sigma = 1$) then $J_s = 0$.

In the intermediate case with coefficient σ,

$$J_s = (1 - \sigma) C_s J_v.$$

This is consistent with the idea expressed by Eq. 5.14 that a fraction $(1 - \sigma)$ of the solute particles can enter the membrane. In that case, C_s is the outside solute concentration on both sides of the membrane, and $C_s(1 - \sigma)$ is the solute concentration inside the membrane. We will develop a detailed model for transport in a right-cylindrical pore in Sect. 5.9. We anticipate that discussion and present a simple justification of the factor $1 - \sigma$. In bulk solution the concentration C_s is obtained by imagining a certain volume of solution, counting the number of solute particles whose centers lie within the volume, and taking the ratio. In a cylindrical pore of radius R_p and length ΔZ, the volume of fluid is $\pi R_p^2 \Delta Z$. The centers of solute particles of radius a cannot be within distance a of the pore wall. The number of solute particles within the pore is therefore $C_s \pi \left(R_p - a\right)^2 \Delta Z$. The concentration in the pore is the number of particles divided by the pore volume:

$$C_{s,\,\text{inside}} = \frac{C_s \pi \left(R_p - a\right)^2 \Delta Z}{\pi R_p^2 \Delta Z}$$

$$= C_s \left(1 - \frac{a}{R_p}\right)^2 = C_s \left(1 - \sigma\right).$$

This correction is called the *steric factor*. Solvent flow within a distance a of the walls contributes to J_v but not to solvent drag. This model will be extended to a volume flow with a parabolic velocity profile in Sect. 5.9.4.

If $J_v = 0$ there will be no solvent drag but there will be diffusion. The solute flux will be proportional to the concentration gradient and therefore to the concentration difference across the membrane: $J_s \propto \Delta C_s$. The proportionality constant depends on properties of the membrane. If the membrane is pierced by pores, for example, it depends on pore size, membrane thickness, number of pores per unit area, and the diffusion constant. The dependence will be derived later in this chapter. It is customary to write the proportionality constant as ωRT: $J_s = \omega RT \Delta C_s$. The factor ω is called the *membrane permeability* or *solute permeability*.

In the linear approximation the fluence rate resulting from both processes is the sum of these two terms:

$$J_s = (1 - \sigma)\overline{C}_s J_v + \omega RT \, \Delta C_s. \qquad (5.18)$$

Here an average value $\overline{C}_s$ has been written for the solvent drag term, because the concentration on each side of the membrane is not necessarily the same. The way that this average is taken will become clearer in the discussion of the pore model described in Sect. 5.9.

The solute equation has been written for both fluence rate and concentration in terms of particles. In terms of molar fluence rate and concentration, it is exactly the same:

$$J_s(\text{molar}) = (1-\sigma)\bar{c}_s J_v + \omega RT\,\Delta c_s. \tag{5.19}$$

Either way, the diffusion proportionality constant is ωRT. It does not change because C_s and J_s(particles) are both written in terms of particles, and c_s and J_s(molar) are both written in terms of moles. Referring to Eq. 5.18, the solvent drag term has units of (particles m^{-3}) (m s^{-1}) = particles m^{-2} s^{-1}. Therefore the factor ωRT has units of m s^{-1}. Since the units of RT are joules or N m (per mole), the units of ω are

$$\frac{\text{mol m s}^{-1}}{\text{N m}} = \text{mol N}^{-1}\,\text{s}^{-1}. \tag{5.20}$$

Further interpretation of ω will be made for specific models.

We have used the same σ in both the solvent drag term and in the preceding section. Although this was made plausible by saying that $1-\sigma$ is the fraction of solute molecules that gets through the membrane, its rigorous proof is more subtle. It has been proved in general using thermodynamic arguments, which can be found in Katchalsky and Curran (1965). It can be proved in detail for specific membrane models.

5.7 Example: The Artificial Kidney

The artificial kidney provides an example of the use of the transport equations to solve an engineering problem. The problem has been extensively considered by chemical engineers, and we will give only a simple description here. Those interested in pursuing the problem further can begin with reviews by Mavroidis (2006) or Lysaght and Moran (2006). The reader should also be aware that this "high-technology" solution to the problem of chronic renal disease is not entirely satisfactory. It is expensive and uncomfortable and leads to degenerative changes in the skeleton and severe atherosclerosis (Lindner et al. 1974). The alternative treatment, a transplant, has its own problems, related primarily to the immunosuppressive therapy. Anyone who is going to be involved in biomedical engineering or in the treatment of patients with chronic disease should read the account by Calland (1972), a physician with chronic renal failure who had both chronic dialysis and several transplants. The distinction between a high-technology treatment and a real conquest of a disease has been underscored by Thomas (1974, pp. 31–36).

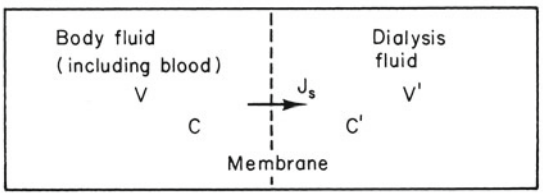


Fig. 5.8 The simplest model of dialysis. All the body fluid is treated as one compartment; transport across the membrane is assumed to take longer than transport from various body compartments to the blood

The simplest model of dialysis is shown in Fig. 5.8. Two compartments, the body fluid and the dialysis fluid, are separated by a membrane that is porous to the small molecules to be removed and impermeable to larger molecules. If such a configuration is maintained long enough, then the concentration of any solute that can pass through the membrane will become the same on both sides. The dialysis fluid is prepared with the desired composition of such small molecules as sodium, potassium, and glucose. Volume V' must be larger than V for effective dialysis to take place; otherwise, the concentration of solutes in the dialysis fluid builds up from the initially prepared values. In early work, V' was up to 100 l (since V is about 40 l). Although the fluid was replaced every two hours or so, it was an excellent medium in which to grow bacteria. Although the bacteria could not get through the membrane, they released exotoxins (or, if they died, endotoxins) which diffused back into the patient and caused fever. Now a continuous flow system has been used in which the solutes are continually metered into flowing dialysis fluid that is then discarded. Because of this, we will assume that there is no buildup of concentration in the dialysis fluid. (Effectively volume V' is infinite.) We will assume that $\Delta p = 0$. (Actually, proteins cause some osmotic pressure difference, which we will ignore.)

Without solvent drag, the solute transport is by diffusion, $J_s = \omega RT(C-C')$, where C is the concentration of solute in the blood and C' is the concentration in the dialysis fluid. If the surface area of the membrane is S, then the rate of change of the number of solute molecules N is

$$\frac{dN}{dt} = -S\omega RT(C-C').$$

If the solute is well mixed in the body fluid compartment, then $N = CV$, and this equation can be written as

$$\frac{dC}{dt} = -\frac{S\omega RT}{V}(C-C').$$

This is the equation for exponential decay. The steady state solution is $C = C'$. The complete solution is (Appendix F)

$$C(t) = \left[C(0) - C'\right]e^{-t/\tau} + C', \tag{5.21}$$

where the time constant is

$$\tau = \frac{V}{S\omega RT}. \tag{5.22}$$

The only things that are adjustable in this equation are the membrane area S and its permeability ω. The size of pores in the membrane is dictated by what solutes are to be retained in the blood. The number of pores per unit area and the thickness of the membrane can be controlled. Typical cellophane membranes have $\omega RT = 5 \times 10^{-6}$ m s^{-1} (with a thickness of 500 μm). The area may be 2 m^2. With a fluid volume $V = 40$ l, this gives

$$\tau = \frac{40 \times 10^{-3}\,\text{m}^3}{(2\,\text{m}^2)\,(5 \times 10^{-6}\,\text{m s}^{-1})} = 4 \times 10^3\,\text{s} = 1.1\,\text{h}.$$

Typically, dialysis requires several hours. This longer period is for two reasons. Some of the larger molecules have smaller permeabilities and therefore longer time constants, and rapid dialysis causes cerebral edema and severe headaches.

The actual apparatus is quite complicated. First, it must be sterile, which requires a sterilized, disposable dialysis membrane. Second, the apparatus causes clots, so the blood must be treated with heparin as it enters the machine, and the heparin must be neutralized with protamine as it returns to the patient.

5.8 Countercurrent Transport

This section considers a problem that demonstrates the principle of *countercurrent* transport. An apparatus (perhaps a dialysis machine or an oxygenator) transports a single solute across a thin membrane of permeability ωRT. On one side of the membrane (the "inside") is a thin layer of solvent that flows along the membrane in the $+x$ direction as shown in Fig. 5.9. On the "outside" is another thin layer of solvent that may be at rest or may flow in either the $+x$ or the $-x$ direction. When it flows in the opposite direction of the fluid inside we have the countercurrent situation.

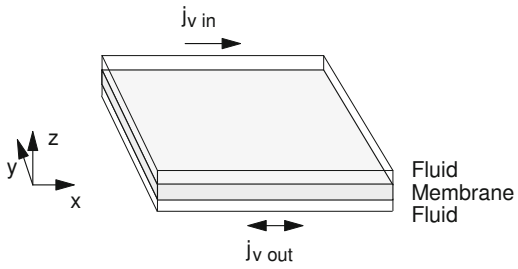
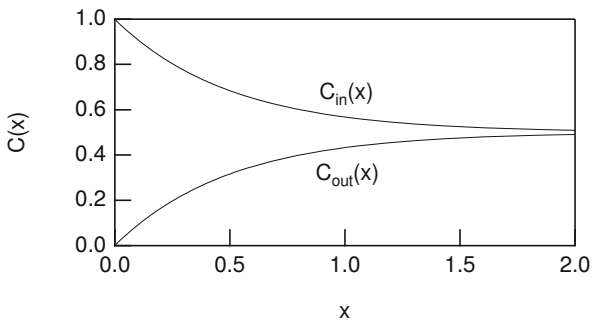


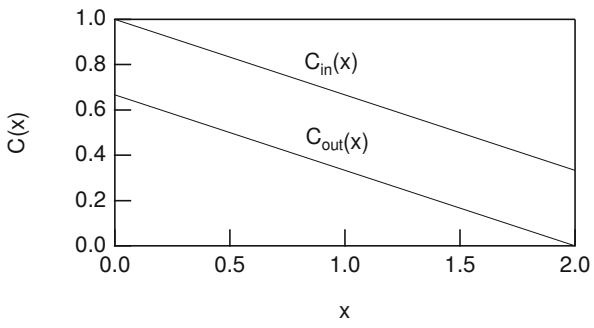
Fig. 5.9 Layers of fluid containing a solute flow parallel to the x axis on either side of a membrane

Suppose that the concentration of solute in the two layers is $C_{\text{in}}(x)$ inside and $C_{\text{out}}(x)$ outside. Solute is transported in the x direction in each fluid layer by pure solvent drag. It diffuses through the membrane from the side with higher concentration to the other. We develop the model below and show that the steady-state concentration profiles are quite different depending on whether the solvent flows are in the same or opposite directions. The results are shown in Fig. 5.10 for the situation in which the value of C_{in} is 1 and the value of C_{out} is 0 where each solvent starts to flow across the membrane. In Fig. 5.10a both layers flow to the right; in Fig. 5.10b they flow in opposite directions. The countercurrent case is more effective in reducing C_{in}. The final value of C_{in} is 0.5 in the first case and 0.33 in the second.

To develop the model, we make the following assumptions. The concentration of solute in each fluid layer is independent of y, z, and t. The thickness of the fluid layer inside is h_{in}. The fluid velocity $j_{v\,\text{in}}$ is everywhere constant. The only important mechanism for solute transport within the fluid is solvent drag. Let the width of the slab in the



(a) Both flows are to the right.



(b) The flows are in opposite directions.

Fig. 5.10 Solute concentration profiles for two different situations where solvent flows parallel to the membrane surface and solute moves through the membrane from inside to outside. **a** Both fluid layers flow to the right. The concentrations rise and falls exponentially, eventually becoming the same on both sides of the membrane. **b** The countercurrent case, in which the solvent flows are in opposite directions. The solvent outside flows from right to left. The concentrations vary linearly

y direction be Y. Inside, the number of particles per second in through the face of the rectangle of area Yh_{in} at x is $C_{\text{in}}(x)\,j_{v\,\text{in}}Yh_{\text{in}}$. The number out through the face at $x+dx$ is $C_{\text{in}}(x+dx)\,j_{v\,\text{in}}Yh_{\text{in}}$. The number through the membrane into the exterior volume is $[C_{\text{in}}(x)-C_{\text{out}}(x)]\,\omega RTYdx$. Combining these we get

$$\frac{dC_{\text{in}}}{dx}=-\frac{\omega RT}{j_{v\,\text{in}}h_{\text{in}}}\left[C_{\text{in}}(x)-C_{\text{out}}(x)\right]. \tag{5.23}$$

A similar expression can be derived for the exterior:

$$\frac{dC_{\text{out}}}{dx}=\frac{\omega RT}{j_{v\,\text{out}}h_{\text{out}}}\left[C_{\text{in}}(x)-C_{\text{out}}(x)\right]. \tag{5.24}$$

Our notation allows j_v to have a different direction (sign). Defining $a=\omega RT/j_vh$ we have the coupled differential equations

$$\begin{aligned}\frac{dC_{\text{in}}}{dx}&=-a_{\text{in}}(C_{\text{in}}-C_{\text{out}}),\\ \frac{dC_{\text{out}}}{dx}&=+a_{\text{out}}(C_{\text{in}}-C_{\text{out}}).\end{aligned} \tag{5.25}$$

We restrict ourselves to the case in which $|a_{\text{in}}|=|a_{\text{out}}|=a$. Changing the direction of j_v changes the sign of a. Assume a is the same on both sides. The equations show that the slope of $C_{\text{in}}(x)$ is minus the slope of $C_{\text{out}}(x)$ if both currents are in the same direction, and the two slopes are the same if the currents are in opposite directions. This can be seen in the solutions in Fig. 5.10.

You can verify that Eqs. 5.26 represent a solution of Eqs. 5.25:

$$\begin{aligned}C_{\text{in}}(x)&=\frac{c_1}{2}\left(1+e^{-2ax}\right)+\frac{c_2}{2}\left(1-e^{-2ax}\right),\\ C_{\text{out}}(x)&=\frac{c_1}{2}\left(1-e^{-2ax}\right)+\frac{c_2}{2}\left(1+e^{-2ax}\right),\end{aligned} \tag{5.26}$$

where c_1 and c_2 are the values of C_{in} and C_{out} at $x=0$. Figure 5.10a shows the concentrations for $c_1=1$ and $c_2=0$ with $a=1$ and $0<x<2$. If the sign of a is changed in the second differential equation, then the fluid outside is flowing in the opposite direction to the fluid inside. Again you can verify that the most general solution is

$$\begin{aligned}C_{\text{in}}(x)&=c_1+(c_2-c_1)ax,\\ C_{\text{out}}(x)&=c_2+(c_2-c_1)ax.\end{aligned} \tag{5.27}$$

Figure 5.10b is a plot with the constants set so that the concentration inside on the left is 1 and on the outside on the right is zero ($c_1=1, c_2=2/3, a=1, 0<x<2$). This configuration is called *countercurrent* flow. We can see from the figure that the transport through the membrane is increased because the concentration difference across the membrane is, on average, greater.

The countercurrent principle is found in the renal tubules (Hall 2011, p. 309; Patton et al. 1989, p. 1081), in the villi of the small intestine (Patton et al. 1989, p. 915), and in the lamellae of fish gills (Schmidt-Nielsen 1972, p. 45). The principle is also used to conserve heat in the extremities—such as people's arms and legs, whale flippers, or the leg of a duck. If a vein returning from an extremity runs closely parallel to the artery feeding the extremity, the blood in the artery will be cooled and the blood in the vein warmed. As a result, the temperature of the extremity will be lower and the heat loss to the surroundings will be reduced.

5.9 A Continuum Model for Volume and Solute Transport in a Pore

In this section we develop a model to predict the values of the phenomenological coefficients of Sects. 5.5 and 5.6. The success of the model depends on its ability to predict behavior, particularly as the size of solute particles is varied. This was an important problem in physiology in the 1960s and 1970s. Instead of comparing the model to experiment, we conclude the section by showing what the forces are on the membrane. This is important because there has been a fair amount of confusion in the literature about the forces on a semipermeable membrane. This section is fairly long. It stands alone; you can skip it if you wish.

The model assumes that the membrane has a particularly simple structure.

1. The membrane is pierced by n circular pores per unit area, all having radius R_p and all being right cylinders. The membrane thickness is ΔZ.
2. The pore and the fluid are electrically neutral. No electrical forces are considered.
3. There is complete mixing on both sides of the pore, so that flow within the liquid on either side can be neglected.
4. The system is in the steady state. There is no variation in flux density (fluence rate) or concentration as a function of time.
5. The pores are large enough so that the bulk flow can be calculated by continuum hydrodynamics.

The quantities considered in this section are summarized in Table 5.1.

5.9.1 Volume Transport

The results of Chap. 1 can be used when the pore is filled with pure water or water and a solute for which $\sigma=0$. From Eq. 1.40 the flux through a single pore is

$$i_v(\text{single pore})=\frac{\pi R_p^4}{8\eta}\frac{\Delta p}{\Delta x}. \tag{5.28}$$

Table 5.1 Symbols used for porous membrane

Quantity	On left	In pore	On right
Total pressure	p		p'
Solute concentration	C_s	$C(z)$	C_s'
Osmotic pressure	$\pi = k_B T C_s$		$\pi' = k_B T C_s'$
Effectively impermeant part of osmotic pressure	$\sigma\pi$		$\sigma\pi'$
Effectively permeant part of osmotic pressure plus water driving pressure	$(1-\sigma)\pi + p_{dw}$	$p_d(z)$	$(1-\sigma)\pi' + p_{dw}'$

The fluence rate through the membrane is obtained by multiplying i_v by n, the number of pores per unit area. The result is

$$J_v = \frac{n\pi R_p^4}{8\eta}\frac{\Delta p}{\Delta Z}$$

so that

$$L_p = \frac{n\pi R_p^4}{8\eta\,\Delta Z}. \tag{5.29}$$

While L_p can be measured fairly easily using Eq. 5.12, it is much more difficult to measure the microscopic quantities needed to test Eq. 5.29. We will not compare the model to experiment here;[8] we will simply give an example of how calculations are done.

A commercial filter used for ultrafiltration might have the property $L_p \approx 1$ ml min^{-1} m^{-2} torr^{-1}. Since 760 torr $= 1 \times 10^5$ Pa, the hydraulic permeability in SI units is

$$\begin{aligned} L_p &= \frac{1\text{ ml}}{1\text{ torr min m}^2}\,\frac{1\text{ min}}{60\text{ s}}\,\frac{10^{-6}\text{ m}^3}{1\text{ ml}}\,\frac{760\text{ torr}}{1\times 10^5\text{ Pa}} \\ &= 1.27\times 10^{-10}\text{ m s}^{-1}\text{ Pa}^{-1}. \end{aligned}$$

The manufacturer's literature[9] can be used to estimate

$$R_p \approx 4.5\text{ nm},$$
$$\Delta Z \approx 10\ \mu\text{m}.^{10}$$

The viscosity of water is 0.9×10^{-3} Pa s at 25 °C. This gives us enough information to estimate n and the fraction of the filter surface that is pores. From Eq. 5.29

$$\begin{aligned} n &= \frac{8\eta\,\Delta Z\,L_p}{\pi R_p^4} = \left(\frac{(8)(0.9\times 10^{-3}\text{ Pa s})(10\times 10^{-6}\text{ m})}{\pi\left(4.5\times 10^{-9}\right)^4\text{ m}^4}\right) \\ &\quad\times (1.27\times 10^{-10}\text{ m s}^{-1}\text{ Pa}^{-1}) \\ &= 7.1\times 10^{15}\text{ m}^{-2}. \end{aligned}$$

Since the area of one pore is $\pi R_p^2 = 6.36\times 10^{-17}$ m^2, the total pore area in 1 m^2 is 0.45 m^2, a number that is not unreasonable.

Next consider the volume flow when the reflection coefficient is not zero. The position within the pore is specified by cylindrical coordinates (r, ϕ, z). The position along the axis of the pore is given by z. The position in a plane perpendicular to the axis of the pore is specified by polar coordinates r and ϕ. Flow of the fluid is described by the vector volume fluence rate $\mathbf{j}_v(r, \phi, z)$. (We use J for fluence rate for the membrane as a whole and j for the fluence rate in bulk solution inside a pore.) It is possible to show rigorously that as long as the pore is a right circular cylinder, $\mathbf{j}_v$ points only along z and is independent of ϕ (the fluid does not flow in a spiral and does not flow into or out of the walls):

$$\mathbf{j}_v(r, \phi, z) = j_v(r, z)\hat{\mathbf{z}}. \tag{5.30}$$

The solution is in a steady state and the flow is not changing with time. Therefore the flux density into a volume at z must be the same as the flux density out at $z + dz$:

$$\frac{\partial j_v}{\partial z} = 0 \tag{5.31}$$

so that j_v is constant along the z axis (although it can be a function of r). This is just what we saw in Chap. 1 for Poiseuille flow; the variation of j_v with r corresponds to the parabolic velocity profile. A value of $j_v(r)$ that is constant in the z direction requires a constant value of $\partial p/\partial z$ inside the pore.

In the pore, the driving pressure is $p_d(z)$. A typical pressure profile is shown in Fig. 5.11. The symbols are defined in Table 5.1. The pressure in the pore has been drawn with constant slope, since $\partial p_d/\partial z$ is constant. Using Eqs. 5.16 and 5.29, we can write

$$J_v = L_p(\Delta p - \sigma k_B T\,\Delta C_s), \tag{5.32}$$

where L_p is given by Eq. 5.29. The value of σ is derived in the next section.

The average value of $j_v(r)$ within the pore will be called $\bar{j}_v$. It is the total flux density through the pore divided by πR_p^2:

$$\bar{j}_v = \frac{i(\text{single pore})}{\pi R_p^2} = \frac{1}{\pi R_p^2}\int_0^{R_p} j_v(r)\,2\pi r\,dr$$

[8] See the third or earlier editions or, for example, Bean (1969, 1972).

[9] Amicon XM-50.

[10] This value may not be consistent with the value of L_p quoted. The pore length ΔZ is not well known, and L_p is variable, depending on experimental conditions.

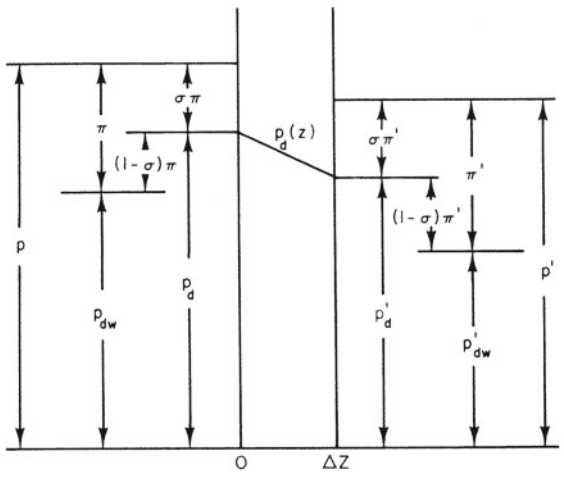


Fig. 5.11 Pressure within a pore and at the boundaries in the steady state

$$= \frac{J_v}{n\,\pi R_p^2} = -\frac{R_p^2}{8\eta}\frac{\partial p_d}{\partial z}. \tag{5.33}$$

5.9.2 Solute Transport

We now consider solute transport in our model pore. The arguments here are very similar to those for combined diffusion and solvent drag that were developed in Sect. 4.12. Those arguments are extended by averaging over the cross section of the pore.

Within the pore, the local solute flux is $\mathbf{j}_s(r, \phi, z)$. Arguments similar to those in the preceding section can be offered to show that $\mathbf{j}_s$ points along the z axis and is independent of ϕ:

$$\mathbf{j}_s(r, \phi, z) = j_s(r, z)\hat{\mathbf{z}}. \tag{5.34}$$

The solute concentration does not depend on ϕ, or else there would be diffusion in the ϕ direction and $\mathbf{j}_s$ would have a ϕ component. So $C = C(r, z)$. The r dependence must be kept because the center of a solute molecule of radius a cannot be within a distance a of the wall. (Recall the discussion of the steric correction on p. 125) Thus $C(r, z) = 0$ if $r > R_p - a$. We write[11]

$$C(r, z) = \begin{cases} 0, & R_p - a < r \\ C(z), & 0 \leqslant r \leqslant R_p - a. \end{cases} \tag{5.35}$$

The solute flux due to solvent drag is $C_s j_v$. For diffusion in one dimension the solute flux along the z axis is $-D(\partial C/\partial z)$. For the cylindrical pore we can combine these and write

$$j_s(r, z) = C(r, z)\,j_v(r, z) - D(r, a, R_p)\frac{\partial C(r, z)}{\partial z}. \tag{5.36}$$

The diffusion constant has been written as a function of r, a, and R_p because in the pore, as distinct from an infinite medium, the constant depends on how close the particle is to the walls. (Remember the relation of D to the viscous drag and the fact that Stokes' law requires modification when the fluid is confined in a tube.)

The preceding section showed that for the steady state j_v is independent of z. A similar argument can be made using the continuity equation for solute particles, implying that j_s is independent of z. Therefore Eq. 5.36 simplifies to

$$D(r, a, R_p)\frac{\partial C(r, z)}{\partial z} - j_v(r)C(r, z) = -j_s(r). \tag{5.37}$$

The easiest way to write $C(r, z)$ in accordance with Eq. 5.35 is

$$C(r, z) = C(z)\Gamma(r),$$

where

$$\Gamma(r) = \begin{cases} 0, & R_p - a < r \\ 1, & 0 \leqslant r < R_p - a. \end{cases} \tag{5.38}$$

With this substitution Eq. 5.37 becomes

$$\Gamma(r)D(r, a, R_p)\frac{dC(z)}{dz} - C(z)\Gamma(r)j_v(r) = -j_s(r).$$

This equation can be multiplied by $2\pi r\,dr$ and integrated from $r = 0$ to $r = R_p$. The result is

$$\left(\int_0^{R_p} \Gamma(r)D(r, a, R_p)2\pi r\,dr\right)\frac{dC(z)}{dz} - \left(\int_0^{R_p} \Gamma(r)\,j_v\,2\pi r\,dr\right)C(z) = -\int_0^{R_p} j_s(r)2\pi r\,dr. \tag{5.39}$$

The physical meaning of this integration can be understood with the aid of Fig. 5.12, which shows a slab of fluid in the pore between z and $z+dz$. Solute does not cross a surface of constant r but moves parallel to the z axis. Diffusion and solvent drag are considered in each shaded area $2\pi r\,dr$. The integration of Eq. 5.39 establishes an average solute fluence rate, since the right-hand side of the equation is the total flux or current of solute particles per second passing through the pore:

$$i_s = \int_0^{R_p} j_s(r)2\pi r\,dr.$$

[11] It can be argued that this is the only possible form for $C(r, z)$. See Levitt (1975, p. 535ff.).

As with the volume fluence rate, it is convenient to call the average solute fluence rate $\overline{j}_s$:

$$\overline{j}_s = \frac{i_s}{\pi R_p^2} = \frac{1}{\pi R_p^2}\int_0^{R_p} j_s(r)2\pi r\, dr. \tag{5.40}$$

The first term of Eq. 5.39 is the diffusive flux at z averaged over the entire cross section of the pore. Define an effective diffusion constant

$$D_{\text{eff}} = \frac{1}{\pi R_p^2}\int_0^{R_p} \Gamma(r)D(r, a, R_p)2\pi r\, dr. \tag{5.41}$$

The second term on the left of Eq. 5.39 is the solvent drag flux averaged over the entire cross section of the pore. The integral is

$$\int_0^{R_p} j_v(r)\Gamma(r)2\pi r\, dr = \int_0^{R_p-a} j_v(r)2\pi r\, dr. \tag{5.42}$$

This integral can be evaluated because we know the velocity profile, $j_v(r)$, Eq. 1.39:[12]

$$j_v(r) = \frac{1}{4\eta}\frac{\Delta p}{\Delta z}\left(R_p^2 - r^2\right). \tag{5.43}$$

We have already defined the average volume fluence rate to be

$$\overline{j_v} = \frac{1}{\pi R_p^2}\int_0^{R_p} j_v(r)2\pi r\, dr.$$

The desired quantity differs only in the limits of integration. To calculate it, write

$$\int_0^{R_p-a} j_v(r)2\pi r\, dr = \pi R_p^2\, \overline{j}_v \frac{\displaystyle\int_0^{R_p-a} j_v(r)2\pi r\, dr}{\displaystyle\int_0^{R_p} j_v(r)2\pi r\, dr}.$$

[12] This ignores the fact that since the walls affect the force on the solute particles, the solute must distort the velocity profile slightly. This point is discussed below.

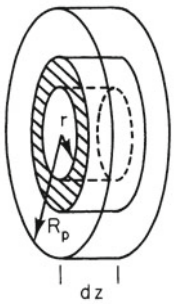


Fig. 5.12 A slab of fluid in a pore between z and $z + dz$, showing how the integration over r is done

The integrals are easily evaluated (see the Problems). The result is

$$\int_0^{R_p} j_v(r)\Gamma(r)2\pi r\, dr = \pi R_p^2\, \overline{j}_v\, f(a/R_p), \tag{5.44a}$$

where the function f is

$$f(\xi) = 1 - 4\xi^2 + 4\xi^3 - \xi^4. \tag{5.44b}$$

When Eqs. 5.40, 5.41, and 5.44a are substituted into Eq. 5.39 and each term is divided by πR_p^2, the result is

$$D_{\text{eff}}\left(\frac{dC}{dz}\right) - j_v f\left(\frac{a}{R_p}\right)C(z) = -\overline{j}_s \tag{5.45a}$$

or

$$\frac{dC}{dz} - \frac{j_v f(a/R_p)}{D_{\text{eff}}}C(z) = -\frac{\overline{j}_s}{D_{\text{eff}}}. \tag{5.45b}$$

This is a differential equation for $C(z)$. The right-hand side is the total solute fluence rate, which is constant. On the left-hand side, C varies along the pore so that the diffusive and solvent-drag fluence rates add up to this constant value. If the constant in front of $C(z)$ is written as

$$\frac{1}{\lambda} = \frac{j_v f(a/R_p)}{D_{\text{eff}}}, \tag{5.46}$$

this is recognized as Eq. 4.58 for drift plus solvent drag in an infinite medium. The results of Sect. 4.13 can be applied here. It is only necessary to determine values for C_0 and C_0'. Recall that in the pore $C(r, z) = C(z)\Gamma(r)$. The function $\Gamma(r)$ takes into account the reflection that occurs because solute particles cannot be closer to the pore wall than their radius. It was also assumed that the solution on either side of the membrane is well stirred. Therefore, $C_0 = C_s$ and $C_0' = C_s'$. Equation 4.70 becomes

$$\overline{j}_s = f\,\overline{j}_v\,\overline{C}_s + D_{\text{eff}}\frac{(C_s - C_s')}{\Delta Z}. \tag{5.47}$$

This is an expression for $\overline{j}_s$, the average solute fluence rate in the pore. To get solute fluence rate in the membrane, it must be multiplied by πR_p^2 and the number of pores per unit area. Since $J_v = n\pi R_p^2\,\overline{j}_v$, we have

$$J_s = f\overline{C}_s\, J_v + \frac{n\pi R_p^2\, D_{\text{eff}}}{\Delta Z}C_s. \tag{5.48}$$

Comparing this with the general phenomenological equation for solute flow, Eq. 5.18,

$$J_s = (1 - \sigma)\overline{C}_s J_v + \omega RT\,\Delta C_s$$

we see that

$$1 - \sigma = f,$$

$$\omega RT = \frac{n\,\pi R_p^2\,D_{\text{eff}}}{\Delta Z}, \tag{5.49}$$

$$\lambda = \frac{D_{\text{eff}}}{\bar{j}_v(1-\sigma)} = \frac{\omega RT(\Delta Z)}{J_v(1-\sigma)}.$$

The average solute concentration C is obtained from Eq. 4.66 with the substitution of ΔZ for the pore length:

$$\overline{C}_s = \frac{C_s e^x - C_s'}{e^x - 1} - \frac{1}{x}(C_s - C_s').$$

This can be rearranged as

$$\overline{C}_s = \tfrac{1}{2}(C_s + C_s') + G(x)(C_s - C_s') \tag{5.50a}$$

with

$$G(x) = \frac{1}{2}\left(\frac{e^x + 1}{e^x - 1}\right) - \frac{1}{x}, \tag{5.50b}$$

where $x = \Delta Z/\lambda$. This is the same function we saw in Fig. 4.17.

The solute concentration away from the sides of the pore is

$$C(z) = \frac{C_s(e^{\Delta Z/\lambda} - e^{z/\lambda}) + C_s'(e^{z/\lambda} - 1)}{e^{\Delta Z/\lambda} - 1}. \tag{5.51}$$

While the concentration profile is not usually measured experimentally, it is useful to plot it to help us visualize the interrelation of diffusion and solvent drag. Call $\phi = C_s'/C_s$. Equation 5.51 can be rearranged as

$$C(z) = C(0)\left(1 - (1-\phi)\frac{e^{z/\lambda} - 1}{e^{\Delta Z/\lambda} - 1}\right). \tag{5.52}$$

We can see several things from this equation. First, if the concentration is the same at each end of the pore, $\phi = 1$, the second term in the large parentheses vanishes, and the concentration is uniform throughout the pore. If $\phi \neq 1$, then the concentration is that at $z = 0$, plus a factor which may be positive or negative, depending on whether ϕ is less than or greater than 1. The ratio of exponentials occurring in that factor is plotted in Fig. 5.13 for different values of $\Delta Z/\lambda$, the ratio of the pore length to the effective diffusion distance.

These curves determine the shape of the concentration profile along the pore. If the flow is zero, $\lambda = D_{\text{eff}}/\bar{j}_v(1-\sigma)$ is infinite and $\Delta Z/\lambda$ is zero. We then have pure diffusion, and the concentration changes uniformly along the pore, corresponding to the straight line in Fig. 5.13. The plots in Fig. 5.14 show what the concentration profiles are like for diffusion to the left and to the right when the flow is to the right. Compare the shape of the concentration profile on the left in Fig. 5.14 with the curve for $\Delta Z/\lambda = 1$ in Fig. 5.13. When the concentration is higher on the left, we have to take the mirror image of Fig. 5.13; the curve for $\Delta Z/\lambda = -1$ gives the concentration profile in Fig. 5.14 on the right.

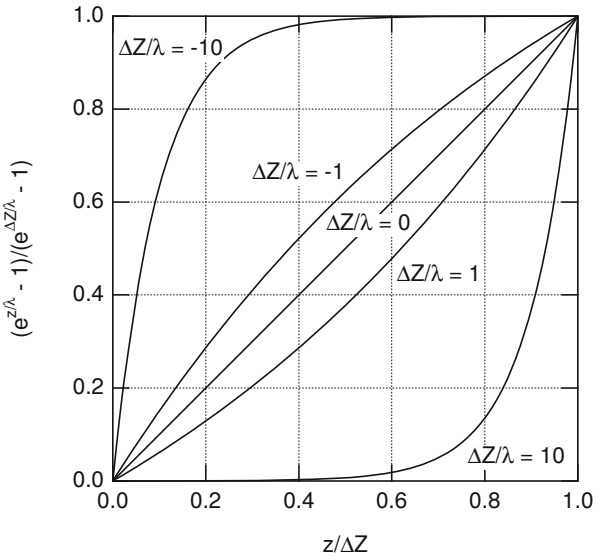


Fig. 5.13 Plot of the factor $(e^{z/\lambda} - 1)/(e^{\Delta Z/\lambda} - 1)$, which appears in Eq. 5.52

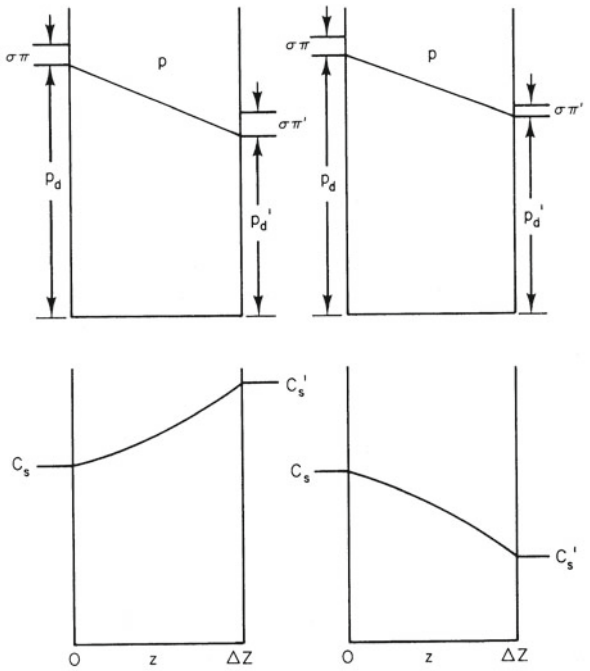


Fig. 5.14 A possible set of values for p, p_d, and C along a pore for diffusion to the left and diffusion to the right. The fluid on each side of the pore is well stirred and of sufficient volume so that concentrations do not change with time

As the pore becomes very long compared to the diffusion length (for example, $|\Delta Z/\lambda| = 10$ or more), the concentration along the pore is nearly that carried into the pore by bulk flow from the left until we get to the far end, where diffusion back up the pore gives a smooth transition to the final concentration on the right.

We can think of the pressure in the pore as being made up of driving pressures due to water and to the solute within the pore:

$$p_d(z) = p_{dw}(z) + p_{ds}(z).$$

Since the effective driving pressure for impermeant solute in the J_v equation is $k_B T\,\Delta C$, it would be nice to be able to write

$$p_d(z) = p_{dw}(z) + (1-\sigma)k_B T\, C(z).$$

This is consistent with the solvent drag flux at position z in the pore, which was given in Eq. 5.45a by

$$\overline{j}_v\, f\, C(z) = \overline{j}_v(1-\sigma)C(z).$$

The "effective" concentration for solvent drag is $(1-\sigma)C(z)$.

5.9.3 Summary

To summarize, the combination of solvent and a solute with reflection coefficient has a volume flux

$$J_v = L_p(\Delta p - \sigma\, k_B T\, \Delta C_s) \tag{5.53}$$

and a solute flux

$$J_s = (1-\sigma)\overline{C}_s J_v + \omega RT\, \Delta C_s. \tag{5.54}$$

The hydraulic permeability is

$$L_p = \frac{n\,\pi R_p^4}{8\eta \Delta Z}. \tag{5.55}$$

The solute permeability is

$$\omega RT = \frac{n\,\pi R_p^2\, D_{\text{eff}}}{\Delta Z}. \tag{5.56}$$

The characteristic length for diffusion is

$$\lambda = \frac{D_{\text{eff}}}{\overline{j}_v(1-\sigma)} = \frac{\Delta Z\,\omega RT}{J_v(1-\sigma)}. \tag{5.57}$$

The average concentration is

$$\overline{C}_s = \tfrac{1}{2}(C_s + C_s') + G(x)\,\Delta C_s, \tag{5.58}$$

where $G(x)$ is given by Eq. 5.50b. The parameter x is

$$x = \frac{J_v(1-\sigma)}{\omega RT} = \frac{\Delta Z}{\lambda}. \tag{5.59}$$

Notice that the solvent drag term as well as the diffusion term depends on ΔC_s, through the factor $\overline{C}_s$.

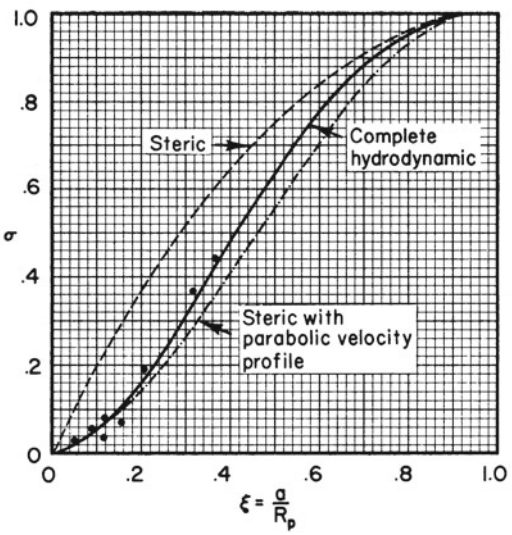


Fig. 5.15 Calculated values of the reflection coefficient are indicated by the lines. Calculations are shown for the simple steric factor, the steric factor weighted by a parabolic velocity profile, Eq. 5.60, and a more detailed calculation, which takes account of the distortion of the velocity profile by the solute particles by Levitt (1975) and by Bean (1972, pp. 29–35). The data points are from Durbin (1960) as reinterpreted by Bean (1972)

5.9.4 Reflection Coefficient

We have referred previously to the fact that the centers of solute particles can occupy only a fraction of the pore volume. A solute particle's center cannot be further from the pore axis than $R_p - a$. The simplest correction is the steric factor, seen on p. 125 . The ratio of effective area to total area approximates $1-\sigma$. If $\xi = a/R_p$, then

$$1-\sigma \approx \frac{\pi(R_p-a)^2}{\pi R_p^2} = 1 - \frac{2a}{R_p} + \frac{a^2}{R_p^2},$$
$$\sigma = 2\xi - \xi^2.$$

A better calculation was seen in the preceding subsection. Accept the fact (quoted from thermodynamic results) that the same σ occurs in the equations for J_v and J_s. We saw that the edges of the pore have less bulk flow than the center, so that the steric effect overestimates how many particles are reflected. From Eq. 5.44b,

$$\sigma = 1 - f = 4\xi^2 - 4\xi^3 + \xi^4. \tag{5.60}$$

These two approximations to σ are plotted in Fig. 5.15, along with the results of a more detailed calculation that takes account of distortions in the velocity profile due to rotation of the solute molecules.

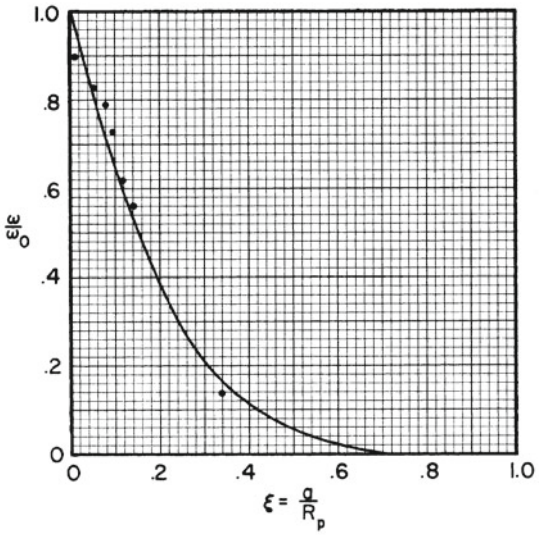


Fig. 5.16 Plot of ω/ω_0 for experimental data by Beck and Schultz (1970) and a calculation by Bean (1972)

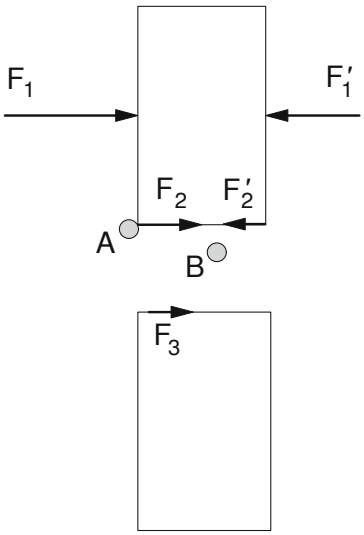


Fig. 5.17 The forces on a membrane with pores. The fluid on the left exerts force $\mathbf{F}_1$ due to the hydrostatic pressure p. A similar force $\mathbf{F}_1'$ is exerted on the right. Solute molecules like A are reflected at the pore edge and exert force $\mathbf{F}_2$. Solute molecule B enters the pore. It contributes to the viscous force of the flowing fluid on the cylindrical walls of the pore, $\mathbf{F}_3$. $\mathbf{F}_3$ is to the right if the fluid flows from left to right through the pore

5.9.5 The Effect of Pore Walls on Diffusion

The solute permeability is given by

$$\omega RT = \frac{n\pi R_p^2 D_{\text{eff}}}{\Delta Z}.$$

The effective diffusion coefficient takes into account the steric factor as well as the drag on the solute particles by the pore walls. If the pore had an infinitely large diameter, the unrestricted permeability would be

$$\omega_0 RT = \frac{n\pi R_p^2 D}{\Delta Z},$$

where D is the diffusion coefficient for an infinite medium. Figure 5.16 shows some data from Beck and Schultz (1970) and a curve for ω/ω_0 calculated by Bean (1972).[13]

In Europe, filtration rather than dialysis is used to treat kidney patients. There is evidence that some as yet unidentified toxin of medium molecular weight accumulates in the blood. Comparison of $1-\sigma$ from Fig. 5.15 with ω/ω_0 from Fig. 5.16 shows that solvent drag removes medium-sized molecules more effectively. The fluid and electrolytes lost by the patient must be replaced.

[13] The steric factor, which Bean includes separately, is built into D_{eff} through the function $\Gamma(r)$.

5.9.6 Net Force on the Membrane

We conclude the section by calculating the force of the fluid on our model membrane. The results give some insight into the nature of osmotic pressure.

A membrane of total area S is pierced by n pores per unit area of radius R_p. The pressures in the fluid on each side of the membrane are p and p'. A solute with reflection coefficient σ has concentration C on the left and C' on the right. We want to calculate the total force exerted by the fluid on the membrane. There are three contributions to this force. These can be understood by referring to Fig. 5.17.

Forces F_1 and F_1' are the forces exerted by the fluid on the walls of the membrane on each side. They are obtained by multiplying the total pressure on each side by the area of the membrane that is not occupied by pores. In a total area S there are nS pores, each of area πR_p^2.

$$F_1 = pS\left(1 - n\pi R_p^2\right)$$

$$F_1' = p'S\left(1 - n\pi R_p^2\right).$$

The net force to the right is

$$F_1 - F_1' = S\left(p - p'\right)\left(1 - n\pi R_p^2\right). \tag{5.61}$$

Forces F_2 and F_2' are exerted by solute molecules reflected from the pore region, such as molecule A in Fig. 5.17. These

are the ones that contribute to the osmotic pressure. The net force to the right is therefore the total pore area $Sn\pi R_p^2$ times the impermeant part of the osmotic pressure difference:

$$F_2 - F_2' = Sn\pi R_p^2 \left(\sigma\pi - \sigma\pi'\right). \tag{5.62}$$

Force F_3 is the viscous drag exerted on the walls of the pores by the water and permeant solute molecules flowing through them. To calculate it we recall that the viscous force per unit area is $-\eta\left(\partial v/\partial r\right)$. The velocity is $v = j_v$. Differentiating Eq. 5.43,we obtain

$$\frac{\partial j_v}{\partial r} = -\frac{1}{4\eta}\frac{\Delta\left(p - \sigma\pi\right)}{\Delta Z}2r.$$

The total force is η times this quantity evaluated at $r = R_p$, times total area of the cylindrical walls of all the pores, which is $(Sn)\left(2\pi R_p \Delta Z\right)$:

$$\begin{aligned} F_3 &= Sn2\pi R_p \Delta Z\eta\left(\frac{1}{4\eta}\frac{(p-p') - \sigma\left(\pi - \pi'\right)}{\Delta Z}2R_p\right) \\ &= Sn\pi R_p^2\left[(p-p') - \sigma\left(\pi - \pi'\right)\right]. \end{aligned} \tag{5.63}$$

The net force on the membrane is the sum of these forces:

$$F_1 - F_1' + F_2 - F_2' + F_3 = S(p - p'). \tag{5.64}$$

We see that the net force on the membrane is the total pressure difference times the total area of the membrane, regardless of the differences in osmotic pressure on each side. Both solute and solvent exert a force on the nonpore area of the membrane. The solute molecules at the membrane surface whose centers are within the area of a pore may be reflected or may enter the pore. If they are reflected, they contribute to the force when they strike the membrane at the edge of a pore. If they are not reflected, they enter the pore and contribute to the viscous drag on the membrane due to flow through the pore.

Symbols Used In Chapter 5

Symbol	Use	Units	First used page
a	Solute particle radius	m	125
a, a_{in}, a_{out}	Parameters	m^{-1}	128
c_1, c_2, c_1', c_2'	Solute concentration	(mole) m^{-3}	118
f	Temporary function		131
h	Thickness of fluid layer	m	127
i	Solute current through membrane	s^{-1}	137
i_s	Solute flow	s^{-1}	130
i_v	Volume flow	$\text{m}^3\ \text{s}^{-1}$	123
j_s, $\mathbf{j}_s$	Solute fluence rate in pore	$\text{m}^{-2}\ \text{s}^{-1}$	130
j_v, $\mathbf{j}_v$	Volume fluence rate in pore	m s^{-1}	129
k_B	Boltzmann's constant	J K^{-1}	118
n	Number of moles		118
n	Number of pores per unit area	m^{-2}	125
p_1, etc.	Pressure	Pa	118
p_d	Driving pressure	Pa	121
p	Total pressure	Pa	121
p_{dw}	Driving pressure of water	Pa	121
r	Radius in cylindrical coordinates	m	129
x, y, z	Position	m	127
x	$\Delta Z/\lambda$		133
z	Distance along pore	m	129
$\hat{\mathbf{z}}$	Unit vector in z direction		129
C, C_s, etc.	Particle concentration of the species indicated by the subscript	(particle) m^{-3}	118
D, D_{eff}	Diffusion constant	$\text{m}^2\ \text{s}^{-1}$	130
F	Force	N	134
G	Factor relating solvent drag and diffusion		132
J_s	Solute fluence rate through membrane	$\text{m}^{-2}\ \text{s}^{-1}$	125
J_v	Volume fluence rate through membrane	m s^{-1}	123
L_p	Hydraulic permeability	$\text{m s}^{-1}\ \text{Pa}^{-1}$	123
N_1, etc.	Number of molecules		118
N_A	Avogadro's number		118
R	Gas constant	$\text{J mol}^{-1}\ \text{K}^{-1}$	118
R_p	Pore radius	m	125
S	Surface area	m^2	123
T	Absolute temperature	K	118
V, V', V^*	Volume	m^3	118
X, Y	Distance	m	127
ΔZ	Pore length	m	125
η	Viscosity	Pa s	128
λ	Effective diffusion distance	m	131
μ	Chemical potential	J molecule^{-1}	120
ξ	a/R_p		133
π	Osmotic pressure	Pa	120
π	Geometric constant		
σ	Reflection coefficient		124
τ	Time constant	s	127
ω	Solute permeability	$\text{mol N}^{-1}\ \text{s}^{-1}$	125
ω_0	Solute permeability in an infinite medium	$\text{mol N}^{-1}\ \text{s}^{-1}$	134
ϕ	Angle in cylindrical coordinates		129
ϕ	C_s'/C_s		132
Γ	Radial dependence of solute concentration		130

Problems

Section 5.3

Problem 1. Use estimates of the size of a water molecule, the osmolarity of body fluids, and the thickness of a cell membrane to decide if Fig. 5.1 is drawn to scale or if

(a) the size of the water molecules has been exaggerated for clarity
(b) the ratio of the number of solute molecules to the number of solvent molecules has been exaggerated for clarity.

Problem 2. Perform the unit conversions to verify that 8.3145 J mol^{-1} K^{-1} is equivalent to 0.08206 atm l mol^{-1} K^{-1}.

Problem 3. The protein concentration in serum is made up of two main components: albumin (molecular weight 75,000) 4.5 g per 100 ml and globulin (molecular weight 170,000) 2.0 g per 100 ml. Calculate the osmotic pressure due to each constituent. (These results are inaccurate because of electrical effects.)

Problem 4. If the osmotic pressure in human blood is 7.7 atm at 37 °C, what is the solute concentration assuming that $\sigma = 1$? What would be the osmotic pressure at 4 °C?

Problem 5. Sometimes after trauma the brain becomes very swollen and distended with fluid, a condition known as cerebral edema. To reduce swelling, mannitol may be injected into the bloodstream. This reduces the driving force of water in the blood, and fluid flows from the brain into the blood. If 0.01 mol l^{-1} of mannitol is used, what will be the approximate osmotic pressure?

Section 5.4

Problem 6. When a person is given an intravenous fluid, the solute concentration in the fluid must be matched to the solute concentration in the blood to avoid problems arising from a change in the blood's osmotic pressure. One such fluid, called "isotonic saline," can be made by adding salt (NaCl) to distilled water. The osmolarity of the blood is about 0.3 osmole.

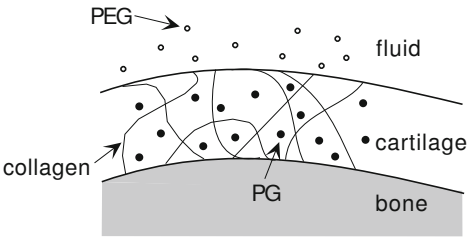
(a) How many grams of NaCl must be added to a liter of water to make isotonic saline? What fraction of the solution's mass is NaCl? (Hint: recall that NaCl dissolves into Na^+ and Cl^-, and both contribute to the osmotic pressure.)

(b) Repeat for dextrose, $C_6H_{12}O_6$, which does not dissociate.

Problem 7. An understanding of osmotic pressure is important in medicine. Consider the case reported by Steinmuller (1998) in the *New England Journal of Medicine*. A 5 % solution of albumin was needed to infuse into a patient with kidney disease (renal insufficiency). No 5 % solution was available, so the hospital pharmacy used 25 % albumin diluted 1:4 with pure water. Injection of the solution into the patient caused renal failure. The albumin in a 25 % albumin solution has an osmolarity of about 36 mosmol. Typically, such a solution also contains about 300 mosmol of other ions (see Problem 6).

(a) Calculate the osmolarity of the solution injected into the patient.
(b) Calculate the osmolarity of the solution if the pharmacy had properly used isotonic saline instead of pure water to perform the 1:4 dilution.

Problem 8. Articular cartilage covers the ends of bones in joints and allows the bones to move smoothly against each other. It contains a network of collagen fibers that can exert a mechanical tensile stress to resist tissue swelling, resulting in a pressure P_c within the cartilage. The collagen fibers do not withstand compression. The cartilage also contains proteoglycan molecules that cause tissue swelling because of their osmotic pressure, π_{PG}. One can determine P_c by placing the cartilage in a polyethylene glycol solution with osmotic pressure π_{PEG}, measuring π_{PG} and π_{PEG}, and using the relationship $P_c = \pi_{PG} - \pi_{PEG}$.



Typical data are

π_{PEG} (atm)	π_{PG} (atm)
0.0	4.0
2.5	5.5
5.0	7.0
7.5	8.5
10.0	10.0

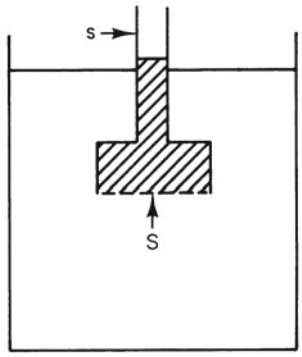
(a) What is the excess pressure P_c exerted by the collagen matrix under normal conditions ($\pi_{PEG} = 0$)?
(b) At what value of π_{PEG} does the collagen matrix exert no tensile stress (become "limp")?
(c) Plot P_c vs π_{PEG}. Find a linear equation that fits the data.
(d) If the collagen in an arthritic joint can only exert a pressure of 2 atm when $\pi_{PEG} = 0$, by how much will the tissue swell (by what percent will its volume change?)

In (b) and (d), assume that only the proteoglycans cause osmotic pressure and that their number does not change, but the tissue volume increases as the tissue swells with water. This problem is based on the work of Basser et al. (1998), but the data have been modified.

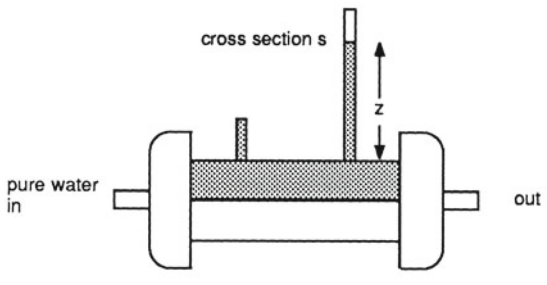
Section 5.5

Problem 9. Suppose that L_p is expressed in m^3 N^{-1} s^{-1} or m s^{-1} Pa^{-1}. Find conversion factors to express it in
(a) ml min^{-1} cm^{-2} $torr^{-1}$.
(b) ml s^{-1} cm^{-2} (in. water)$^{-1}$.
(c) ml s^{-1} cm^{-2} (lb $in.^{-2}$)$^{-1}$.

Problem 10. An ideal semipermeable membrane is set up as shown. The membrane surface area is S; the cross-sectional area of the manometer tube is s. At $t = 0$, the height of fluid in the manometer is zero. The density of fluid is ρ. Show that the fluid height rises to a final value with an exponential behavior. Find the final value and the time constant. Ignore dilution of the solute.



Problem 11. Consider the design of a lecture demonstration apparatus to show osmotic pressure that uses a commercially available filter as shown in the drawing. Assuming well-stirred fluid on both sides of the membrane and neglecting the change of solute concentration in the manometer tube as water flows in, one finds that height z increases to the equilibrium value exponentially, with a time constant obtained in the previous problem. What would be the time constant if one used a membrane for which $L_p = 1$ ml min^{-1} m^{-2} torr^{-1}, and the total membrane area is $S = 0.2$ m^2. Suppose that the inner radius of the manometer tube is 1 mm. (One could not use sucrose as a solute, because this particular membrane is permeable to molecules of molecular weight less than 50,000.)



Problem 12. A cell has variable volume V and fixed surface area S. The total hydrostatic pressure p is the same inside and outside the cell, and there is complete and instantaneous mixing. Initially the interior and exterior are both pure water. The initial volume of the cell is V_0. At $t = 0$ the exterior is bathed in a solution containing an impermeant solute of concentration C_0.

(a) Does the cell shrink to zero volume or expand to its maximum volume, which is a sphere of surface area S?

(b) Derive a differential equation for the volume change and integrate it to find how long it takes for the cell to reach zero or maximum volume.

Problem 13. A cell has variable volume V and fixed surface area S. The total hydrostatic pressure p is always the same both inside and outside the cell. There is complete and instantaneous mixing both inside and out. An impermeant solute has an initial concentration $C(0)$ both inside and outside. The initial cell volume is V_0. At $t = 0$ the exterior solute is removed.

(a) Does the cell shrink to zero volume or expand to its maximum volume, which is a sphere of surface area S?

(b) Derive a differential equation for $V(t)$ and find how long it takes for the cell to reach zero or maximum volume.

Section 5.6

Problem 14. Two membranes have permeabilities $\omega_1 RT$ and $\omega_2 RT$. Find the permeability of a two-layered membrane in terms of ω_1 and ω_2.

Problem 15. Solute is carried through a pipe by solvent drag. The radius of the pipe is b. The average flow along the pipe is $\bar{j}_v$ (independent of r because it has been averaged over r). Assume that within the pipe the concentration of solute is independent of radius and can be written as $C(z)$. The solute is carried along purely by solvent drag. Solute concentration outside the pipe is zero. Solute diffuses through the wall of the pipe, which has solute permeability ωRT. In terms of $\bar{j}_v$, b, and ωRT, obtain a differential equation for $C(z)$ and show that C decays exponentially along the pipe. Find the decay constant.

Section 5.7

Problem 16. A kidney machine has a membrane permeability $\omega RT = 0.5 \times 10^{-3}$ cm s^{-1}. If the membrane area is 1 m^2, the volume of body fluid is 40 l, and the volume of dialysate is effectively infinite, what is the time constant? How long will it take to reduce the BUN (blood urea nitrogen) concentration from 120 mg per 100 ml to 20 mg per 100 ml?

Problem 17. Find the pair of coupled differential equations for C and C' for a dialysis machine in which V' is not infinite.

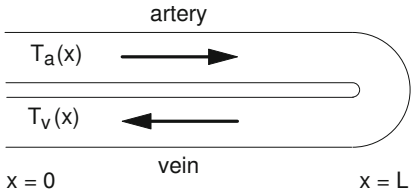
Section 5.8

Problem 18. In the countercurrent model (Eq. 5.25) the total current i through the membrane when its length is X is

$$i = \omega RTY \int_0^X [C_{\text{in}}(x) - C_{\text{out}}(x)]\, dx.$$

Solve this integral for the two cases given by Eqs. 5.26 and 5.27. Show that the current ratio in these two cases is 1.36 when $a = 1$ and $X = 2$.

Problem 19. The countercurrent model applies to the transport of heat as well as particles, with temperature taking the place of concentration. Consider a countercurrent heat exchanger, which represents the arrangement of blood vessels in the flipper of a whale (Schmidt-Nielsen 1972).



The temperatures of the arterial and venous blood are governed by equations similar to Eq. 5.27:

$$T_a = c_1 + (c_2 - c_1)ax,$$
$$T_v = c_2 + (c_2 - c_1)ax.$$

Assume that the arterial blood at $x = 0$ is at the warm temperature of the whale's body, T_w. The arterial blood at $x = L$ enters the capillaries at temperature $T_a(L)$ and is cooled to the temperature of the surrounding ocean water, T_c, by the time it enters the vein at $x = L$.

(a) Determine c_1 and c_2 in terms of T_w, T_c, a, and L.
(b) Plot $T_a(x)$ and $T_v(x)$ for $T_w = 37\,^{\circ}\text{C}$, $T_c = 7\,^{\circ}\text{C}$, $a = 1\ \text{mm}^{-1}$ and $L = 3\ \text{mm}$.
(c) The loss of heat from the body to the surroundings is proportional to $\Delta T = T_a(L) - T_c$. Find an expression for ΔT. What does ΔT reduce to if $aL \gg 1$? Interpret these results physically. To minimize heat loss to the ocean should aL be large or small?
(d) The energy the body must supply to heat the returning venous blood is proportional to $\Delta T' = T_w - T_v(0)$. Find an expression for $\Delta T'$.

Section 5.9

Problem 20. Derive Eqs. 5.44a and 5.44b.

Problem 21. Show that Eq. 5.51 gives $C(z) = \text{const}$ when $\lambda = 0$ (pure solvent drag) and gives $dC/dz = \text{const}$ when $\lambda \to \infty$ (pure diffusion).

Problem 22. Obtain expressions for J_s when $\lambda = 0$ and $\lambda \to \infty$.

Problem 23. Show that for very large pores when $\sigma = 0$ the parameter $x = \Delta Z/\lambda = J_v/\omega RT$ depends only on pore radius, solute particle radius, pressure difference and temperature, and not on viscosity, the number of pores per unit area, or the membrane thickness.

Problem 24. When $C_s' = 0$, what are the limiting values of $\overline{C_s}$ as $x \to 0$? As $x \to \infty$? As $x \to -\infty$?

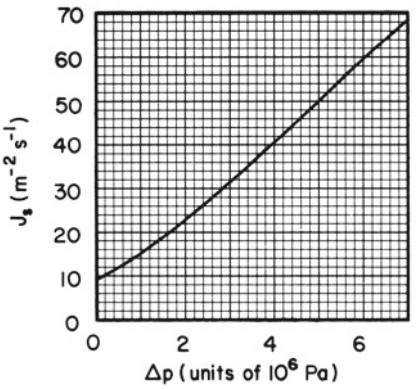
Problem 25. (a) Write J_s in terms of C_s, C_s', J_v and x.
(b) Specialize to the case $C_s' = 0$.

Problem 26. (a) Find the ratio $(1 - \sigma)\overline{C_s}J_v/[\omega RT(C_s - C_s')]$ in terms of x, C_s and C_s'.
(b) Specialize to the case $C_s' = 0$ and discuss limiting values for small and large x.

Problem 27. (a) Show that

$$J_s = \omega RT\left(C_s\frac{xe^x}{e^x - 1} - C_s'\frac{x}{e^x - 1}\right)$$

where $x = J_v(1 - \sigma)/\omega RT$.
(b) Discuss the special case $C_s' = 0$ in the limits $x \to 0$ and $x \to \infty$.
(c) From the data shown, estimate L_p and ωRT. The data are for the transport of radioactive water with a concentration of 10^{15} molecules m^{-3} on one side of the membrane and zero on the other.



Problem 28. Consider the following cases for transport of water through a membrane.

(a) Water flows by bulk flow through the membrane with $\Delta p = 0$. There is an impermeant solute ($\sigma = 1$) on the right with concentration C_{big} and zero concentration on the left. Find the particle fluence rate of water in terms of L_p.
(b) There is no volume flow through the membrane ($J_v = 0$). Some of the water molecules on the left are tagged with radioactive hydrogen (tritium). The concentration of tagged water molecules is C_s on the left and 0 on the right. Find the particle fluence rate of tagged water in terms of L_p and ωRT.
(c) There is volume flow, as in case (a), and there are also tagged water molecules on the left. Find the particle fluence rate of tagged water in terms of L_p and ωRT.
(d) Restate the answers in terms of the parameters of a collection of n pores per unit area of radius R_p and length ΔZ.
(e) Estimate the value of x for part (c) if $R_p = 10^{-8}$ m and $c_{\text{big}} = c_s = 0.1\ \text{mol}\ 1^{-1}$.

Problem 29. Construct diagrams analogous to Fig. 5.14a when the total pressure is the same on both sides and $\pi' = 0$ and (b) when $(p - \sigma\pi) < p'$ and $\pi' = 0$.

Problem 30. Consider the case of water permeability shown in Fig. 5.1c. Water and solute molecules move through the membrane in the same way. They "dissolve" from solution into the membrane. Assume that the concentration of water molecules just inside the membrane is proportional to the pressure just outside: $C = \alpha p$. The membrane has thickness ΔZ and the diffusion constant for water in the membrane material is D. Under steady-state conditions, derive an expression for L_p.

Problem 31. Consider the case in which solute moves along a tube by a combination of diffusion and solvent drag. Ignore radial diffusion within the tube, but assume that solute is moving out through the walls so that j_s is changing with position in the tube. In particular, the number of solute particles passing out through the wall in length dz in time dt is $CA2\pi R_p\, dz\, dt$, where A is related to the permeability of the wall. Consider a case in which C does not change with time, but depends only on position along the tube.

(a) Write down the conservation equation for an element of the tube and show that

$$-\frac{\partial j}{\partial z} - \frac{2AC}{R_p} = 0.$$

(b) Combine the results of part (a) with Eq. 5.45a and show that $C(z)$ must satisfy the differential equation

$$\frac{\partial^2 C}{\partial z^2} - \frac{\bar{j}_v f}{D}\frac{\partial C}{\partial z} - \frac{2A}{DR_p}C = 0.$$

Show that this equation will be satisfied if the concentration decreases exponentially along the tube as $C(z) = C_0 e^{-\alpha z}$, where

$$\alpha = \frac{\bar{j}_v f}{2D}\left[-1 + \left(1 + \frac{8AD}{R_p \bar{j}_v^2 f^2}\right)^{1/2}\right].$$

Problem 32. The volume of a water molecule is V_w and the volume of a solute molecule is V_s. Define a new quantity J_w that is the number of water molecules per unit area per second passing through the membrane. What is J_w in terms of J_v and J_s?

References

Agre P, King LS, Yasui M, Guggino WB, Ottersen OP, Fujiyoshi Y, Engel A, Nielsen S (2002) Aquaporin water channels: From atomic structure to clinical medicine. J Physiol 542:3–16

Basser PJ, Schneiderman R, Bank RA, Wachtel E, Maroudas A (1998) Mechanical properties of the collagen network in human articular cartilage as measured by osmotic stress technique. Arch Biochem Biophys 351:207–219

Bean CP (1969) Characterization of cellulose acetate membranes and ultrathin films for reverse osmosis. Research and Development Progress Report No. 465 to the U.S. Department of Interior, Office of Saline Water. Contract No. 14-01-001-1480. Washington, Superintendent of Documents, October 1969.

Bean CP (1972) The physics of porous membranes—neutral pores. In: Eisenman G (ed) Membranes, vol 1. Dekker, New York, pp 1–55

Beck RE, Schultz JS (1970) Hindered diffusion in microporous membranes with known pore geometry. Science 170:1302–1305

Calland CH (1972) Iatrogenic problems in end-stage renal failure. N Engl J Med 287:334–336

Cokelet GR (1972) The rheology of human blood. In: Fung YC et al (eds) Biomechanics, its foundations and objectives. Prentice Hall, Englewood Cliffs (especially p. 77, Fig. 4)

Durbin RP (1960) Osmotic flow of water across permeable cellulose membranes. J Gen Physiol 44:315–326

Fishman RA (1975) Brain edema. N Engl J Med 293:706–711

Gennari FJ, Kassirer JP (1974) Osmotic diuresis. N Engl J Med 291:714–720

Guyton AC, Taylor AE, Granger HJ (1975) Circulatory physiology II. Dynamics and control of the body fluids. Saunders, Philadelphia

Hall JE (2011) Guyton and Hall textbook of medical physiology, 12th edn. Saunders/Elsevier, Philadelphia

Hildebrand JH, Scott RL (1964) The solubility of nonelectrolytes, 3rd edn. Dover, New York

Katchalsky A, Curran PF (1965) Nonequilibrium thermodynamics in biophysics. Harvard University Press, Cambridge

Kramer EM, Myers DR (2012) Five popular misconceptions about osmotic pressure. Am J Phys 80:694–699

Levitt DG (1975) General continuum analysis of transport through pores. I. Proof of Onsager's reciprocity postulate for uniform pore. Biophys J 15:533–551

Lindner A, Charra B, Sherrard DJ, Scribner BH (1974) Accelerated atherosclerosis in prolonged maintenance hemodialysis. N Engl J Med 290:697–701

Lysaght MJ, Moran J (2006) Peritoneal dialysis equipment. In: Bronzino JD (ed) The biomedical engineering handbook, 3rd edn. Vol 3: Tissue engineering and artificial organs. CRC Press, Boca Raton, pp 68-1–68-13

Mavroidis C (2006) Artificial kidney. In Bronzino JD (ed) The biomedical engineering handbook, 3rd edn. Vol 3: Tissue engineering and artificial organs. CRC Press, Boca Raton, pp 67-1–67-24

Murata K, Mitsuoka K, Hirai T, Walz T, Agre P, Heymann JB, Engel A, Fujiyoshi Y (2000) Structural determinants of water permeation through aquaporin-1. Nature 407:599–605

Parisi M, Dorr RA, Ozu M, Toriano R (2007) From membrane pores to aquaporins: 50 years measuring water fluxes. J Biol Phys 33(5–6):331–343

Patton HD, Fuchs AF, Hille B, Scher AM, Steiner RF (eds) (1989) Textbook of physiology, 21st edn. Saunders, Philadelphia

Preston GM, Carroll TP, Guggino WB, Agre P (1992) Appearance of water channels in *Xenopus* oocytes expressing red cell CHIP28 protein. Science 256:385–387

Schmidt-Nielsen K (1972) How animals work. Cambridge University Press, Cambridge

Silverstein ME, Ford CA, Lysaght MJ, Henderson LW (1974). Treatment of severe fluid overload by ultrafiltration. N Engl J Med 291:747–751

Steinmuller DR (1998) A dangerous error in the dilution of 25 percent albumin. N Engl J Med 338:1226–1227

Thomas L (1974) Lives of a cell. Viking, New York

White RJ, Likavec MJ (1992) Current concepts: the diagnosis and initial management of head injury. N Engl J Med 327:1507–1511

Zeuthen T (2010) Water-transporting proteins. J Membrane Biol 234:57–73

Impulses in Nerve and Muscle Cells

6

A nerve cell conducts an electrochemical impulse because of changes that take place in the cell membrane. These allow movement of ions through the membrane, setting up currents that flow through the membrane and along the cell. Similar impulses travel along muscle cells before they contract. This chapter reviews the basic properties of electric fields and currents that are needed to understand the propagation of the nerve- or muscle-cell impulse.

Section 6.1 introduces the physiology of nerve conduction. The next eight sections develop the electrostatics and the physics of current flow needed to understand how the action potential propagates along the cell.

The next sections deal with the charge distribution on a resting cell membrane (Sect. 6.10) and the cable model of the axon (Sect. 6.11). If the membrane properties do not change as the voltage across the membrane changes, this leads to electrotonus or passive spread (Sect. 6.12). If the membrane properties do change, a signal can propagate without change of shape. Section 6.13 tells how Hodgkin and Huxley developed equations to describe the membrane changes, and Sects. 6.14 and 6.15 apply their results to the propagation of a nerve impulse. The chapter to this point forms an integrated story of conduction in an unmyelinated axon.

Propagation in a myelinated axon is described in Sect. 6.16. Section 6.17 examines the capacitance of a bilayer membrane that has layers with different properties. Section 6.18 shows how minor alterations in the membrane properties can transform the Hodgkin–Huxley model to one that displays repetitive electrical activity.

Section 6.19 illustrates how tabulated solutions to the electrical capacitance of conductors in different geometries can be used to solve diffusion problems with similar geometric configurations.

6.1 Physiology of Nerve and Muscle Cells

A nerve[1] consists of many parallel, independent signal paths, each of which is a nerve cell or fiber. Each cell has an input end (*dendrites*), a *cell body*, a long conducting portion or *axon*, and an output end. The axon portion of the cell can transmit a nerve pulse in either direction. The ends give the cell its unidirectional character. The input end can be a transducer (stretch receptor, temperature receptor, etc.) or a junction (*synapse*) with another cell. A threshold mechanism is built into the input end; when an input signal exceeding a certain level is received, the nerve fires an *impulse* or *action potential* of fixed size and duration that travels down the axon. There may be several inputs that can either aid or inhibit each other, depending on the nature of the synapses.

Muscle cells are also long and cylindrical. An electrical impulse travels along a muscle cell to initiate its contraction. This chapter concentrates on the propagation of the action potential in a nerve cell, but the discussion can be regarded as a model for what happens in muscle cells as well.

The axon transmits the impulse without change of shape. The axon can be more than a meter in length, extending from the brain to a synapse low in the spinal cord or from the spinal cord to a finger or toe. Bundles of axons constitute a nerve. The output end branches out in fine nerve endings, which appear to be separated by a gap from the next nerve or muscle cell that they drive.

The long cylindrical axon has properties that are in some ways similar to those of an electric cable. Its diameter may range from less than one micron (1 μm) to as much as 1 mm

[1] A good discussion of the properties of nerves and the Hodgkin–Huxley experiments is found in Katz (1966). More modern descriptions of nerves and nerve conduction are found in many books, such as Patton et al. (1989) or Nicholls et al. (2011).

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_6, © Springer International Publishing Switzerland 2015

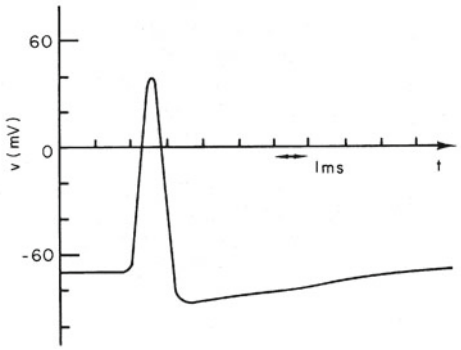


Fig. 6.1 A typical nerve impulse or action potential, plotted as a function of time

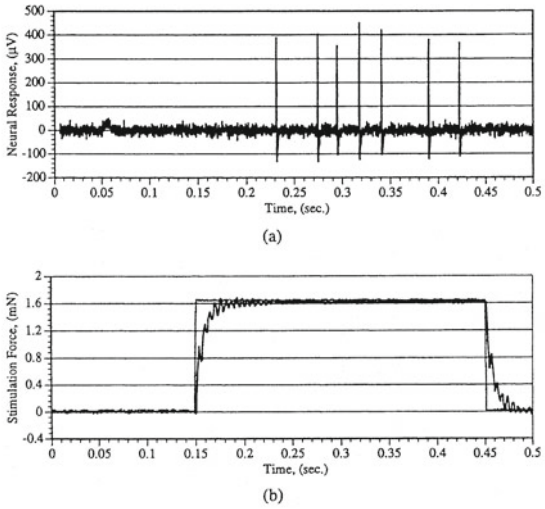


Fig. 6.2 The response of a mechanical receptor in the cornea to an applied force. **a** The impulses recorded on the surface of the nerve bundle. **b** The applied force. Impulses occur while the force is applied. (Source: Kane et al. 1995) © 1995 IEEE. Reprinted by permission

for the giant axon of a squid; in humans the upper limit is about 20 μm. Pulses travel along it with speeds ranging from 0.6 to 100 m s^{-1}, depending, among other things, on the diameter of the axon. The axon core may be surrounded by either a membrane (for an *unmyelinated* fiber) or a much thicker sheath of fatty material (*myelin*) that is wound on like tape. A myelinated fiber has its sheath interrupted at intervals and replaced by a short segment of membrane similar to that on an unmyelinated fiber. These interruptions are called *nodes of Ranvier*. A typical human nerve might contain twice as many unmyelinated fibers as myelinated. We will see in Sect. 6.16 that the myelin gives a faster impulse conduction speed for a given axon radius. Myelinated fibers conduct information where speed is important, such as motor information; unmyelinated fibers conduct information such as temperature, for which speed is not important. A typical unmyelinated axon might have a radius of 0.7 μm with a membrane thickness of 5–10 nm. Myelinated fibers have a radius of up to 10 μm, with nodes spaced every 1–2 mm. We will find later that the spacing of the nodes is about 140 times the inner radius of the fiber, a fact that is quite important in the relationship between conduction speed and fiber radius.

A microelectrode inserted inside a resting axon records an electrical potential that is about 70 mV less than outside the cell. (We will define electrical potential difference in Sect. 6.4.) A nerve impulse or action potential or spike in an unmyelinated axon is shown as a function of time in Fig. 6.1. As the impulse passes by the electrode, the potential rises in a millisecond or less to about +40 mV. The potential then falls to about −90 mV and then recovers slowly to its resting value of −70 mV. The membrane is said to *depolarize* and then *repolarize*.

The history of recording the action potential has been described by Geddes (2000). The propagation speed of the action potential was first measured by Helmholtz around 1850. The measurement technology steadily improved, culminating in the use of a microelectrode inserted by Hodgkin and Huxley (1939) into the cut end of the giant axon of the squid to record the action potential directly.

The information sent along a nerve fiber is coded in the repetition rate of these pulses, all of which are the same shape. Figure 6.2 shows the response of a low-threshold mechanoreceptor in the cornea to a mechanical stimulus. The heavy curve in the bottom panel shows the applied force, and the upper panel shows the impulses.

Comparison of the *intracellular fluid* or *axoplasm* with the *extracellular* fluid surrounding each axon shows an excess of potassium and a deficit of sodium and chloride ions within the cell, as shown in Fig. 6.3. The regenerative action

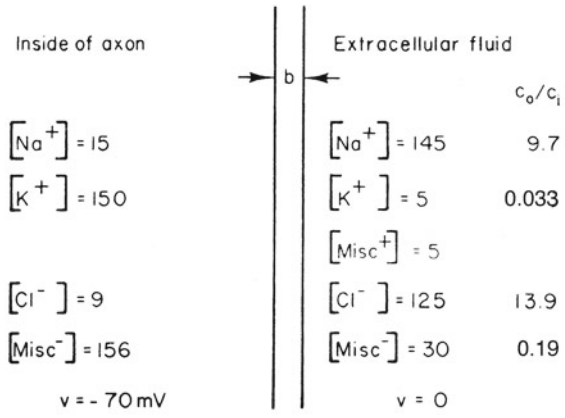


Fig. 6.3 Ion concentrations in a typical mammalian nerve and in the extracellular fluid surrounding the nerve. Concentrations are in mmol l^{-1}; c_o/c_i is the concentration ratio. The membrane thickness is b

that produces the sudden changes of membrane potential is caused by changing permeability of the membrane to ions—primarily sodium and potassium—as discussed in Sects. 6.13 and 6.14. On the molecular level these permeability changes are due to the opening and closing of selected *ion channels*, discussed in more detail in Chap. 9.

The axon can be removed from the rest of the cell and it will still conduct nerve impulses. The speed and shape of the action potential depend on the membrane and the concentration of ions inside and outside the cell. The axoplasm has been squeezed out of squid giant axons and replaced by an electrolyte solution without altering appreciably the propagation of the impulses—for a while, until the ion concentrations change significantly. The axoplasm does contain chemicals essential to the long-term metabolic requirements of the cell and to maintaining the ion concentrations.

At the end of a nerve cell the signal passes to another nerve cell or to a muscle cell across a *synapse* or junction. A few synapses in mammals are electrical; most are chemical (Nolte 2002, p. 193; Hall 2011, Chap. 45). In electrical synapses, channels connect the interior of one cell with the next. In the chemical case a neurotransmitter chemical is secreted by the first cell. It crosses the synaptic cleft (about 50 nm) and activates or inhibits the next cell.

At the neuromuscular junction the transmitter is *acetylcholine* (ACh). ACh increases the permeability of nearby muscle to sodium, which then enters and depolarizes the muscle membrane. The process is quantized.[2] Packets of acetylcholine of definite size are liberated (Katz 1966, Chap. 9; Patton et al. 1989, Chap. 6).

There are a number of neurotransmitters in the central nervous system. *Glutamate* is a common excitatory neurotransmitter in the central nervous system. It increases the membrane permeability to sodium ions, which enhances depolarization. *Glycine*, on the other hand, is an inhibitory neurotransmitter. It causes the interior potential to become more negative (hyperpolarized) and firing is inhibited. A number of other chemical mediators such as norepinephrine, epinephrine, dopamine, serotonin, histamine, aspartate, and gamma-aminobutyric acid, are also found in the nervous system (Hall 2011, Chap. 45).

If the potential becomes high enough (that is, more positive or less negative), the regenerative action of the membrane takes over, and the cell initiates an impulse. If the input end of the cell acts as a transducer, the interior potential rises when the cell is stimulated. If the input is from another nerve, the signal may cause the potential to increase by a subthreshold amount so that two or more stimuli must be received simultaneously to cause firing, or it may decrease the potential and inhibit stimulation by another nerve at the synapse.

[2] See Prob. 3 in Appendix J.

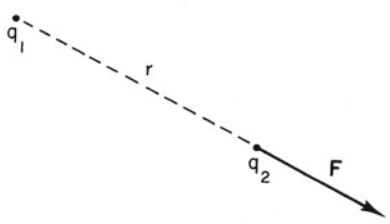


Fig. 6.4 Force **F** is exerted by charge q_1 on charge q_2. It points along a line between them. An equal and opposite force $-\mathbf{F}$ is exerted by q_2 on q_1

This makes possible the logic network that comprises the central nervous system.

6.2 Coulomb's Law, Superposition, and the Electric Field

Coulomb's law relates the electrical force between two charged objects. If two objects have electrical charge q_1 and q_2, respectively, and are separated by a distance r, then there is a force between them, the magnitude of which is given by

$$|\mathbf{F}| = \left(\frac{1}{4\pi\epsilon_0}\right)\frac{q_1 q_2}{r^2}. \tag{6.1}$$

When the charge is measured in coulombs (C), F in newtons (N), and r in meters (m), the constant has the value

$$\frac{1}{4\pi\epsilon_0} \approx 9 \times 10^9 \text{ N m}^2 \text{ C}^{-2} \tag{6.2}$$

to an accuracy of 0.1 %. The direction of the force is along the line between the two charges as shown in Fig. 6.4. If the charges are both positive or both negative, the force is repulsive, which is consistent with assigning a positive sign to **F**. If one is positive and the other negative, then the force is attractive, and **F** is assigned a minus sign. Force **F** is exerted by charge q_1 on charge q_2. The force exerted by q_2 on q_1 has the same magnitude but points in the opposite direction. The forces on both charges act to separate them if they have the same sign and to attract them if the signs are opposite.

If two or more charges exert a force on the particular charge being considered, the total force is found by applying Coulomb's law to each charge (paired with the one on which we want to find the force) and adding the vector forces that are so calculated. An example of this is shown in Fig. 6.5. Charges q_1, q_2, and q_3 are $+1.0 \times 10^{-6}$, -2.0×10^{-6}, and $+3.0 \times 10^{-6}$ C, respectively. The magnitude of the force that q_1 exerts on q_3 is

$$F_{1 \text{ on } 3} = \frac{(9 \times 10^9)(1 \times 10^{-6})(3 \times 10^{-6})}{(2 \times 10^{-2})^2} = 67.5 \text{ N}.$$

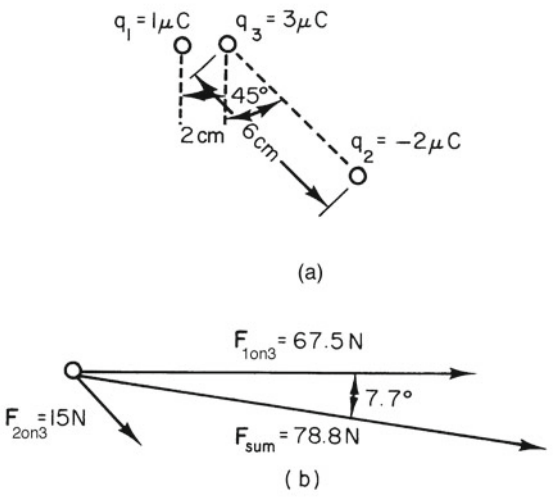


Fig. 6.5 An example of applying Coulomb's law and adding forces on q_3 due to charges q_1 and q_2. **a** The arrangement of charges. **b** The forces on q_3

Similarly, the force exerted by q_2 on q_3 is

$$F_{2\text{ on }3} = \frac{(9\times 10^9)(-2\times 10^{-6})(3\times 10^{-6})}{(6\times 10^{-2})^2} = -15\text{ N}.$$

The minus sign means that the force is attractive, that is, toward q_2. The two forces are shown in Fig. 6.5b, along with their vector sum. The sum can be found by components as in Chap. 1. The result is 78.8 N at an angle of 7.7° clockwise from the direction of $\mathbf{F}_{1\text{ on }3}$.

If a collection of charges causes a force to act on some other charge (a *test charge*) located somewhere in space, we say that the collection of charges produces an *electric field* at that point in space. One can think, for example, of charge q_1 producing an electric field vector, of magnitude

$$|\mathbf{E}_1| = \frac{1}{4\pi\epsilon_0}\frac{q_1}{r^2} \tag{6.3}$$

pointing radially away from q_1 (if q_1 is positive) or radially toward q_1 (if q_1 is negative). The force on test charge q_2 placed at the observing point is then

$$\mathbf{F} = q_2\mathbf{E}_1. \tag{6.4}$$

6.3 Gauss's Law

It is possible to derive a theorem about the electric field from a collection of charges, known as *Gauss's law*. Rather than derive it from Coulomb's law, we will state it and show that Coulomb's law can be derived from it. Then we will consider some examples of its use.

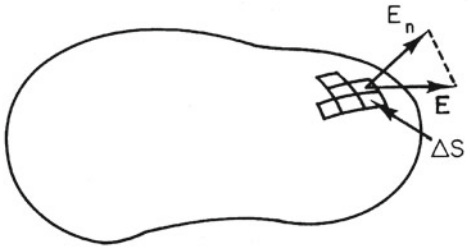


Fig. 6.6 Calculating the integral of the normal component of **E** through a surface

Divide up *any* closed surface into elements of surface area, such as ΔS in Fig. 6.6. For each element ΔS, calculate the component of **E** normal to the surface, E_n, and multiply it by the magnitude of the surface area ΔS. Add these quantities for the entire closed surface, calling them positive if the normal component of **E** points outward and negative if **E** points inward. Gauss's law says that the resulting sum is equal to the total charge inside the surface, divided by ϵ_0. In symbols,[3]

$$\iint E_n\, dS = \frac{q}{\epsilon_0} = \frac{4\pi q}{4\pi\epsilon_0}. \tag{6.5}$$

This surface integral is exactly the same as the flux of the continuity equation, Eq. 4.4. It is in fact called the electric field flux.[4] The surface is called a *Gaussian surface*.

While Gauss's law is always *true*, it is not always *useful*. It is helpful only in cases where **E** is constant over the entire surface of integration, or when the surface can be divided into smaller surfaces, on each of which E_n can be argued to be constant or zero. One of the few cases in which Gauss's law is useful to calculate **E** is the case of a point charge, and another is related to the cell membrane. In each case, the symmetry of the problem allows the surface of integration to be specified so that E_n is either constant or zero.

The first example is a point charge in empty space. Since such a charge has no preferred orientation (it is a point), and since there is nothing else around to specify a preferred direction in space, the electric field must point radially toward or away from the charge and must depend only on distance from the charge. Therefore, if the Gaussian surface is a sphere centered on the charge, E_n is the same everywhere on the sphere.

[3] Some books use one integral sign in this equation and others use two. Strictly speaking the integral over a surface is a two-dimensional integral.

[4] Additional discussion and examples can be found in Schey (2004).

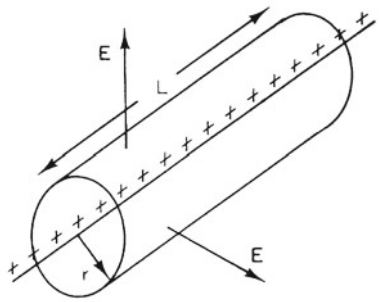


Fig. 6.7 Gauss's law is used to calculate the electric field from an infinite line of charge. The Gaussian surface is a segment of a cylinder concentric with the line of charge

It can be taken outside the integral in Eq. 6.5 to give

$$\iint E_n \, dS = E \iint dS.$$

The integral of dS over the entire surface of the sphere is just the surface area of the sphere, $4\pi r^2$ (see Appendix L). Gauss's law gives

$$4\pi r^2 E = \frac{q}{\epsilon_0}$$

or

$$E = \frac{q}{4\pi\epsilon_0 r^2}.$$

Gauss's law implies Coulomb's law for the case of a point charge.

If the charge in this problem is not a point charge, nothing changes in the argument as long as the charge distribution is spherically symmetric. The electric field at a distance r from the center of the distribution is the same as if all the charge within the sphere of radius r were located at the center of the sphere.

Next, consider a problem with cylindrical symmetry rather than spherical symmetry. An example is an infinitely long line of charge. For a segment of the line of charge of length L, the amount of charge is proportional to L, $q = \lambda L$, where λ is the linear charge density in units C m^{-1}. Symmetry shows that **E** must point radially outward (or inward) and be perpendicular to the line. Therefore if the Gaussian surface is a cylinder of length L and radius r, the axis of which is the line of charge, one can argue that on the end caps $E_n = 0$, while on the wraparound surface of the cylinder $E_n = |\mathbf{E}|$. This is shown in Fig. 6.7. The total integral is therefore the integral for the wraparound surface, which is $E \iint dS$. The surface area of the cylinder is its circumference $(2\pi r)$ times its length (L). Therefore Gauss's law becomes $2\pi r L \, E = \lambda L/\epsilon_0$, or

$$E = \frac{\lambda}{2\pi\epsilon_0 r}. \tag{6.6}$$

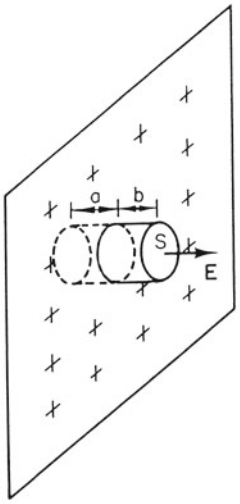


Fig. 6.8 A portion of an infinite sheet of charge and the appropriate Gaussian surface

Since the constant $1/4\pi\epsilon_0$ is so easily remembered, it is convenient to write this as

$$E = \frac{1}{4\pi\epsilon_0}\frac{2\lambda}{r}. \tag{6.7}$$

Consider next an infinite sheet of charge, with charge per unit area σ C m^{-2}. The symmetry of the situation requires that **E** be perpendicular to the sheet. To see why, suppose that **E** is not perpendicular to the sheet. I stand on the sheet looking in such a direction that **E** points diagonally off to my left. If I turn around in place, I see **E** pointing diagonally off to my right. Since the charge per unit area is constant and extends an infinite distance in every direction, the charge distribution looks exactly the same as it did before I turned around. The only way to resolve this contradiction (that **E** changed direction while the charge distribution did not change) is to have **E** perpendicular to the sheet.

The Gaussian surface can be a cylinder with end caps of area S and sides perpendicular to the sheet. Let the end caps be a distance a from the charge sheet on one side and b from the charge sheet on the other, as in Fig. 6.8. Since there is no component of **E** across the sides of the cylinder, changing b or a does not change the total flux through the surface. Since the charge inside the volume does not change, E must be independent of distance from the sheet of charge. (This is true only because the charge sheet is infinite.) By symmetry, the flux through each end cap is the same, as may be seen from the cross section of the surface in Fig. 6.9. The total flux is therefore $2ES$, while the charge within the volume is σS. Therefore, Gauss's law gives

$$E = \frac{\sigma}{2\epsilon_0} = \frac{1}{4\pi\epsilon_0} 2\pi\sigma. \tag{6.8}$$

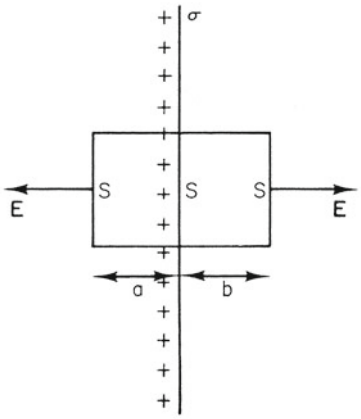


Fig. 6.9 A side view of the Gaussian surface in Fig. 6.8

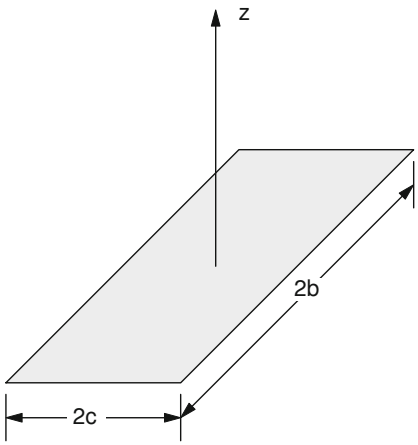


Fig. 6.10 A rectangular sheet of charge. The electric field along the z axis is shown in Fig. 6.11 for $2b = 200\,\text{m}$ and $2c = 2\,\text{m}$

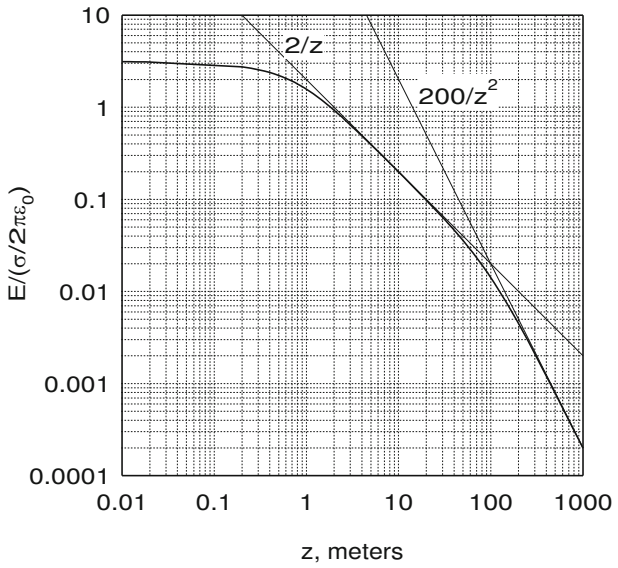


Fig. 6.11 A log-log plot of the electric field from a sheet of charge of width 2 m and length 200 m, measured along the perpendicular bisector of the sheet (Fig. 6.10). Much closer than 1 m, the field is constant. Around 10 m the field is proportional to $1/r$, the field from a line charge. Farther away than 100 m the field is proportional to $1/r^2$, the field from a point charge

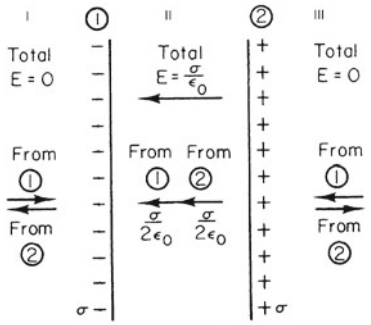


Fig. 6.12 The electric field due to two infinite sheets of charge of opposite sign

There is, of course, something quite unreal about a sheet of charge extending to infinity. However, it is a good approximation for an observation point close to a finite sheet of charge. If the sheet is limited in extent and the observation point is far away, the distance to all parts of the sheet from the observation point is nearly the same, and the charge sheet may be regarded as a point charge. If one considers a rectangular sheet of charge lying in the xy plane of width $2c$ and length $2b$, as shown in Fig. 6.10, it is possible to calculate exactly the **E** field along the z axis. By symmetry, the field points along the z axis. The surface charge density is σ. The distance is $r = \left(x^2 + y^2 + z^2\right)^{1/2}$. The component of **E** parallel to the z axis is $E\cos\theta = Ez/r$. Therefore, if the charge in element of area $dx\,dy$ is $\sigma\,dx\,dy$, the field is

$$E = \frac{\sigma z}{4\pi\epsilon_0}\int_{-b}^{b}\int_{-c}^{c}(x^2 + y^2 + z^2)^{-3/2}\,dxdy. \qquad (6.9)$$

This integral can be evaluated (see Problem 7). The result is

$$E = \frac{4\sigma}{4\pi\epsilon_0}\tan^{-1}\left(\frac{bc}{z\sqrt{c^2 + b^2 + z^2}}\right). \qquad (6.10)$$

This is plotted in Fig. 6.11 for $c = 1\,\text{m}$, $b = 100\,\text{m}$. Close to the sheet ($z \ll 1$) the field is constant, as it is for an infinite sheet of charge. Far away compared to 1 m but close compared to 100 m, the field is proportional to $1/r$ as with a line charge. Far away compared to 100 m, the field is proportional to $1/r^2$, as from a point charge.

As a final example, consider two infinite sheets of charge, one with density $-\sigma$ and the other with density $+\sigma$, as shown in Fig. 6.12. This can be solved by using the result for a single

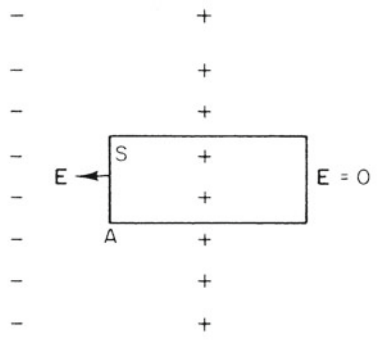


Fig. 6.13 A Gaussian surface to determine the electric field between two sheets of charge

sheet of charge, Eq. 6.8, and the principle of superposition. Consider first the region I of Fig. 6.12. There, the negative charge will give an **E** field that has magnitude $\sigma/2\epsilon_0$ and points toward the right, while the positive sheet of charge will give an **E** field of $\sigma/2\epsilon_0$ pointing to the left. The total **E** field in region I is zero. A similar argument can be made in region III with the field of the negative charge pointing left and that of the positive charge pointing to the right. Again the sum is zero. In region II, however, the two **E** fields point in the same direction, and the total field is

$$E = \frac{\sigma}{\epsilon_0} = \frac{1}{4\pi\epsilon_0} 4\pi\sigma. \tag{6.11}$$

Notice the factor of 2 difference between Eqs. 6.8 and 6.11. Another way to explain the difference is that there is no **E** in region III, so that a Gaussian surface can be constructed as shown in Fig. 6.13. Then the flux is zero through every surface except cap A. The charge within the volume is σS, while the flux through cap A is ES. Therefore, $E = \sigma/\epsilon_0$.

Within a cell membrane of 6 nm thickness surrounding a cell of radius 5 μm or 5 000 nm, the electric field can be calculated by making the approximation that the sheets of charge are infinite. Suppose that the electric field within the membrane is 1.17×10^7 N C^{-1}. (We will learn how to determine this value later.) From Eq. 6.11 the charge density is

$$\sigma = \frac{E}{4\pi(1/4\pi\epsilon_0)} = \frac{1.17 \times 10^7}{4\pi(9 \times 10^9)} = 1.03 \times 10^{-4} \text{ C m}^{-2}.$$

This tells us something about the makeup of the cell. The membrane is in contact with atoms, each of which has a diameter of about 10^{-10} m. Therefore there are approximately 10^{20} atoms (in water molecules, as ions, etc.) in contact with 1 m^2 of the membrane surface. Suppose that the excess charge that causes the electric field in the membrane resides in these atoms and that each atom is either neutral or a monovalent ion. The number of atoms in the square meter which are charged is

$$\frac{1.03 \times 10^{-4} \text{ C m}^{-2}}{1.6 \times 10^{-19} \text{ C atom}^{-1}} = 6.4 \times 10^{14} \text{ atoms m}^{-2}.$$

The fraction of atoms that are charged is $(6.4 \times 10^{14})/(10^{20}) = 6.4 \times 10^{-6}$. Roughly 1 in every 10^5 atoms in contact with the membrane carries an unneutralized charge. (This result is modified by partial neutralization of this external charge by charge movement within the membrane. See Eq. 6.35 and the footnote on p. 157.)

6.4 Potential Difference

It is often convenient to talk about the *electrical potential difference*, or *voltage difference,* instead of the electric field. The potential is related to the difference in energy of a charge when it is at different points in space. Suppose that an electric field **E** of magnitude E_x points along the x axis. A positive charge is located at point A. A force $\mathbf{F}_{\text{ext}}$ must be applied to the charge by something besides the electric field, or else the charge will be accelerated to the right by the force qE_x. The charge can be moved slowly to the right at a constant speed so that its kinetic energy remains fixed, if the external force is always to the left and its magnitude is adjusted so that $F_{\text{ext}} = -qE_x$.

This situation is shown in Fig. 6.14. The external force does work on the charge. One can either say that the total work done on the charge by both forces is equal to zero, or one can ignore the work done by the electric force and invent the idea of potential energy—energy of position—due to the electric field. The increase in potential energy[5] as the charge moves a distance dx is

$$dU = F_{\text{ext}}\, dx = -qE_x\, dx.$$

If E_x varies with position, the total change in potential energy when the particle is moved without acceleration from A to B is given by

$$\Delta U = U(B) - U(A) = -q \int_A^B E_x(x)\, dx. \tag{6.12}$$

For example, in a constant electric field of 1.4×10^7 N C^{-1}, a particle with charge $q = 1.6 \times 10^{-19}$ C experiences an electric force equal to 2.24×10^{-12} N. If it is moved 5 nm along the x axis, the electric force does 1.12×10^{-20} J of work on it, increasing its kinetic energy. To prevent this

[5] In earlier chapters the potential energy was called E_p, and the total energy was called U. For the next few pages U will be used for potential energy, to avoid confusion with a component of the electric field.

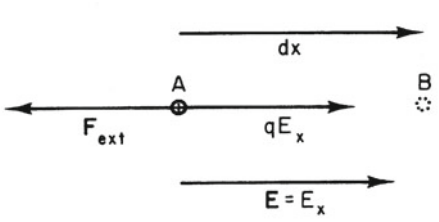


Fig. 6.14 A charge q is moved from A to B, a distance dx in the x direction. External force $\mathbf{F}_{\text{ext}}$ keeps the charge from being accelerated

increase in kinetic energy, F_{ext} must be applied. The external force does work -1.12×10^{-20} J. We can either say that the total work done by both forces is zero or we can ignore the electrical force and say that the external force changed the potential energy of the particle by -1.12×10^{-20} J as the particle moved from A to B.

If the displacement of the particle is perpendicular to the direction of the electric field, it is also perpendicular to the direction of $\mathbf{F}_{\text{ext}}$. Therefore neither force does work on the particle and the potential energy is unchanged. This fact can be used to prove (Serway and Jewett 2013, p. 567) that in three dimensions,

$$\begin{aligned} \Delta U &= U(B) - U(A) \qquad (6.13) \\ &= -q\left(\int_A^B E_x\,dx + \int_A^B E_y\,dy + \int_A^B E_z\,dz\right). \end{aligned}$$

Using the notation of a "dot" or scalar product of two vectors (Sect. 1.9), this can also be written as a line integral along any path from A to B:

$$\Delta U = -q\int_A^B \mathbf{E}\cdot d\mathbf{r}. \qquad (6.14)$$

It is easier to evaluate the integral along some paths than along others.

The potential energy difference is measured in joules. It is always proportional to the charge of the particle that is moved in the electric field. It is convenient to define the *potential difference* Δv to be the potential energy difference per unit charge. When the energy difference is in joules and the charge is in coulombs, the ratio is J C^{-1}, which is called a volt (V):

$$\Delta v \quad (\text{V}) = \frac{\Delta U \quad (\text{J})}{q \quad (\text{C})}. \qquad (6.15)$$

To move a charge of +3 C from point A to point B where the potential is 5 V higher requires that 15 J of work be done on the charge. If the charge is then allowed to move back to point A under the influence of only the electric field, its kinetic energy increases by 15 J as the potential energy decreases by the same amount.

This definition of potential, when combined with the definition of potential energy, Eq. 6.12, gives

$$\Delta v = -\int_A^B E_x\,dx$$

or

$$E_x = -\frac{\partial v}{\partial x}. \qquad (6.16a)$$

That is, the component of the electric field in any direction is the negative of the rate of change of potential in that direction. The units of $\mathbf{E}$ were seen earlier to be N C^{-1} (from $\mathbf{F} = q\mathbf{E}$). Equation 6.16 shows that the units of $\mathbf{E}$ are also V m^{-1}. In three dimensions this relationship becomes

$$\mathbf{E} = -\operatorname{grad} v = -\nabla v, \qquad (6.16b)$$

where grad is the gradient operation defined in Eq. 4.19.

Notice that only differences in potential energy and differences in potential (or colloquially, differences in voltage) are meaningful. We can speak of the potential at a point only if we have previously agreed that the potential at some other point will be called zero. Then we are really speaking of the difference of potential between the reference point and the point in question.

In many cases, it is customary to define the potential to be zero at infinity. Then the potential at point B is

$$v(B) = -\int_\infty^B E_x\,dx.$$

If you try to apply this equation to the infinite line and sheet of charge, you will discover that it does not work. The reason is that you cannot get infinitely far away from a charge distribution that extends to infinity.

6.5 Conductors

In some substances, such as metals or liquids containing ions, electric charges are free to move. When all motion of these charges has ceased and static equilibrium exists, there is no net charge within the conductor. To see why there is not, consider a small volume within the conductor. If there were an electric field within that region, the charges there would experience an electric force. Since they are free to move, this force would accelerate them. This force will vanish only when the electric field within the conductor is zero. Therefore, in the static case the electric field within a conductor is zero. Now apply Gauss's law to a small volume within the conductor. Since the electric field in the conductor is zero everywhere, the flux through the Gaussian surface is zero, and the net charge within the volume is zero.

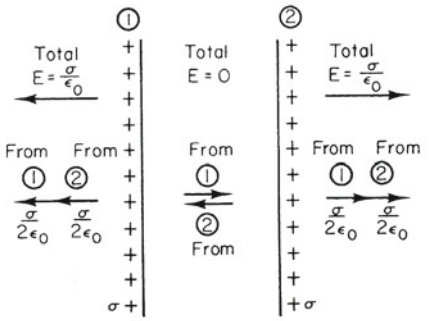


Fig. 6.15 The electric field in and around an infinite plane conductor carrying a charge on each surface

At the surface of the conductor, there may well be charge that gives rise to electric fields outside the conductor. Consider, for example, an infinite sheet of metal that has positive charge on it. The positive charge will distribute itself as shown in Fig. 6.15, and either superposition or Gauss's law may be used to show that the electric field outside the conductor is σ/ϵ_0.

Because the electric field is zero throughout a conductor in equilibrium, no work is required to move a charge from one point to another. All parts of the conductor are at the same potential. This statement is true only if the charges are not moving. We will see later that if they are (that is, if a current is flowing), then the electric field in the conductor is not zero and the potential in the conductor is not the same everywhere.

6.6 Capacitance

Suppose that two conductors are fixed in space, with charge $+Q$ on one and $-Q$ on the other. The potential difference v between the conductors is proportional to Q. The proportionality constant depends on the geometrical arrangement of the conductors. When the proportionality is written as

$$Q = Cv \tag{6.17}$$

the proportionality constant C is called the *capacitance*. The units of capacitance are C V^{-1} or farads (F).

As an example of capacitance, consider two parallel conducting plates side by side. Let the area of each be S and the separation be b. The charge layers of Fig. 6.13 might be charge on the inner surface of each conductor. The total charge on each plate has magnitude σS. The electric field between the plates is σ/ϵ_0 and the potential difference is $v = Eb = \sigma b/\epsilon_0$. (Note that the potential difference is proportional to the charge per unit area.) The capacitance is

$$C = \frac{Q}{v} = \frac{\sigma S\epsilon_0}{\sigma b} = \frac{\epsilon_0 S}{b}. \tag{6.18}$$

If the plates are separated further with a fixed charge on them, the potential difference increases and the capacitance is decreased. Increasing the area and charge of the plates with fixed σ and fixed b increases Q and C but not v.

6.7 Dielectrics

Charges rearrange themselves so that there is no static electric field within a conductor. In a dielectric, charges are not free to move far enough to completely cancel the effect of any external electric field, but they can move far enough to cause a partial cancellation.[6]

The partial neutralization of the external electric field can be understood from the following model. Consider a dielectric in the absence of external fields. The electron distribution of each atom is centered on the nucleus so that there is no electric field (at least when we average over a region containing many atoms). This is shown schematically in Fig. 6.16a, in which each + sign represents a nucleus and each circle represents a distribution of negative charge in an atom. Figure 6.16b shows some external charges producing an electric field. If the dielectric is introduced in the space where this electric field exists, the negative electron clouds are shifted with respect to the nuclei, as shown in Fig. 6.16b. The result is a polarization electric field $\mathbf{E}_p$, which is in the opposite direction to the external electric field. The total field within the dielectric is the vector sum of these two fields:[7]

$$\mathbf{E}_{\text{tot}} = \mathbf{E}_{\text{ext}} + \mathbf{E}_p. \tag{6.19}$$

In simple materials all three vectors are parallel and E_p is proportional to E_{tot}. Then we can define the *electric susceptibility* χ by the equation

$$E_p = -\chi E_{\text{tot}}.$$

This can be combined with the previous equation to get

$$E_p = -\frac{\chi}{1+\chi} E_{\text{ext}}.$$

[6] In some materials an electric field applied along one direction can cause charge displacement in a different direction. This book deals only with cases in which the induced electric field is parallel to the applied electric field.

[7] In most textbooks, it is customary to define the polarization by $\mathbf{P} = -\mathbf{E}_p$ or $\mathbf{P} = -\epsilon_0\mathbf{E}_p$. We have not done that in order to make the phenomenon easier to understand.

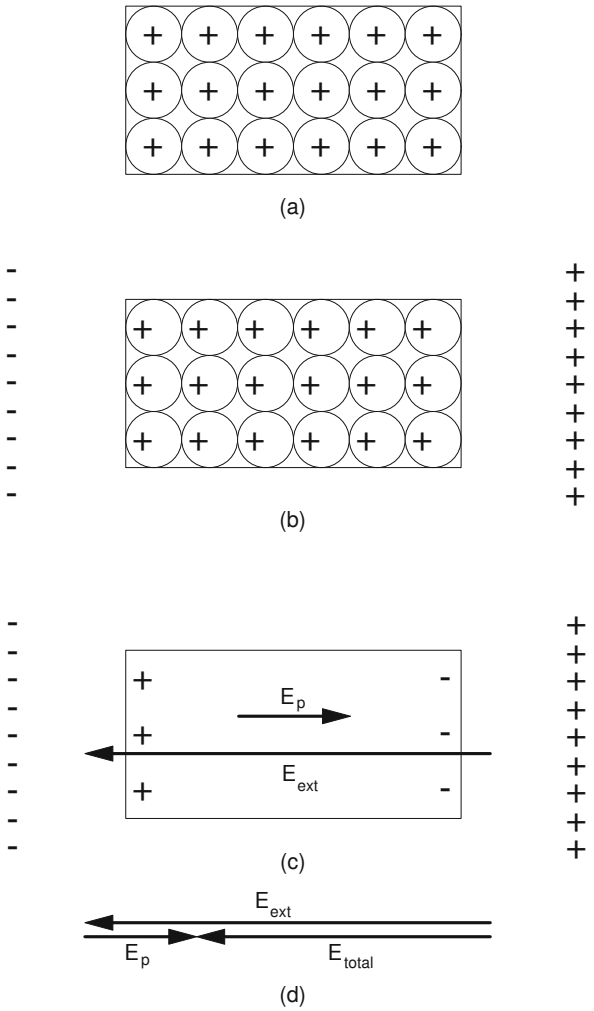


Fig. 6.16 The polarization of a dielectric by an external electric field. **a** Atoms in the absence of an external field. **b** An external electric field causes a shift of each electron cloud relative to the positively charged nucleus. **c** There is a net buildup of positive charge at the left edge of the dielectric and of negative charge at the right edge. **d** The total electric field within the dielectric is the sum of the external electric field and the polarization electric field induced in the dielectric

The polarization electric field is thus proportional to both the total electric field (proportionality constant $-\chi$) and the external field [proportionality constant $-\chi/(1+\chi)$]. The former relationship is more fundamental, since the field displacing charges in one atom is the total field, due to both external charges and to the charges in neighboring atoms.

The total field within the dielectric is

$$E_{\text{tot}} = E_{\text{ext}} - \frac{\chi}{1+\chi}E_{\text{ext}} = \frac{1}{1+\chi}E_{\text{ext}} = \frac{1}{\kappa}E_{\text{ext}}. \quad (6.20)$$

The factor $\kappa = 1 + \chi$ is called the *dielectric constant* of the dielectric. The electric field within the dielectric is reduced by the factor $1/\kappa$ from that which would exist without

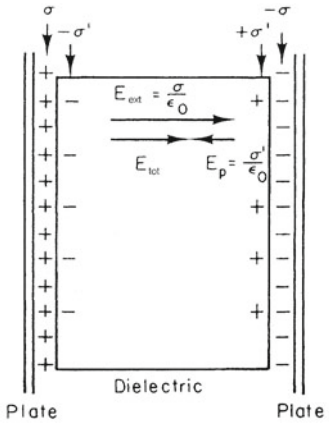


Fig. 6.17 The polarization electric field reduces the electric field between the plates. The conducting plates could be extracellular and intracellular fluid, and the dielectric could be the cell membrane

the dielectric. The dielectric constant for typical nerve membranes[8] is about 7. The dielectric constant of water is quite high (around 80) because the water molecules can easily reorient their charged ends.

The relationship between the applied field, the polarization field, and the total field can be seen in the following example. The electric field between two parallel sheets of charge of density $+\sigma$ and $-\sigma$ per unit area has magnitude $E_{\text{ext}} = \sigma/\epsilon_0$. If there is dielectric between them (such as a cell membrane) and if the polarization in the dielectric is uniform, then there is effectively a charge $\pm\sigma'$ induced on the surface of the dielectric that partially neutralizes the external charges. This is shown in Fig. 6.17. The total electric field within the membrane is $E_{\text{tot}} = \left|\mathbf{E}_{\text{ext}} + \mathbf{E}_p\right| = \sigma/\epsilon_0 - \sigma'/\epsilon_0 = \sigma_{\text{net}}/\epsilon_0 = E_{\text{ext}}/\kappa$.

To recapitulate, in Fig. 6.17 E_{ext} is σ/ϵ_0 and depends on the external charge distribution; the potential difference between the plates depends on the total field, and its magnitude is E_{tot} times the plate separation.

It is customary to refer to two different kinds of charge. The *free charge* is the charge that we bring into a region. We have some control over it. The *bound charge* is the charge induced in the dielectric by the movement or distortion of atoms and molecules in the dielectric in response to the free charge that has been introduced. Gauss's law can be written either in terms of the total charge (free plus bound)

$$\iint E_n\, dS = \frac{q_{\text{tot}}}{\epsilon_0} = \frac{q_{\text{free}} + q_{\text{bound}}}{\epsilon_0} \quad (6.21a)$$

[8] This value is high compared to the dielectric constant for a pure lipid, which is between 2 and 3. See the discussion in Sect. 6.17.

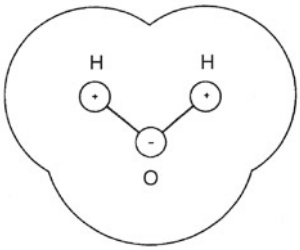


Fig. 6.18 A schematic diagram of a water molecule. The hydrogen nuclei are 96.5 pm from the oxygen nucleus; the included angle is about 104°. The radius of each hydrogen atom is about 120 pm; the radius of the oxygen atom is about 140 pm. The water molecule has a permanent electric dipole moment because the oxygen atom carries a partial negative charge and each hydrogen atom carries a partial positive charge

or in terms only of the free charge

$$\iint \kappa E_n \, dS = \frac{q_{\text{free}}}{\epsilon_0}. \tag{6.21b}$$

The dielectric constant is placed inside the integral sign because the Gaussian surface could pass through materials with different values of the dielectric constant.

As another example of the effect of a dielectric, consider a spherical ion of radius a in which all the charge resides on the surface. In a vacuum, the potential at distance r is $v = q/4\pi\epsilon_0 r$, so on the surface of the ion, the potential is $q/4\pi\epsilon_0 a$. The work required to bring to the surface an additional charge dq is $dW = v dq = q dq/4\pi\epsilon_0 a$. The total work required to place charge Q on the ion is therefore

$$W = \int dW = \frac{1}{4\pi\epsilon_0 a} \int_0^Q q \, dq = \frac{\frac{1}{2}Q^2}{4\pi\epsilon_0 a}.$$

If the sphere is immersed in a uniform dielectric the total electric field and therefore the potential is reduced by a factor κ. The energy required to assemble the ion is then

$$W = \frac{\frac{1}{2}Q^2}{4\pi\epsilon_0 \kappa \, a}. \tag{6.22}$$

This is called the *Born charging energy*. For an ion of radius 0.2 nm (200 pm) and $Q = 1.6 \times 10^{-19}$ C, the Born charging energy in a vacuum is 5.8×10^{-19} J ion^{-1}. Multiplying by Avogadro's number gives 3.5×10^5 J mol^{-1}. Often in problems involving charges of a few times the electronic charge, it is convenient to use the energy unit electron volt: 1 eV$= 1.6 \times 10^{-19}$ J. For this problem, the Born charging energy is 3.6 eV ion^{-1}.

If the ion is in a dielectric with $\kappa = 2$ (a lipid, for example), the Born charging energy is reduced to 1.8 eV ion^{-1}. Water has a very high dielectric constant (about 80) because the water molecules look roughly like that in Fig. 6.18, and the molecules can easily align with an applied electric field. The same ion in water has a Born charging energy of 0.045 eV. At room temperature, the Boltzmann factor for the energy required to create the ion in vacuum is 3.32×10^{-61}. In a lipid, it is 5.76×10^{-31}, and in water, it is 0.175. This explains why it is easy to form ionic solutions in water but not in lipids.

6.8 Current and Ohm's Law

In the electrostatic case, there are no moving charges and no electric field within a conductor. When a current flows in a conductor, charges are moving and there is an electric field.

The electric current i in a wire is the amount of charge per unit time passing a point on the wire. If the amount of charge in time dt is dQ, the current is

$$i = \frac{dQ}{dt}. \tag{6.23}$$

The units of the current are C s^{-1} or amperes (A) (sometimes called amps). The current density j (or j_Q in the notation of Chap. 5) is the current per unit area, i/S. The units are C m^{-2} s^{-1} or A m^{-2}. In an extended medium, the current density is a vector **j** at each point in the medium. The direction of **j** is the direction charge is moving at that point.

If there is no electric field in the conductor, there is no average motion of the charges. (There will be random thermal motion, but it will be equally likely in every direction. This random motion of charges is one cause of "noise" in electrical circuits.) To have a current there must be an electrical field in the conductor; this means that there will be a potential difference between two points in the conductor. If there is no potential difference between two points in the conductor, there is no current. For the simple conductor of Fig. 6.19, the current is found to be proportional to the voltage difference between the ends of the conductor. The current is shown flowing from B to A. When $v(B)$ is greater than $v(A)$, v is positive and the current is positive. When v is negative, the current is in the other direction and is also negative.

For the wire of Fig. 6.19, the relationship between current and voltage difference is linear. In that case, we can write *Ohm's law*:

$$i = \frac{1}{R} v = G v \tag{6.24a}$$

or

$$v = i R. \tag{6.24b}$$

R is called the *resistance* of the conductor. Since the current is measured in amps and the voltage in volts, its units are V A^{-1} or ohms (Ω). The reciprocal of the resistance is the *conductance* G. Its units are Ω^{-1} or siemens (S).

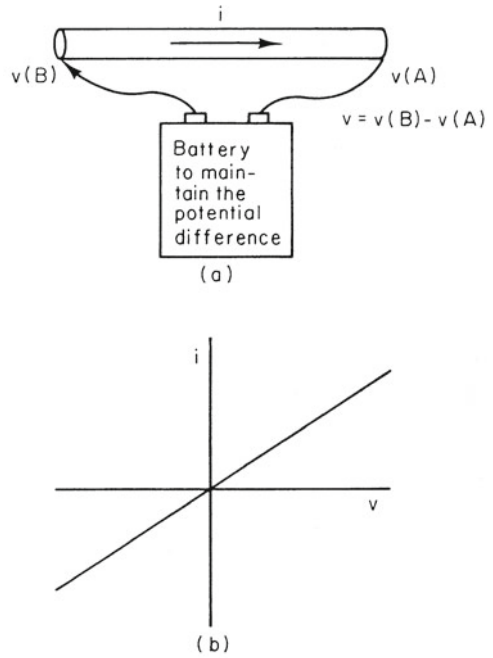


Fig. 6.19 A current flows in the wire as long as the battery or some other device maintains a potential difference between two points on the wire. The potential difference means that there is an electric field within the wire. If the wire obeys Ohm's law, the current is proportional to the potential difference

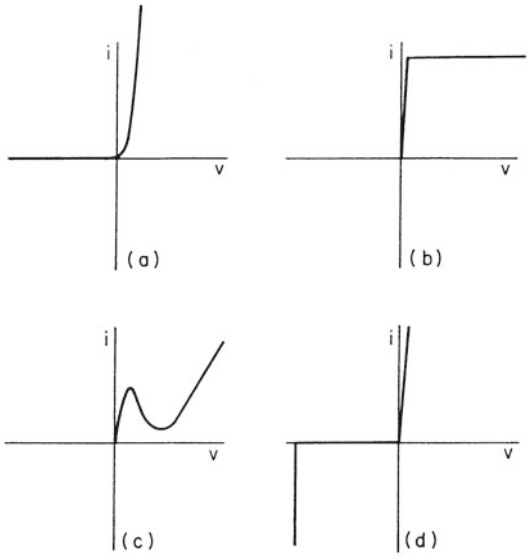


Fig. 6.20 Current–voltage relationships for some nonlinear devices used in electronic circuits. **a** Diode. **b** Transistor. **c** Tunnel diode. **d** Zener diode

Ohm's law is not universal. It describes only certain types of conductors. Figure 6.20 shows the current–voltage characteristics of several devices that have nonlinear behavior and that make modern electronic circuits possible (Horowitz and Hill 2015). They are shown here not for their own sake, but to emphasize the limited validity of Ohm's law. The nerve cell membrane is not linear.

It is possible to write Ohm's law in another form. Placing two identical wires in parallel in the circuit of Fig. 6.19 would cause twice as much current to flow (assuming that the battery maintains the voltage difference at the original level). The current density j remains constant as the cross-sectional area of the wire is changed, when the wire length and voltage difference are held fixed. Similarly, to maintain the same current through a single wire twice as long requires a voltage difference twice as great. Therefore, it is voltage per unit length that determines the current. In this spirit, Ohm's law can be written as

$$j = \frac{i}{S} = \frac{v(B) - v(A)}{SR}.$$

If L is the length of the wire and x the position along it, this can be written as

$$j_x = -\frac{L}{SR}\frac{v(x = L) - v(x = 0)}{L} = -\frac{L}{SR}\frac{\partial v}{\partial x}, \quad (6.25a)$$

$$j_x = -\sigma\frac{\partial v}{\partial x}. \quad (6.25b)$$

In three dimensions this alternative statement of Ohm's law becomes

$$\mathbf{j} = \sigma\mathbf{E}. \quad (6.26)$$

The σ in this equation[9] is the electrical conductivity, measured in (A m^{-2})/(V m^{-1}) or S m^{-1}. Its reciprocal is the resistivity of the material, ρ. The units of resistivity are Ω m. For a cylindrical conductor, the resistivity and the resistance are related by

$$\frac{1}{\rho} = \frac{L}{SR}$$

or

$$R = \rho\frac{L}{S}. \quad (6.27)$$

This shows that making the conductor longer increases its resistance, while increasing the cross-sectional area lowers the resistance.

Suppose that an electric field acts on a charge moving in a medium that obeys Ohm's law. The electric field does work on the charge, but the energy is continually transferred to the medium by collisions between the charge and other particles

[9] Note that σ has now been used for two things in this chapter: surface charge per unit area and conductivity. This notation is standard in the literature. You can tell from the context which is meant. Similarly, the symbol ρ is used for charge per unit volume and for resistivity (and for mass density in other chapters). These double usages are found frequently in the literature.

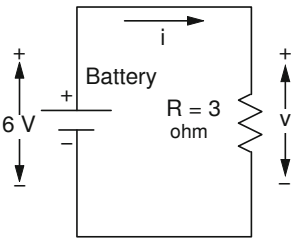


Fig. 6.21 A resistor connected to a battery

in the medium. If a charge Q moves to a lower potential, all the energy it gained is transferred to heat. The rate of energy dissipation is the *power*

$$P = vi. \tag{6.28}$$

The units of power are J s^{-1} or watts (W). For a material that obeys Ohm's law, Eq. 6.28 can be combined with Ohm's law to give

$$P = i^2 R \tag{6.29}$$

or

$$P = \frac{v^2}{R}. \tag{6.30}$$

This type of energy loss has clinical significance. If a patient contacts a source of very high voltage such as an 11,000-V power line, the strong electric fields will cause current to flow throughout the patient's body or limb, because $\mathbf{j} = \sigma \mathbf{E}$. The resistive heating can be enough to boil water within the tissues. If the limb is x rayed, the steam bubbles will look very much like the bubbles that appear in *clostridium* (gas gangrene) infections; if the x ray is deferred a few days, it will be impossible to tell from the x ray whether the bubbles are due to the electrical injury or subsequent infection.

6.9 The Application of Ohm's Law to Simple Circuits

The ultimate goal of this chapter is to apply Ohm's law to the axon. Before doing that, however, it is worthwhile to see how it can be applied to some simpler circuits in which the current and voltage are not changing with time.

The simplest circuit is a resistance R connected across a battery, as shown in Fig. 6.21. The battery voltage of 6 V is the potential difference across the resistor. If the resistance is 3 Ω, the current is $i = v/R = 6/3 = 2$ A. The rate of heat production in the resistor is $P = vi = (6)(2) = 12$ W. This could also have been calculated from $P = v^2/R = 36/3$, or $P = i^2 R = (4)(3)$. A current of 2 A means that every second 2 C of charge leave the positive terminal of the battery and flow through the resistor. When the charge arrives at the other end of the resistor, it has lost 12 J of energy to heat. The 2 C then travel through the battery back to the positive terminal, gaining 12 J from a chemical reaction within the battery.

This example has been stated as though the positive charge moves. In a metallic conductor negative charges (electrons) move from the negative terminal of the battery through the resistor to the positive terminal. In salt water and most body fluids, both positive and negative ions move. From a macroscopic point of view, we cannot tell the difference between the transport of a charge $-q$ from point A to point B, and the transport of a charge $+q$ from point B to point A. Both processes make the total charge at B less positive and the total charge at A more positive by an amount q.

Two fundamental principles used in this discussion have not been explicitly stated. The first is the *conservation of electric charge*: all charge that leaves the battery passes through the resistor. The second is the *conservation of energy*: a charge that starts at some point in the circuit and comes back to its starting point has neither lost nor gained energy. (The energy gained by a charge in the battery is equal to the energy lost by it in passing through the resistor.) These principles become less obvious and more useful in a circuit that is more complicated than the one considered above. They are known as *Kirchhoff's laws*.

In a more complicated circuit, Kirchhoff's first law (conservation of charge) takes the following form. Any junction where the current can flow in different paths is called a *node*. The algebraic sum of all the currents into a node is zero. (By algebraic sum we mean that currents into the node are positive, while currents leaving the node are negative, or vice versa.) This ensures that no charge will accumulate at the node.[10]

As an example of Kirchhoff's first law, consider the node in Fig. 6.22. Conservation of charge requires that $2+3+i = 0$ or $i = -5$ A. (In this case positive currents flow into the node; the negative current means that 5 A is flowing out of the node as current i.)

Kirchhoff's second law was used implicitly in the example above to say that the voltage across the resistor is 6 V. In general, Kirchhoff's second law says that if one goes around any closed path in a complicated circuit, the total voltage change is zero.

[10] More generally, the node could represent a conductor, such as the plate of a capacitor, on which charge can accumulate. In that case the charge Q changes with time:

$$\frac{dQ}{dt} = \sum(\text{all currents into the node}).$$

This statement is quite similar to the continuity equation of Sect. 4.1.

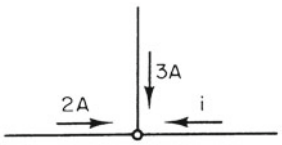


Fig. 6.22 Conservation of charge means that current i is -5 A

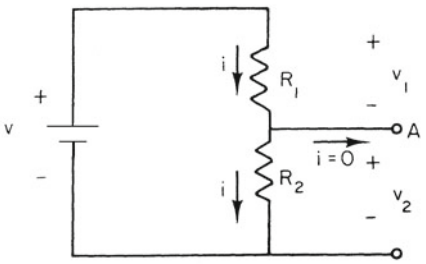


Fig. 6.23 A more complicated circuit, sometimes called a voltage divider

Kirchhoff's laws can be applied to show that the total resistance of a set of resistors in series is

$$R = R_1 + R_2 + R_3 + \cdots .$$

This follows from the definition of resistance, the fact that the same current flows in each resistor, and the total potential difference across the set of resistors is the sum of the potential difference across each one. Kirchhoff's laws can also be used to show that for a collection of resistors in parallel, the total resistance is given by

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots ,$$

(see Problem 24).

Consider a more complicated example in which two resistors are connected across a battery. The battery voltage is v, and the resistances are R_1 and R_2, as shown in Fig. 6.23. If no current flows out lead A, then conservation of charge requires that the same current i flows in each resistor. The sum of the voltages v_1 and v_2 is v. Therefore, $i = v_1/R_1 = v_2/R_2$ and $v = v_1 + v_2 = iR_1 + iR_2 = i(R_1 + R_2)$. The voltage across R_2 is iR_2 or

$$v_2 = \frac{R_2}{R_1 + R_2} v. \tag{6.31}$$

6.10 Charge Distribution in the Resting Nerve Cell

The axon consists of an ionic intracellular fluid and an ionic extracellular fluid, separated by a membrane. The intracellular and extracellular media are electrical conductors. When the cell is in equilibrium there is no current and no electric field in these regions. There will be a field and currents when an impulse is traveling along the axon.

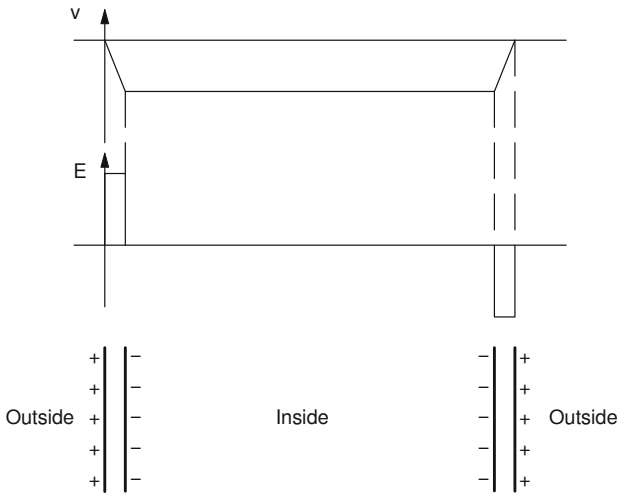


Fig. 6.24 The potential, electric field, and charge at different points on the diameter of a resting nerve cell. Portions of the cell membrane on opposite sides of the cell are shown. Outside the cell on the left the potential and electric field are zero. As one moves to the right into the cell, the electric field in the membrane causes the potential to decrease to -70 mV. Within the cell the field is zero and the potential is constant. Moving out through the right-hand wall the potential rises to zero because of the electric field within the membrane

Because the electric field in the resting cell is zero, there is no net charge in the fluid. Positive ions are neutralized by negative ions everywhere except at the membrane. A layer of charge on each surface generates an electric field within the membrane and a potential difference across it.

Measurements with a microelectrode show that the potential within the cell is about 70 mV less than outside. If the potential outside is taken to be zero, then the interior resting potential is -70 mV. Figure 6.24 shows a slice across the cell, showing the membrane on opposite sides of the cell and the charges and electric field. If the potential drops 70 mV as one enters the cell on the left, if the membrane thickness is 6 nm, and if the electric field within the membrane is assumed to be constant, then

$$E = -\frac{dv}{dx} = -\frac{-70 \times 10^{-3}\ \text{V}}{6 \times 10^{-9}\ \text{m}} = 1.17 \times 10^{7}\ \text{V m}^{-1}. \tag{6.32}$$

This is how the value of E was determined for use on p. 147.

Except for the layers of charge on the inside and outside of the membrane, which are shown in Fig. 6.24 and which give rise to the electric field and potential difference, the extracellular and intracellular fluids are electrically neutral. However, the ion concentrations are quite different in each (Fig. 6.3). There is an excess of sodium ions outside and an excess of potassium ions inside.

It is possible to see which concentrations (if any) are consistent with the hypothesis that the ions can pass freely through the membrane. If a species is in equilibrium, the concentration ratio c_i/c_o across the membrane is given by a Boltzmann factor or the Nernst equation (see Chap. 3). The potential energy of the ion is zev, where z is the valence of the ion, e the electronic charge (1.6×10^{-19} C), and v the potential in volts. Using subscripts i and o to represent inside and outside the cell, we have

$$\frac{c_i}{c_o} = \frac{e^{-zev_i/k_BT}}{e^{-zev_o/k_BT}} = e^{-ze(v_i - v_o)/k_BT}. \tag{6.33}$$

Here k_B is Boltzmann's constant, 1.38×10^{-23} J K^{-1}. For a situation in which $T = 310$ K and $v_i - v_o = -70 \times 10^{-3}$ V, c_i/c_o is 13.7 for univalent positive ions and $1/13.7 = 0.073$ for negative ions. The ratios in Fig. 6.3 are 0.103 for sodium, 30 for potassium, and 0.071 for chloride. The chloride concentration ratio is consistent with equilibrium, while the sodium concentration ratio is much too small (too few sodium ions inside) and the potassium concentration ratio is too large (too many potassium ions inside). A potential of −90 mV would bring the potassium concentration ratio into equilibrium, but then chloride would not be in equilibrium and sodium would be even farther from equilibrium. In fact, tracer studies show that potassium leaks out slowly and sodium leaks in slowly. The resting membrane is not completely impermeable to these ions (Hodgkin 1964, Chap. 6; Laüger 1991). To maintain the ion concentrations a membrane protein called the sodium–potassium pump uses metabolic energy to pump potassium into the cell and sodium out. The usual ratio of sodium to potassium ions in this active transport is 3 sodium to 2 potassium ions (Patton et al. 1989, Vol. 1, p. 27).

The intracellular and extracellular fluids can be modeled as two conductors separated by a fairly good insulator. The conductors have a capacitance between them. We can estimate this capacitance in two ways. We can either regard the membrane as a plane insulator sandwiched between plane conducting plates (as if the membrane had been laid out flat as in Fig. 6.25), or we can treat it as a dielectric between concentric cylindrical conductors. The text will use the first approximation, while the second will be left to a problem. Suppose that two parallel plates have area S and charge $\pm Q$, respectively, then the charge density on each is $\sigma = \pm Q/S$. Equation 6.11 gives the electric field without a dielectric between the conductors: $E_{\text{ext}} = \sigma/\epsilon_0 = Q/\epsilon_0 S$.

With the dielectric of dielectric constant κ, the field is reduced to $E = E_{\text{ext}}/\kappa = \sigma/\kappa\epsilon_0 = Q/\kappa\epsilon_0 S$ as was seen in Eq. 6.20. The magnitude of the potential difference is E times the plate separation b: $v = Eb = Qb/\kappa\epsilon_0 S$. The capacitance is $C = Q/v$:

$$C = \frac{Q\kappa\epsilon_0 S}{Qb} = \frac{\kappa\epsilon_0 S}{b}. \tag{6.34}$$

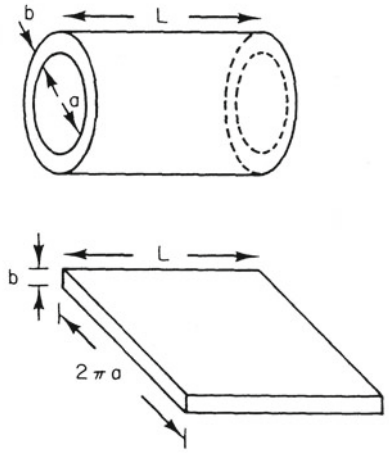


Fig. 6.25 A portion of a cell membrane of length L, in its original configuration and laid out flat. The membrane thickness is b and the radius of the axon is a. The plane approximation is used to calculate both the capacitance and resistance of the membrane

The charge density on the surface of the membrane is obtained from $\sigma = Q/S = Cv/S = \kappa\epsilon_0 v/b$.

Measurements of the dielectric constant κ for axon membrane show it to be about 7. Using values of −70 mV for v and 6 nm for b, the capacitance per unit area of membrane can be calculated, as can σ:

$$\frac{C}{S} = \frac{(7)(8.85 \times 10^{-12})}{6 \times 10^{-9}} = 0.01 \text{ F m}^{-2} = 1 \text{ μF cm}^{-2},$$

$$\sigma = (0.01)(70 \times 10^{-3}) = 7 \times 10^{-4} \text{ C m}^{-2}. \tag{6.35}$$

This value for the surface charge density is larger by a factor of 7 than that calculated in Sect. 6.3. The reduction of the electric field by polarization of the dielectric has been taken into account in the present calculation. A larger external charge is required to give the same field within the dielectric.

The value of b for myelinated fibers is much greater, typically 2000 nm instead of 6 nm. This reduces the capacitance per unit area by a factor of 300.

6.11 The Cable Model for an Axon

We now consider the rather complicated flow of charge in the interior of an axon, through the membrane, and in the conducting medium outside the cell during departures from rest. We will model the axon by electric conductors that obey Ohm's law inside and outside the cell and a membrane that has capacitance and also conducts current. We will apply

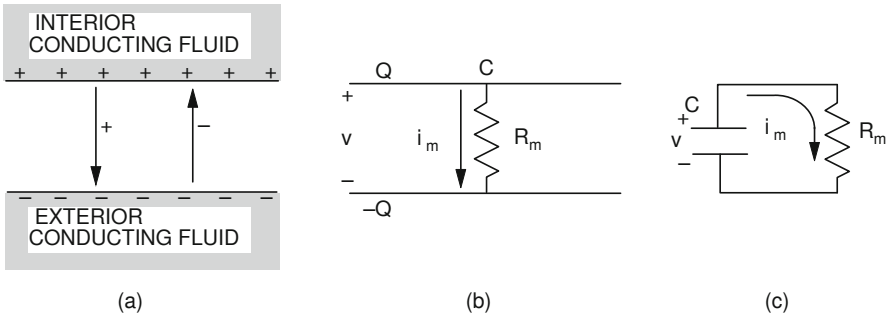


Fig. 6.26 Leakage currents through the membrane. **a** The flow of positive and negative ions. **b** The membrane capacitance is represented by the parallel plates and the leakage resistance by a single resistor. **c** The capacitance and resistance are usually drawn like this

Kirchhoff's laws—conservation of energy and charge—to a small segment of the axon. The result will be a differential equation that is independent of any particular model for the cell membrane. This is called the *cable model* for an axon. We will then apply the cable model in two cases. The first case is when the voltage change does not alter the properties of the membrane. The second case is a voltage change that changes the ionic permeability of the membrane, thereby generating a nerve impulse.

Consider the small segment of membrane shown in Fig. 6.26a. For the moment we ignore the resting potential on the membrane. We will see later that accounting for the resting potential requires only a small change to the model. The upper capacitor plate, corresponding to the inside of the membrane, carries a charge Q. The lower capacitor plate (the outside of the membrane) has charge $-Q$. The charge on the membrane is related to the potential difference across the membrane by the membrane capacitance C_m: $Q = C_m v$. Figure 6.26a shows positive ions on the inside and negative ions on the outside of the membrane. (In a resting nerve cell, there is negative charge on the inside of the membrane, Q is negative, $-Q$ is positive, and $v < 0$.)

If the resistance between the plates of a capacitor is infinite, no current flows, and the charge on the capacitor plates remains constant. However, a membrane is not a perfect insulator; if it were, there would be no nerve conduction. Some current flows through the membrane. We call this current i_m and *define* outward current to be positive, as in Fig. 6.26b.

Imagine for now that there is no current along the axon. In that case i_m discharges the membrane capacitance, and the charge and potential difference fall to zero as charge flows through the resistor. When i_m is positive, Q and v decrease with time:

$$-i_m = \frac{dQ}{dt} = C_m \frac{dv}{dt}. \tag{6.36}$$

Let us explore the behavior of this isolated segment of axon a bit further. For now we think of the total leakage current as being through a single effective resistance R_m. This is shown in Fig. 6.26b. It is customary to draw the resistance separately, as in Fig. 6.26c. The current is then $i_m = v/R_m$ and $C_m(dv/dt) = -i_m = -v/R_m$,

$$\frac{dv}{dt} = -\frac{1}{R_m C_m} v. \tag{6.37}$$

This is the familiar equation for exponential decay of the voltage (see Chap. 2). If the initial voltage at $t = 0$ is v_0, the solution is

$$v(t) = v_0 e^{-t/\tau}, \tag{6.38}$$

where the time constant τ is given by

$$\tau = R_m C_m. \tag{6.39}$$

Referring to Fig. 6.25, we saw that if we have a section of membrane of area S and thickness b the capacitance is given by Eq. 6.34. For a conductor of the same dimensions we saw [Eq. 6.27] that the resistance is $R_m = \rho_m b/S$, so the time constant is

$$\tau = R_m C_m = \frac{\rho_m b}{S} \frac{\kappa \epsilon_0 S}{b} = \kappa \epsilon_0 \rho_m. \tag{6.40}$$

We have the remarkable result that the time constant is independent of both the area and thickness of the membrane. Doubling the area S doubles the amount of charge that must leak off, but it also doubles the membrane current. Doubling b doubles the resistance, but it also makes the membrane capacitance half as large. In each case the factors S and b cancel in the expression for the time constant.

If a very thin lipid membrane is produced artificially, it is found to have a very high resistivity—about 10^{13} Ω m (Scott 1975, p. 493). Certain proteins added to the lipid material reduce the resistivity by several orders of magnitude. For natural nerve membrane the resistivity is about

$$\rho_m = 1.6 \times 10^7 \ \Omega \text{ m}. \tag{6.41}$$

This is the effective resistivity for resting membrane, taking into account all of the ion currents. If ρ_m had this constant value the time constant would be $\tau = \kappa \epsilon_0 \rho_m =$

Table 6.1 Properties of a typical unmyelinated nerve

a	Axon radius	5×10^{-6} m
b	Membrane thickness	6×10^{-9} m
ρ_i	Resistivity of axoplasm	0.5 Ω m
$r_i = \rho_i/\pi a^2$	Resistance per unit length inside axon	6.4×10^9 Ω m^{-1}
κ	Dielectric constant of membrane	7[a]
ρ_m	Resistivity of membrane	16×10^6 Ω m
$\kappa\rho_m$		112×10^6 Ω m
$c_m = \kappa\epsilon_0/b$	Membrane capacitance per unit area	10^{-2} F m^{-2}
$2\pi\kappa\epsilon_0 a/b$	Membrane capacitance per unit length of axon	3×10^{-7} F m^{-1}
$g_m = 1/\rho_m b$	Conductance per unit area of membrane	10 S m^{-2}
$1/g_m$	Reciprocal of conductance per unit area	0.1 Ω m^2
$2\pi a/\rho_m b$	Membrane conductance per unit length of axon	3.2×10^{-4} S m^{-1}
v_r	Resting potential inside axon	−70 mV
$E = v_r/b$	Electric field in membrane	1.2×10^7 V m^{-1}
$\kappa\epsilon_0 v_r/b$	Charge per unit area on membrane surface	7×10^{-4} C m^{-2}
	Net number of univalent ions per unit area	4.4×10^{15} m^{-2}
	Net number of univalent ions per unit length	6.6×10^7 m^{-1}

[a]See Sect. 6.17 for a discussion of the dielectric constant.

$(7)(8.85 \times 10^{-12})(1.6 \times 10^7) = 1 \times 10^{-3}$ s. (Actually, the resistivity changes drastically as the potential across the membrane changes during the propagation of a nerve impulse.) Since we observe a potential difference across the membrane, there must be a mechanism for renewing the charge on the membrane surface.

The resistance and capacitance of the portion of the axon membrane in Fig. 6.25 can be written in terms of the axon radius a and the length L of the segment by noting that $S = 2\pi aL$. Then one has

$$C_m = \frac{\kappa\epsilon_0 2\pi aL}{b}, \quad R_m = \frac{\rho_m b}{2\pi aL}.$$

It is convenient to recall that $v = iR$ can be written as $i = Gv$ and introduce the conductance of the membrane segment

$$G_m = \frac{2\pi aL}{\rho_m b}. \tag{6.42}$$

Both the capacitance and the conductance are proportional to the area of the segment S. It is also convenient to introduce the lowercase symbols c_m and g_m to stand for the membrane capacitance and membrane conductance per unit area:

$$c_m = \frac{C_m}{S} = \frac{\kappa\epsilon_0}{b}, \tag{6.43}$$

$$g_m = \frac{G_m}{S} = \frac{1}{\rho_m b} = \frac{\sigma_m}{b}. \tag{6.44}$$

(Remember that $\sigma_m = 1/\rho_m$ is the electrical conductivity, the reciprocal of the resistivity. It is *not* the charge per unit area. σ is frequently used for both quantities in the literature.)

Both c_m and g_m depend on the membrane thickness as well as the dielectric constant and resistivity of the membrane. The units of c_m and g_m are, respectively, F m^{-2} and S m^{-2}. Be careful: many sources give them per square centimeter instead of per square meter.

We can rewrite Eq. 6.36 in terms of the current density j_m, which is proportional to the capacitance per unit area, c_m:

$$-j_m = c_m \frac{dv}{dt}. \tag{6.45}$$

Table 6.1 shows typical values for these quantities and some to be discussed later for an unmyelinated axon.[11] These values should not be associated with a particular species. Parameters such as the resistance and capacitance per unit length of the axon are measured directly. Others, such as ρ_m, require an estimate of membrane thickness and are less well known

Now let us consider current that flows inside and outside the axon. Assume that the currents inside are longitudinal, that is, parallel to the axis of the axon. A discussion of departures from this assumption is found in Scott (1975, p. 492). With this assumption, the interior fluid can be regarded as a resistance of length L and radius a as shown in Fig. 6.27. The resistance of such a segment is $R_i = \rho_i L/S = \rho_i L/\pi a^2$. It is convenient to work with the resistance per unit length, r_i:

$$r_i = \frac{R_i}{L} = \frac{\rho_i}{\pi a^2} = \frac{1}{\pi a^2 \sigma_i}. \tag{6.46}$$

[11] Some insight into the magnitude of the charge on the membrane can be obtained from these numbers. The excess charge on the surface of the membrane is 7×10^{-4} C m^{-2} for the unmyelinated fiber. This corresponds to 4.4×10^{15} ions m^{-2}, if each ion has a charge of 1.6×10^{-19} C. Each atom or ion in contact with the membrane surface occupies an area of about 10^{-20} m^2; thus there are about 10^{20} atoms or ions in contact with a square meter of membrane surface. These may be neutral or positively or negatively charged. If charged, most are neutralized by the presence of a neighbor of opposite charge. The excess charge density that is required can be obtained if $4.4 \times 10^{15}/10^{20}$ or roughly one out of every 20, 000 of the atoms in contact with the surface is ionized and not neutralized.

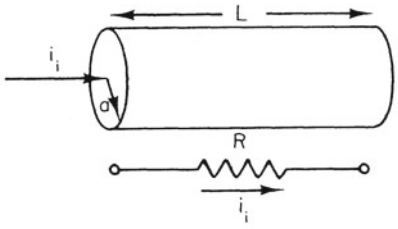


Fig. 6.27 Axoplasm of length L and radius a can be treated like a simple resistor

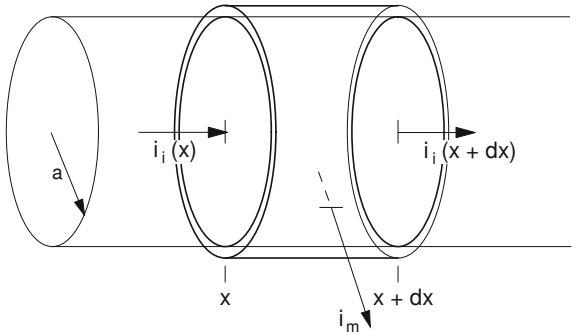


Fig. 6.28 The membrane surrounding a small portion of an axon is shown, along with the longitudinal currents in and out of the segment

The question of resistance of the extracellular fluid for currents outside the axon is more complicated. If the extracellular fluid were infinite in extent, the longitudinal resistance outside the cell would be very small (see Chap. 7). On the other hand, in a nerve or a muscle the axons or muscle cells are packed close together, there is not very much extracellular fluid, and the external resistance per unit length can be significant. There are some important effects that occur because of this. We will discuss them in Chap. 7.

Now we can consider the effect of both membrane and longitudinal currents. Figure 6.28 shows a small region of the axon between x and $x + dx$ and the surrounding membrane. Current i_i flows longitudinally along the axon on the inside. The current through the membrane is i_m. The potential difference across the membrane is $v = v_i - v_o$. In this section no attempt will be made to relate i_m or j_m to v. Charge Q resides on the inside surface of the membrane and can be either negative or positive. An equal and opposite charge $-Q$ resides on the outer surface of the membrane.

Because the capacitance can charge or discharge, Kirchhoff's law (conservation of charge) does not say that the sum of the currents is zero. Rather, it says that the net current into the volume of axoplasm between x and $x + dx$ changes the charge on the interior surface of the membrane:

$$i_i(x) - i_i(x + dx) - i_m = \frac{dQ}{dt} = C_m \frac{d(v_i - v_o)}{dt}. \quad (6.47a)$$

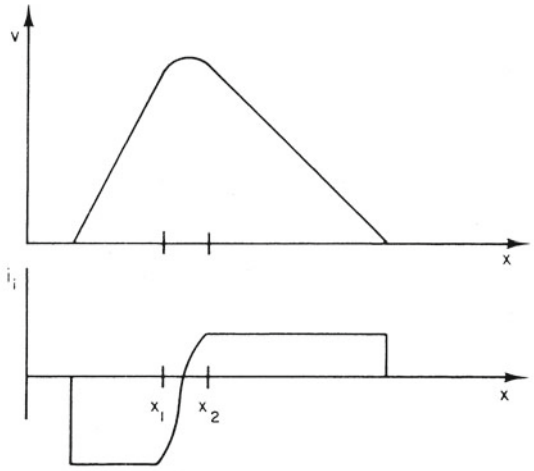


Fig. 6.29 A hypothetical plot of $v_i(x)$ and the longitudinal current i_i associated with it

When $i_i(x) = i_i(x + dx)$ this gives Eq. 6.36. The membrane current i_m represents an average value for the segment of membrane between x and $x + dx$. It is also a function of x.

We can define $di_i = i_i(x + dx) - i_i(x)$ as the increase in i_i along segment dx. Then we can rewrite Eq. 6.47a as

$$-di_i = C_m \frac{dv}{dt} + i_m. \quad (6.47b)$$

This is an important equation. It says that when the current flowing inside the axon decreases in a small distance dx, part of the current charges the capacitance of that segment of membrane, and the rest flows through the membrane.

Consider a small segment of axoplasm of length dx. The intracellular voltage at the left end is $v_i(x)$; at the right end it is $v_i(x + dx)$. The current along the segment is the voltage difference between the ends divided by the resistance of the segment. The resistance is $R_i = r_i\, dx$. Therefore the current is

$$i_i(x) = \frac{v_i(x) - v_i(x + dx)}{r_i\, dx} = -\frac{1}{r_i} \frac{dv_i}{dx}. \quad (6.48)$$

The voltage must change along the axon for a current to flow within it. The minus sign in Eq. 6.48 shows that a current flowing from left to right (in the $+x$ direction) requires a voltage that decreases from left to right, and vice versa. Figure 6.29 shows a hypothetical plot of $v_i(x)$ and the current which would accompany it. Notice that the current is flowing from the region of higher voltage to lower voltage–towards both ends from the region between x_1 and x_2. In that region either the charge on the membrane is changing or current is flowing through the membrane.

Consider again the cylindrical geometry shown in Fig. 6.28. The surface area of this portion of membrane is

$2\pi a\,dx$. Dividing each term of Eq. 6.47a by the area and remembering the definitions of j_m and c_m we obtain

$$c_m \frac{\partial v}{\partial t} = -j_m + \frac{1}{2\pi a}\left[\frac{i_i(x) - i_i(x+dx)}{dx}\right]. \tag{6.49}$$

It is necessary to use partial derivatives because the current and voltage depend on both x and t as an impulse travels down the nerve. As $dx \to 0$

$$\frac{i_i(x+dx) - i_i(x)}{dx} \to \frac{\partial i_i}{\partial x}.$$

This can be evaluated using the expression for Ohm's law in the axoplasm, Eq. 6.48:

$$\frac{\partial i_i}{\partial x} = -\frac{1}{r_i}\frac{\partial^2 v_i}{\partial x^2}. \tag{6.50}$$

When this is inserted in Eq. 6.49 the result is

$$c_m \frac{\partial (v_i - v_o)}{\partial t} = -j_m + \frac{1}{2\pi a\, r_i}\frac{\partial^2 v_i}{\partial x^2}. \tag{6.51}$$

In many cases the extracellular potential is small. In that case the voltage across the membrane, v, is approximately the same as the intracellular voltage, v_i, so we can rewrite Eq. 6.51 as

$$c_m \frac{\partial v}{\partial t} = -j_m + \frac{1}{2\pi a\, r_i}\frac{\partial^2 v}{\partial x^2}. \tag{6.52}$$

This rather formidable looking equation is called the *cable equation* or *telegrapher's equation*. It was once familiar to physicists and electrical engineers as the equation for a long cable, such as a submarine cable, with capacitance and leakage resistance but negligible inductance (Jeffreys and Jeffreys 1956, p. 602). It has the form of Fick's second law of diffusion, Eq. 4.26, with the addition of the j_m term.

It is worth recalling the origin of each term and verifying that the units are consistent. The term on the left is the rate at which the membrane capacitance is gaining charge per unit area. Therefore all terms in the equation have the units of current per unit area. The first term on the right is the current per unit area through the membrane in the direction that discharges the membrane capacitance. The second term on the right gives the buildup of charge on this area of the membrane because of differences in current along the axon. If $v(x)$ were constant, there would be no current along the inside of the axon. If function $v(x)$ had constant slope, the current along the inside of the axon would be the same everywhere and there would be no charge buildup on the membrane. It is only because $v(x)$ changes slope that i_i is different at two neighboring points in the axon and charge can collect on the membrane.

Now, for the units. Since $i = C(dv/dt)$, the units of $c_m \partial v/\partial t$ are current per unit area. The j_m term is by definition current per unit area. Since r_i has the units of $\Omega\ \mathrm{m}^{-1}$, the term $2\pi a r_i$ has the units of Ω. When this is combined with $\partial^2 v/\partial x^2$, which has units $\mathrm{V\ m}^{-2}$, the result is $\mathrm{A\ m}^{-2}$ as required.

This is a very general equation stating Kirchhoff's laws for a segment of the axon. The only assumptions are that the currents depend only on time and position along the axon and that voltage changes on the outside of the axon can be neglected. Particular models for nerve conduction use different relations between j_m and $v(x,t)$.

6.12 Electrotonus or Passive Spread

The simplest membrane model is one that obeys Ohm's law. This approximation is valid if the voltage changes are small enough so that the membrane conductance does not change, or if something has been done to inactivate the normal changes of membrane conductance with voltage. It is also useful for myelinated nerves between the nodes of Ranvier. This is called *electrotonus* or *passive spread*.

In its quiescent state, the voltage all along the inside of the axon has the constant resting value v_r. Both $\partial v/\partial t$ and $\partial^2 v/\partial x^2$ are zero. Equation 6.52 can be satisfied only if $j_m = 0$. Although j_m is zero, it may be made up of several leakage components. In this section we simply assume that j_m is proportional to $v - v_r$:

$$j_m = g_m(v - v_r). \tag{6.53}$$

This simple model does predict that $j_m = 0$ when $v = v_r$. It also predicts that the current will be positive (outward) if $v > v_r$ and negative (inward) if $v < v_r$. It does not explain the propagation of an all-or-nothing nerve impulse. The conductance per unit area, g_m, is assumed to be independent of v and of the past history of the membrane. This is a good assumption only for very small voltage changes. With this assumption, Eq. 6.52 becomes

$$c_m \frac{\partial v}{\partial t} = -g_m(v - v_r) + \frac{1}{2\pi a\, r_i}\frac{\partial^2 v}{\partial x^2}. \tag{6.54}$$

This equation is usually written in a slightly different form by dividing through by g_m:

$$\frac{1}{2\pi a\, r_i g_m}\frac{\partial^2 v}{\partial x^2} - v - \frac{c_m}{g_m}\frac{\partial v}{\partial t} = -v_r.$$

It is also customary to make the assignments

$$\lambda^2 = \frac{1}{2\pi a\, r_i g_m},$$

$$\tau = \frac{c_m}{g_m},$$

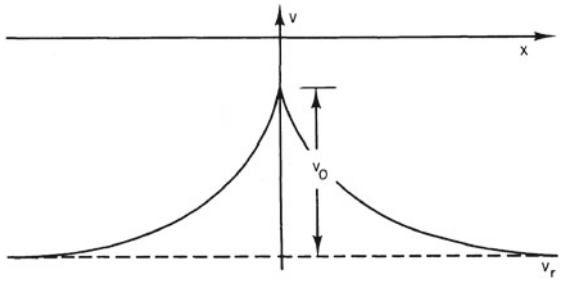


Fig. 6.30 The voltage distribution along an axon in electrotonus when the membrane capacitance is charged and the voltage is not changing with time

so that the equation becomes

$$\lambda^2 \frac{\partial^2 v}{\partial x^2} - v - \tau \frac{\partial v}{\partial t} = -v_r. \tag{6.55}$$

In terms of the primary axon parameters, the parameters in Eq. 6.55 are

$$\lambda^2 = \frac{ab\rho_m}{2\rho_i}, \tag{6.56}$$

$$\tau = \kappa \epsilon_0 \rho_m. \tag{6.57}$$

The time constant was seen before in Eq. 6.40. Equation 6.55 has a steady-state solution $v = v_r$. If a new variable $v' = v - v_r$ is used, it becomes the homogeneous version of the same equation with a steady-state solution $v' = 0$.

For nerve conduction, the inhomogeneous equation with various exciting terms corresponding to physiological stimuli was discussed by Davis and Lorente de Nó (1947) and by Hodgkin and Rushton (1946). Their work is summarized by Plonsey (1969, p. 127).

Before considering general solutions to Eq. 6.55, consider some special cases. If $c_m = 0$, so that $\tau = 0$, or if enough time has elapsed so that the voltage is not changing with time and $\partial v/\partial t = 0$, the equation reduces to

$$\lambda^2 \frac{\partial^2 v}{\partial x^2} - v = -v_r.$$

You can verify by substitution that this has a solution

$$v - v_r = \begin{cases} v_0 e^{-x/\lambda}, & x > 0 \\ v_0 e^{x/\lambda}, & x < 0. \end{cases} \tag{6.58}$$

If the voltage is held at a constant value $v = v_r + v_0$ at some point on the axon, the voltage will decay exponentially to v_r in both directions from that point. This is shown in Fig. 6.30.

Next suppose that $v(x, t)$ does not depend on x, so that there is no longitudinal current in the axon and $\partial^2 v/\partial x^2 = 0$. This can be accomplished experimentally by threading a wire axially along the axon, if the axon is fat enough. The equation reduces to

$$\tau \frac{\partial v}{\partial t} + v = v_r.$$

This is the familiar equation for exponential decay. If v were held at $v_0 + v_r$ and then the constraint were removed at $t = 0$, the voltage would decay exponentially back to v_r

$$v - v_r = v_0 e^{-t/\tau}.$$

This represents the discharge of the membrane capacitance through the membrane resistance.

The behavior of $v(x, t) - v_r$ at various times after an excitation is applied is shown in Fig. 6.31. The excitation is a constant current injected at $x = 0$ for all time $t > 0$. After a long time, the curve is identical to that in Fig. 6.30, as the membrane capacitance has fully charged. Only the membrane leakage current attenuates the signal. At earlier times the solution is not precisely exponential; the analytic solution involves error functions (Prob. 36). The change of voltage with time at fixed positions along the cable is also shown. Both the finite propagation time and the attenuation of the signal are evident.

6.13 The Hodgkin–Huxley Model for Membrane Current

If the voltage at some point along the axon changes by a few millivolts from the resting value, the voltage at other points

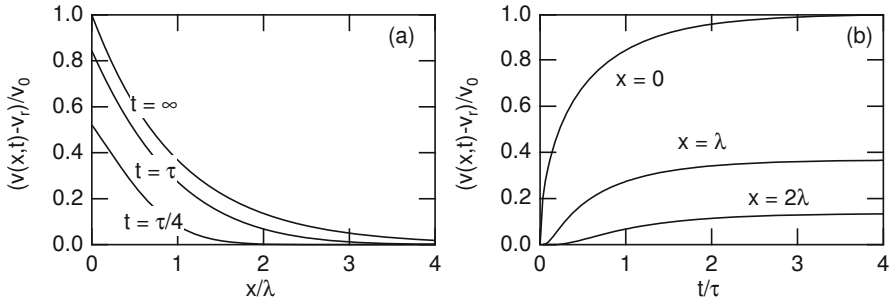


Fig. 6.31 Some representative solutions to the problem of electrotonus after the application of a constant current at $x = 0$. **a** The voltage along the axon at different times. **b** Voltage at a fixed point on the axon as a function of time

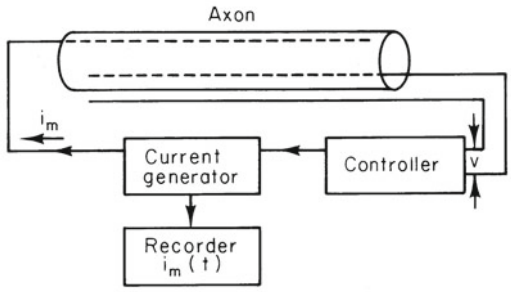


Fig. 6.32 Apparatus for voltage-clamp measurements

along the axon is described by electrotonus. However, if the inside voltage rises from the resting value by 20 mV or more, a completely different effect takes place. The potential rises rapidly to a positive value, then falls to about −80 mV, and finally returns to the resting value (Fig. 6.1). This behavior is attributable to a very nonlinear dependence of membrane current on transmembrane voltage.

Considerable work was done on nerve conduction in the late 1940s, culminating in a model that relates the propagation of the action potential to the changes in membrane permeability that accompany a change in voltage. The model (Hodgkin and Huxley 1952) does not explain why the membrane permeability changes; it relates the shape and conduction speed of the impulse to the observed changes in membrane permeability. Nor does it explain all the changes in current. (For example, the potassium current does fall eventually, and there are some properties of the sodium current that are not adequately described.) Nonetheless, the work was a triumph that led to the Nobel Prize for Alan Hodgkin and Andrew Huxley.

Most of the experiments that led to the Hodgkin–Huxley model were carried out using the giant axon of the squid. This is a single cell several centimeters long and up to 1 mm in diameter that can be dissected from the squid. The removal of axoplasm from the preparation and its replacement by electrolytes has shown that the critical phenomena all take place in the membrane. The important results are reviewed in many places (Katz 1966, Chaps. 5 and 6; Plonsey 1969, p. 127; Plonsey and Barr 2007, Chap. 4; Scott 1975, pp. 495–507).

6.13.1 Voltage Clamp Experiments

Voltage-clamp experiments were particularly illuminating. Two long wire electrodes were inserted in the axon and connected to the apparatus shown in Fig. 6.32. The resistance of the wires was so low that the potential at all points along the axon was the same at any instant of time. The potential depended only on time, and not on position. This is called a *space-clamped* experiment. One electrode, paired with an electrode in the surrounding medium, measured the voltage difference across the membrane. The other electrode was used to inject or remove whatever current was necessary to keep this voltage difference constant. Measurement of this current allowed calculation of the membrane conductance. This technique is called *voltage clamping*. The experiment described here was both voltage- and space-clamped.

When the membrane potential was raised abruptly from the resting value to a new value and held there, the resulting current was found to have three components:

1. A current, lasting a few microseconds, that changed the surface charge on the membrane.
2. A current flowing inward which lasted for 1 or 2 ms. Various experiments, such as replacing the sodium ions in the extracellular fluid with some other monovalent ion, showed that this was due to the inward flow of sodium ions. (Had the potential not been voltage-clamped by the electronic apparatus, this inrush of positive charge would have raised the potential still further.)
3. An outward current that rose in about 4 ms and remained steady for as long as the potential was clamped at this value. Tracer studies showed that this current was due to potassium ions. (Over a time scale of several tens of milliseconds, the potassium current, like the sodium current, does fall back to zero.)

The first current is the $c_m(\partial v/\partial t)$ term of Eq. 6.52; the second and third currents together constitute j_m. Because of the clamping wires, the $\partial^2 v/\partial x^2$ term is zero.

The next step is to develop a model that describes the major ionic constituents of the current. The sodium and potassium contributions to the current will be considered separately; all other contributions will be combined in a *leakage term*:

$$j_m = j_{Na} + j_K + j_L. \tag{6.59}$$

The leakage includes charge movement due to chloride ions and any other ions that can pass through the membrane.

Consider movement of sodium through the membrane. Similar considerations apply to potassium. The concentrations of sodium inside and out are $[\mathrm{Na}_i]$ and $[\mathrm{Na}_o]$. It will be seen later that the total number of ions moving through the membrane during a nerve pulse in a squid giant axon is too small to change the concentrations significantly. Therefore, the concentrations are fixed.

There would be no movement of sodium ions through the membrane, regardless of how permeable it is, when the concentrations and potential are related by the Boltzmann factor or Nernst equation (Eq. 6.33) with $v = v_i - v_o$:

$$\frac{[\mathrm{Na}_i]}{[\mathrm{Na}_o]} = e^{-ev/k_BT}.$$

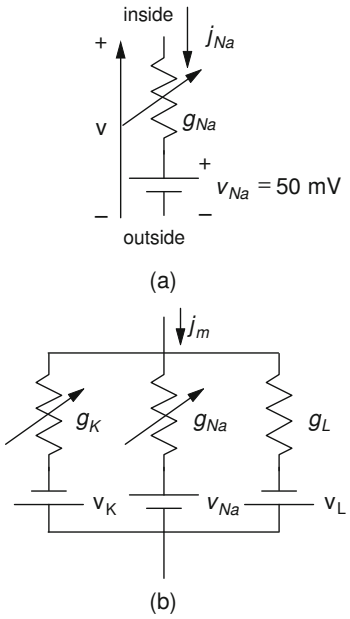


Fig. 6.33 Equivalent circuits for the membrane current. **a** The sodium current–voltage relationship of Eq. 6.61 is the same as that for a variable resistance in series with a battery at the sodium Nernst potential. **b** The total membrane current can be modeled with three such equivalent circuits. See the discussion of the sign of the potassium and leakage Nernst potentials in the text

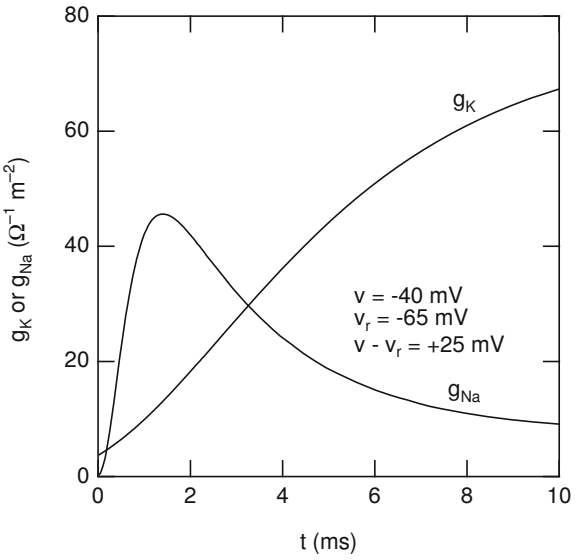


Fig. 6.34 The behavior of the sodium and potassium conductivities with time in a voltage-clamp experiment. At $t = 0$ the voltage was raised by 25 mV from the resting potential. The values are calculated from Eqs. 6.64–6.70 and are representative of the experimental data

For given concentrations, the sodium equilibrium or Nernst potential is

$$v_{Na} = \frac{k_B T}{e} \ln\left(\frac{[\mathrm{Na}_o]}{[\mathrm{Na}_i]}\right). \tag{6.60}$$

The sodium Nernst potential is usually about 50 mV. If $v = v_{Na}$ there is no current of sodium ions, regardless of the membrane permeability to sodium. If v is greater than v_{Na} (more positive), j_{Na} flows outward. If $v < v_{Na}$, the sodium current is inward. These currents can be described by

$$j_{Na} = g_{Na}(v - v_{Na}). \tag{6.61}$$

The coefficient g_{Na} is the sodium conductance per unit area. It is not constant but depends on the value of v and, in fact, on the past history of v. Defining the conductance this way makes the functional form of g_{Na} less complex; in particular, it does not have to change sign as v moves through v_{Na} and the sodium current reverses direction.

This equation can be multiplied by the membrane area to give a current–voltage relationship . Many authors draw a circuit diagram to represent the current flow through the membrane and along the axon. The sodium voltage–current relationship can be represented by a variable resistance corresponding to g_{Na} in series with a battery at the sodium Nernst potential, as shown in Fig. 6.33a.

An expression similar to Eq. 6.61 can be written for the potassium current density:

$$j_K = g_K(v - v_K). \tag{6.62}$$

The potassium Nernst potential is negative—about −77 mV—so the polarity of the potassium battery in Fig. 6.33b has been reversed. The leakage term will be considered later.

To summarize: v is the instantaneous voltage across the membrane. Both v_K and v_{Na} are constants depending on the relative ion concentrations inside and outside the cell and the temperature. The conductances per unit area depend on both the present value of v and its past history.

We can now describe the results of the voltage clamp experiments. The voltage in each experiment was changed from the resting value by an amount Δv. Therefore, $v - v_{Na}$ and $v - v_K$ had constant values after the change, and the changes in current density mirrored the changes in conductivity. Typical results for $\Delta v = 25$ mV and $T = 6\,^{\circ}\mathrm{C}$ are shown in Figs. 6.34 and 6.35. [The method of distinguishing sodium from potassium current is described in the original papers, or in Hille (2001, p. 39).] For a voltage clamp experiment the current and conductance have the same time variation. The sodium conductance rises from nearly zero and then falls, while the potassium conductance rises more slowly from a small initial resting value. (The potassium current before the voltage clamp was applied was small, because the resting potential was close to the potassium Nernst potential.) Measurements for longer times

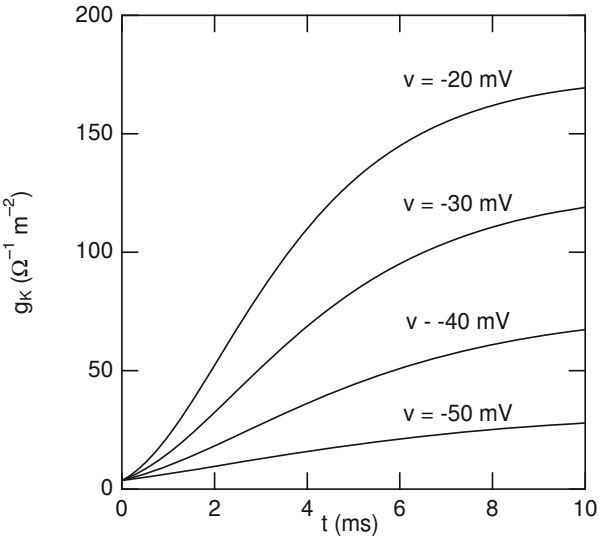


Fig. 6.35 The behavior of the potassium conductance for different values of the clamping voltage. These are representative curves calculated from Eqs. 6.64–6.66

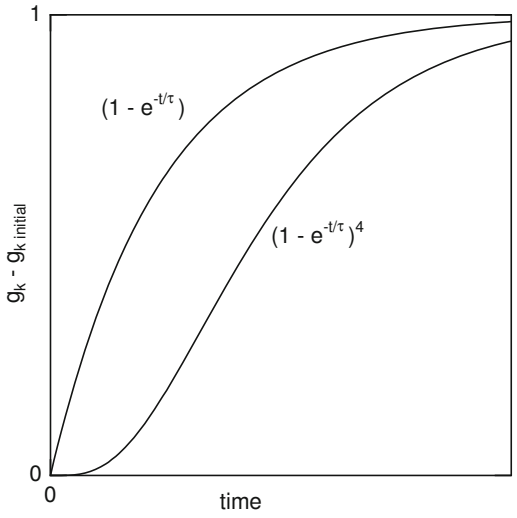


Fig. 6.36 A comparison of $(1 - e^{-t/\tau})$ with $(1 - e^{-t/\tau})^4$. The latter more closely approximates the shape of the potassium conductance in Fig. 6.34

show that the potassium conductivity rises to a constant value. Measurements for much longer times show that the potassium current falls after tens of milliseconds. For other values of Δv the conductance changes are different.

6.13.2 Potassium Conductance

Hodgkin and Huxley wanted a way to describe their extensive voltage-clamp data, similar to that in Figs. 6.34 and 6.35, with a small number of parameters. If we ignore the small nonzero value of the conductance before the clamp is applied, the potassium conductance curve of Fig. 6.34 is reminiscent of exponential behavior, such as $g_K(v,t) = g_K(v)(1-e^{-t/\tau(v)})$, with both $g_K(v)$ and $\tau(v)$ depending on the value of the voltage. A simple exponential is not a good fit. Figure 6.36 shows why. The curve $(1 - e^{-t/\tau})$ starts with a linear portion and is then concave downward. The potassium conductance in Figs. 6.34 and 6.35 is initially concave upward. The curve $(1 - e^{-t/\tau})^4$ in Fig. 6.36 more nearly has the shape of the conductance data. This suggests that we try to describe the conductance by

$$g_K(v,t) = g_{K\infty}\left[n_\infty(v)(1 - e^{-t/\tau(v)})\right]^N. \tag{6.63}$$

In this expression, $g_{K\infty}$ is the largest possible conductance per unit area. The value of $n_\infty(v)$ varies between 0 and 1 and determines the asymptotic value of the conductance change for a particular value of the voltage step. Hodgkin and Huxley found a good fit to their data with $N = 4$. If the initial value of the conductance were zero, our empirical fit to the potassium conductance data would be

$$g_K(v,t) = g_{K\infty}n^4(v,t), \tag{6.64a}$$

$$n(v,t) = n_\infty(v)(1 - e^{-t/\tau(v)}). \tag{6.64b}$$

But the initial potassium conductance was not zero. How should this be handled? Hodgkin and Huxley assumed that n is a measure of some fundamental property of the potassium channels, and that the conductance is always described by Eq. 6.64a. When the clamp voltage changes, the subsequent change of n is described by an exponential decay with the appropriate values of $n_\infty(v)$ and $\tau(v)$. If the initial value of n is n_0, the expression for $n(v,t)$ after the voltage clamp change is

$$n(v,t) = n_\infty(v)\left[1 - \left(\frac{n_\infty(v) - n_0}{n_\infty(v)}\right)e^{-t/\tau(v)}\right]. \tag{6.64c}$$

The function n is a solution to the differential equation

$$\frac{dn}{dt} = -\frac{n}{\tau} + \frac{n_\infty}{\tau}. \tag{6.65a}$$

Hodgkin and Huxley wrote this instead in the form

$$\frac{dn}{dt} = \alpha_n(1 - n) - \beta_n n. \tag{6.65b}$$

The subscript n on α_n and β_n distinguishes them from similar parameters for the sodium conductance.

The dependence of α_n and β_n on voltage is quite pronounced. With v in mV and α_n and β_n in ms^{-1}, the equations used by Hodgkin and Huxley to describe their experimental

values of α_n and β_n are

$$\alpha_n(v) = \frac{0.01\,[10-(v-v_r)]}{\exp\left(\dfrac{10-(v-v_r)}{10}\right)-1},$$
$$\beta_n(v) = 0.125\exp\left(\frac{-(v-v_r)}{80}\right). \tag{6.66}$$

The quantities α_n and β_n are rate constants in Eq. 6.65b. Like all chemical rate constants, they depend on temperature. The values above are correct when $T = 279$ K (6.3 °C). Hodgkin and Huxley assumed that the temperature dependence was described by a Q_{10} of 3. This means that the reaction rate increases by a factor of 3 for every 10 °C temperature rise. The rate at temperature T is obtained by multiplying rates obtained from Eq. 6.66 by

$$3^{(T-6.3)/10}. \tag{6.67}$$

For example, if the temperature is 18.5 °C, the rate must be multiplied by $3^{1.22} = 3.82$.

The variable n is often called the *potassium gate* or the *n gate*. It takes values between zero (a *closed* gate) and 1 (an *open* gate). The n gate is partially open at rest, making the resting membrane somewhat permeable to potassium. As v becomes more positive than the resting potential ("depolarizes"), the n gate opens further or "activates."

The behavior of α_n and β_n was determined from voltage-clamp experiments. In an actual nerve-conduction process, v is not clamped. Hodgkin and Huxley *assumed* that when v varies with time, the correct value of n can be obtained by integrating Eq. 6.65b. At each instant of time the values of α_n and β_n are those obtained from Eq. 6.66 for the voltage at that instant. This was a big assumption—but it worked. The value of $g_{K\infty}$ that they chose was 360 S m^{-2}.

6.13.3 Sodium Conductance

The sodium conductance was described by two parameters: one reproducing the rise and the other the decay of the conductance. The equation was

$$g_{Na} = g_{Na\infty} m^3 h.$$

The parameters m and h obeyed equations similar to that for n:

$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m, \tag{6.68}$$
$$\frac{dh}{dt} = \alpha_h(1-h) - \beta_h h. \tag{6.69}$$

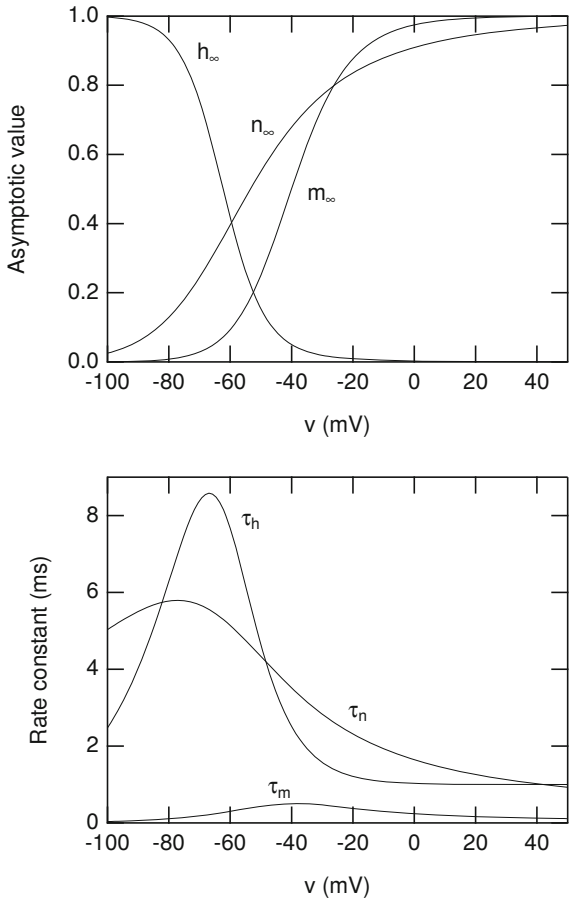


Fig. 6.37 Plots of the sodium and potassium conductance parameters versus the transmembrane potential

The v dependences were

$$\alpha_m = \frac{0.1[25-(v-v_r)]}{\exp\left(\dfrac{25-(v-v_r)}{10}\right)-1},$$
$$\beta_m = 4\exp\left(\frac{-(v-v_r)}{18}\right),$$
$$\alpha_h = 0.07\exp\left(\frac{-(v-v_r)}{20}\right), \tag{6.70}$$
$$\beta_h = \frac{1}{\exp\left(\dfrac{30-(v-v_r)}{10}\right)+1}.$$

These values for α and β are also for a temperature of 6.3 °C. The temperature scaling of Eq. 6.67 must be used for other temperatures. The value of $g_{Na\infty}$ is 1200 S m^{-2}. Figure 6.37 plots the time constants and asymptotic values as a function of membrane potential. These are the parameters for the equations in the form of Eq. 6.65a rather than Eq. 6.65b.

The variable m (called the *sodium activation gate* or *m gate*) is nearly closed at rest, preventing the resting membrane from being permeable to sodium. As v is depolarized m opens, allowing sodium to rush in. The sodium ions carry positive charge, so this inward current causes v to depolarize further (if there is not a voltage clamp), causing m to increase even more. This positive feedback (see Chap. 10) is responsible for the rapid upstroke of the action potential. The inward sodium current ends when v approaches the sodium Nernst potential, about 50 mV.

Variable h (the *sodium inactivation gate* or *h gate*) is different than the n- and m gates because it is open at rest but closes upon depolarization. However, it is slow compared to the m gate (see Fig. 6.37), so during an action potential it does not fully close until after the m gate has opened completely. Once the action potential is finished and v has returned to the resting value, the slow h gate takes a few milliseconds to completely re-open. During this time, the membrane cannot generate another action potential (it is *refractory*) because the closed h gate suppresses the sodium current.

6.13.4 Leakage Current

All other contributions to the current (such as movement of chloride ions) were lumped in the leakage term $j_L = g_L(v - v_L)$. The empirical value for g_L is 3 S m^{-2}. The parameter v_L was adjusted to make the total membrane current equal zero when $v = v_r$. For example, with the data given, zero current is obtained with $v_r = -65$ mV and $v_L = v_r + 10.6 = -54.4$ mV. The three contributions to the membrane current can be thought of as the circuit shown in Fig. 6.33b.

The Hodgkin–Huxley parameters have been used for a wide variety of nerve and muscle systems, even though they were obtained from measurements of the squid axon. A number of other models have since been developed that incorporate the sodium–potassium pump, calcium, etc. They have also been developed for various muscle and cardiac cells (Demir et al. 1994; Luo and Rudy 1994; Wilders et al. 1991).

6.14 Voltage Changes in a Space-Clamped Axon

A space-clamped axon has an interior potential $v(t)$ which does not depend on x. If such an axon is stimulated, a voltage pulse is observed. The first test we can make of the Hodgkin–Huxley model is to see if the parameters from the voltage-clamp experiments can also explain this pulse. To do so, it is necessary to insert Eq. 6.59, with all the other equations that are necessary to use it, in Eq. 6.52. Life is made somewhat simpler by the fact that the spatial derivative in Eq. 6.52 vanishes when the wire is in the axon. The result is

$$c_m \frac{\partial v}{\partial t} = -g_{Na}(v - v_{Na}) - g_K(v - v_K) - g_L(v - v_L). \quad (6.71)$$

When $v = v_r$ the right-hand side of this equation is zero and v does not change. It is necessary to introduce a stimulus to cause the pulse. This has been done in the computer program of Fig. 6.38, which solves Eq. 6.71. This program is not the most efficient that can be used; it has been written for ease of understanding. A stimulus of 10^{-4} A cm^{-2} = 1 A m^{-2} is applied between 0.5 and 0.6 ms. This is an additional term in Eq. 6.71, so that in the program, Eq. 6.71 becomes

```
dvdt = (-jMemb + jStim)/Cmemb;
```

In this statement `dvdt` stands for $\partial v/\partial t$, `jMemb` stands for j_m, `Cmemb` for c_m, and `jStim` for the stimulus current. The equation is solved by repeated application of the approximation

```
v = v + dvdt * deltat;
```

which stands for

$$v(t + \Delta t) = v(t) + \left(\frac{\partial v}{\partial t}\right) \Delta t.$$

The program uses $\Delta t = 10^{-6}$ s. The present value of v is used to calculate the rate constants in procedure `Calcab`. These are then used to calculate the present value of each conductivity. The membrane current is then calculated, and the entire process is repeated for the next time step. The results are tabulated in Fig. 6.39 and plotted in Fig. 6.40.

One can see from the plot that j_m is proportional to $\partial v/\partial t$. Note that although g_{Na} is a smooth curve, j_{Na} has an extra wiggle near $t = 2$ ms, caused by the rapid decrease in the magnitude of $v - v_{Na}$ as the voltage approaches the sodium Nernst potential. The initial depolarization is due to an inrush of sodium ions. But there is still a considerable sodium current during the potassium current. The sodium and potassium currents are nearly balanced throughout most of the pulse. The pulse lasts about 2 ms.

If the temperature is raised, the pulse is much shorter. Figure 6.41 shows the impulse when the temperature is 18.5 °C, calculated by multiplying each of the α and β values by $3^{(18.5-6.3)/10} = 3.82$.

The potassium current is not actually needed to create a nerve impulse because of the leakage current (primarily chloride) and the fact that the sodium conductance decreases after the initial depolarization. The potassium current speeds up the repolarization process. It is easy to modify the program of Fig. 6.38 to show this.

```
//program HodgkinHuxley
//Calculates Hodgkin-Huxley
//space clamped axon at  6.3°C

#include <stdio.h>
#include <math.h>

const float
  deltat =   1e-6,//deltat for integration
  tPStep =   1e-4,//printthis often
  vRest  = -65e-3,//Resting Potential
  Cmemb  =  1e-6,//Membrane capacitance
  tMax   =   5e-3,//Time to quit
  vNa    = 50e-3,//Sodium Nernst pot.
  vK    =-77e-3;//Potassium Nernst pot.

double
  n, m, hh,
  an, am, ah,
  bn, bm, bh,
  dndt, dmdt, dhdt, dvdt,
  gK, gNa,
  jK, jNa, jL, jMemb,
  voltage, t,
  jStim,    //Stimulus current
  tPrint;   //Time interval to print

void Calcab (void)
/*   Calculates the alpha and betas for
n, m, h, using the Hodgkin-Huxley eqns.
The original eqns. were in mV and ms;
these are in volts and seconds  */
  {
  an = (10*(-1000*(voltage-vRest)+10))
       /(exp((-1000*(voltage-vRest)+10)
       /10)-1);
  am = (100*(-1000*(voltage-vRest)+25))
       /(exp((-1000*(voltage-vRest)+25)
       /10)-1);
  ah = 70*exp(-1000*(voltage-vRest)
       /20);
  bn = 125*exp(-1000*(voltage-vRest)
       /80);
  bm = 4000*exp(-1000*(voltage-vRest)
       /18);
  bh = 1000/(exp((-1000*(voltage-vRest)
       +30)/10)+1);
  }
void Calc_Init_Values(void) //Calculates
// initial values of n, m, hh
  {
  Calcab();
  n = an/(an+bn);
  m = am/(am+bm);
  hh = ah/(ah+bh);
  }
```

```
void Calc_Curr(void)
  // Calculate conductances in siemens
  //per sq cm and current densities
  {
  gK  = 36e-3*pow(n, 4);
  gNa = 120e-3*pow(m, 3)*hh;
  jK  = gK *(voltage-vK);
  jNa = gNa*(voltage-vNa);
  jL  = 3e-4*(voltage-vRest-10.6e-3);
  jMemb = jK+jNa+jL;
  }
main(void)
  {
  //Print Table Headings
  printf("time    v       jMemb     gNa
    jNa       gK       jK     jL\n");
  t = 0;
  voltage = vRest;
  tPrint = 0;
  Calc_Init_Values();
  while (t < tMax) //Step through times
    {
    Calc_Curr();//Calc. membrane
          // current from conductances
    if (t >= tPrint)  //Print at
          //certain times
      {
      printf("%4.1f %1s %5.1f %1s ",
1000*t, "", 1000*voltage, "");
      printf("%8.2e %1s %8.2e %1s
%8.2e", jMemb, "", gNa, "", jNa );
      printf("%1s %8.2e %1s %8.2e %1s
%8.2e\n","", gK, "",  jK,"",jL);
      tPrint = tPrint+tPStep;
      }     // end if

    if ((t >= 5e-4) && (t < 6e-4))
    //Stimulus current at beginning
      jStim = 1e-4;
    else
      jStim = 0;//End stimulus current
    dvdt = (-jMemb+jStim)/Cmemb;
    voltage = voltage+dvdt*deltat;
    Calcab();//Calc alpha, beta
    dndt = an*(1-n)-bn*n;
    dmdt = am*(1-m)-bm*m;
    dhdt = ah*(1-hh)-bh*hh;
    n  = n+dndt*deltat;
    m  = m+dmdt*deltat;
    hh = hh+dhdt*deltat;
    t  = t+deltat;
    }       //end while
  }         //end main
```

Fig. 6.38 The computer program used to calculate the response of a space-clamped axon to a stimulus. The results are shown in Figs. 6.39 and 6.40

6.15 Propagating Nerve Impulse

If the wire is not inserted along the axon, the voltage changes in the x direction. A strong enough stimulus at one point results in a pulse that travels along the axon without change of shape. The basic equation that describes it is again Eq. 6.52 with the spatial term and with the Hodgkin–Huxley model for the membrane current:

$$\frac{\partial v}{\partial t} = -\frac{j_m}{c_m} + \frac{1}{2\pi a\, r_i c_m}\frac{\partial^2 v}{\partial x^2},$$
$$j_m = g_{Na}(v - v_{Na}) + g_K(v - v_K) + g_L(v - v_L). \tag{6.72}$$

These can be solved numerically by setting up arrays for values of v, n, m, and h at closely spaced discrete values of x along the axon. If index i distinguishes different values of x, then the discrete equation is

```
dvdt[i]=-jMemb[i]/Cmemb
        +(1/(6.28*a*ri*Cmemb*dx*dx))
        *(v[i+1]-2*v[i]+v[i-1]).
```

Figure 6.42 shows each term in Eq. 6.72 multiplied through by c_m to have the dimensions of current per unit area. The term

$$c_m \frac{\partial v}{\partial t}$$

is the rate at which charge per unit area on the membrane must change in order to change the membrane potential at the rate $\partial v/\partial t$,

$$-j_m = -g_{Na}(v - v_{Na}) - g_K(v - v_K) - g_L(v - v_L)$$

is the rate of charge buildup because of current through the membrane, and

$$\frac{1}{2\pi a\, r_i}\frac{\partial^2 v}{\partial x^2}$$

is the rate of charge buildup on the inner surface of the membrane because the longitudinal current is not uniform.

time	v	jMemb	gNa	jNa	gK	jK	jL
ms	mV	A/sq m	S/sq m	A/sq m	S/sq m	A/sq m	A/sq m
0.0	-65.0	-3.24e-06	1.06e-01	-1.22e-02	3.67e+00	4.40e-02	-3.18e-02
0.2	-65.0	-2.90e-06	1.06e-01	-1.22e-02	3.67e+00	4.40e-02	-3.18e-02
0.4	-65.0	-2.68e-06	1.06e-01	-1.22e-02	3.67e+00	4.40e-02	-3.18e-02
0.6	-55.3	5.75e-02	1.99e-01	-2.09e-02	3.74e+00	8.11e-02	-2.71e-03
0.8	-55.9	5.17e-03	7.08e-01	-7.50e-02	4.02e+00	8.48e-02	-4.59e-03
1.0	-55.6	-3.69e-02	1.19e+00	-1.25e-01	4.31e+00	9.22e-02	-3.58e-03
1.2	-54.5	-7.33e-02	1.69e+00	-1.77e-01	4.62e+00	1.04e-01	-2.79e-04
1.4	-52.6	-1.24e-01	2.46e+00	-2.52e-01	5.01e+00	1.22e-01	5.49e-03
1.6	-49.2	-2.30e-01	4.02e+00	-3.99e-01	5.52e+00	1.54e-01	1.56e-02
1.8	-42.2	-5.34e-01	8.59e+00	-7.92e-01	6.34e+00	2.21e-01	3.67e-02
2.0	-22.3	-1.73e+00	3.14e+01	-2.27e+00	8.15e+00	4.45e-01	9.62e-02
2.2	28.7	-2.08e+00	1.83e+02	-3.89e+00	1.48e+01	1.56e+00	2.49e-01
2.4	38.7	2.13e-01	3.17e+02	-3.60e+00	3.06e+01	3.53e+00	2.79e-01
2.6	32.0	4.18e-01	2.98e+02	-5.35e+00	5.05e+01	5.51e+00	2.59e-01
2.8	22.7	5.00e-01	2.50e+02	-6.81e+00	7.10e+01	7.08e+00	2.31e-01
3.0	12.5	5.19e-01	2.04e+02	-7.67e+00	8.93e+01	7.99e+00	2.01e-01
3.2	2.2	5.06e-01	1.65e+02	-7.91e+00	1.04e+02	8.25e+00	1.70e-01
3.4	-7.7	4.80e-01	1.32e+02	-7.63e+00	1.15e+02	7.97e+00	1.40e-01
3.6	-17.1	4.56e-01	1.04e+02	-6.97e+00	1.22e+02	7.31e+00	1.12e-01
3.8	-26.0	4.43e-01	7.93e+01	-6.03e+00	1.25e+02	6.39e+00	8.52e-02
4.0	-35.0	4.63e-01	5.71e+01	-4.85e+00	1.25e+02	5.26e+00	5.82e-02
4.2	-45.0	5.54e-01	3.55e+01	-3.38e+00	1.22e+02	3.90e+00	2.81e-02
4.4	-57.5	6.86e-01	1.45e+01	-1.56e+00	1.16e+02	2.25e+00	-9.43e-03
4.6	-70.0	4.64e-01	1.98e+00	-2.37e-01	1.07e+02	7.47e-01	-4.68e-02
4.8	-75.2	1.04e-01	8.09e-02	-1.01e-02	9.64e+01	1.76e-01	-6.23e-02

Fig. 6.39 Results of the calculation for a space-clamped axon at 6.3 °C

6.16 Myelinated Fibers and Saltatory Conduction

We have so far been discussing fibers without the thick myelin sheath. Unmyelinated fibers constitute about two-thirds of the fibers in the human body. They usually have radii of 0.05–0.6 μm. The conduction speed in m s^{-1} is given approximately by $u \approx 1800\sqrt{a}$, where a is the axon radius in meters.[12] (Strictly speaking, in this formula a should be replaced by the outer radius $a + b$ including the membrane thickness, but for an unmyelinated fiber $b \ll a$.)

Myelinated fibers are relatively large, with outer radii of 0.5–10 μm. They are wrapped with many layers of myelin between the nodes of Ranvier, as shown in Fig. 6.43. Typically, the outer radius is $a + b \approx 1.67a$ and the spacing between nodes is proportional to the outer diameter $D = 200(a + b) \approx 330a$ (See Prob. 69). These empirical proportionalities between node spacing and radius and between myelin thickness and radius will be very important to our understanding of the conduction speed. The conduction speed in a myelinated fiber is given approximately by $u \approx 12 \times 10^6(a + b) \approx 20 \times 10^6 a$. The conduction speeds of myelinated and unmyelinated fibers are compared in Fig. 6.44.

In the myelinated region the conduction of the nerve impulse can be modeled by electrotonus because the conductance of the myelin sheath is independent of voltage. At each node a regenerative Hodgkin–Huxley-type (HH-type) conductance change restores the shape of the pulse. Such conduction is called *saltatory conduction* because *saltare* is the Latin verb "to jump."

We saw that electrotonus is described by

$$\lambda^2 \frac{\partial^2 v}{\partial x^2} - v - \tau \frac{\partial v}{\partial t} = -v_r, \tag{6.73}$$

where the time constant is

$$\tau = \kappa \epsilon_0 \rho_m \tag{6.74}$$

and the space constant is

$$\lambda = \sqrt{\frac{ab\,\rho_m}{2\rho_i}}. \tag{6.75a}$$

The results of Problem 68 can be used to show that when the myelin thickness is appreciable compared to the inner axon radius, the space constant should be modified:

$$\lambda_{\text{thick}} = \sqrt{\frac{\ln(1 + b/a)\,\rho_m}{2\rho_i}}\, a. \tag{6.75b}$$

For a case in which $a = 5$ μm and $b = 3.3$ μm, the change is not very large. The thin membrane equation contains the quantity $ab = 17 \times 10^{-12}$ m^2 and the thick myelin equation contains $a^2 \ln(1 + b/a) = 12.8 \times 10^{-12}$ m^2.

We now want to understand the different dependence on radius of the conduction speed in the two kinds of fibers. We could do computer modeling for the unmyelinated fiber using Eq. 6.72 with axons of different radii, but this would not

[12] Values quoted in the literature range from $u = 1000\sqrt{a}$ (Plonsey and Barr 2007) to $u = 3000\sqrt{a}$ (Rushton 1951).

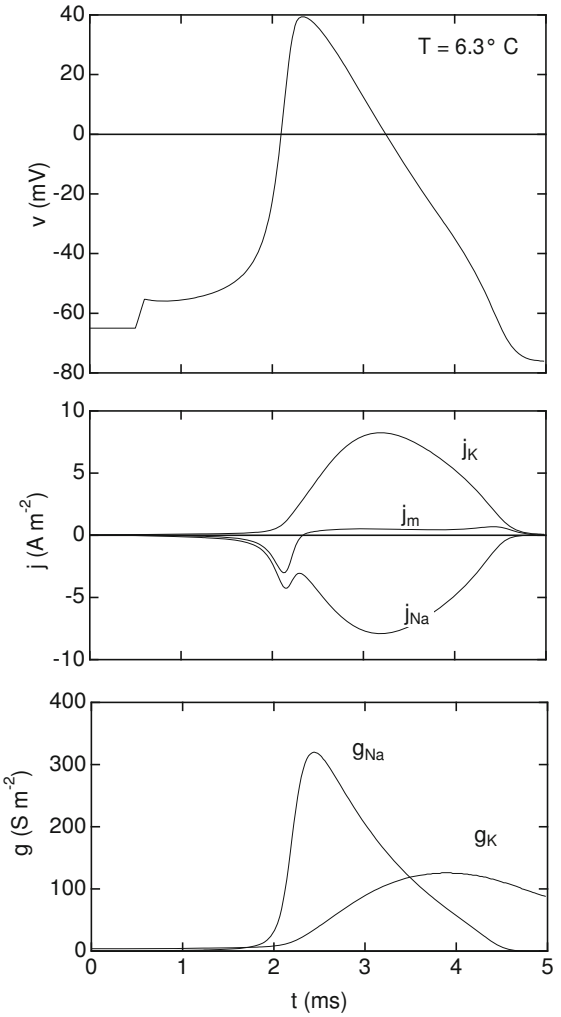


Fig. 6.40 A plot of the computation presented in Fig. 6.38 for a pulse in a space-clamped squid axon at $T = 6.3\,^{\circ}\mathrm{C}$. The axon was stimulated at $t = 0.5$ ms for 0.1 ms

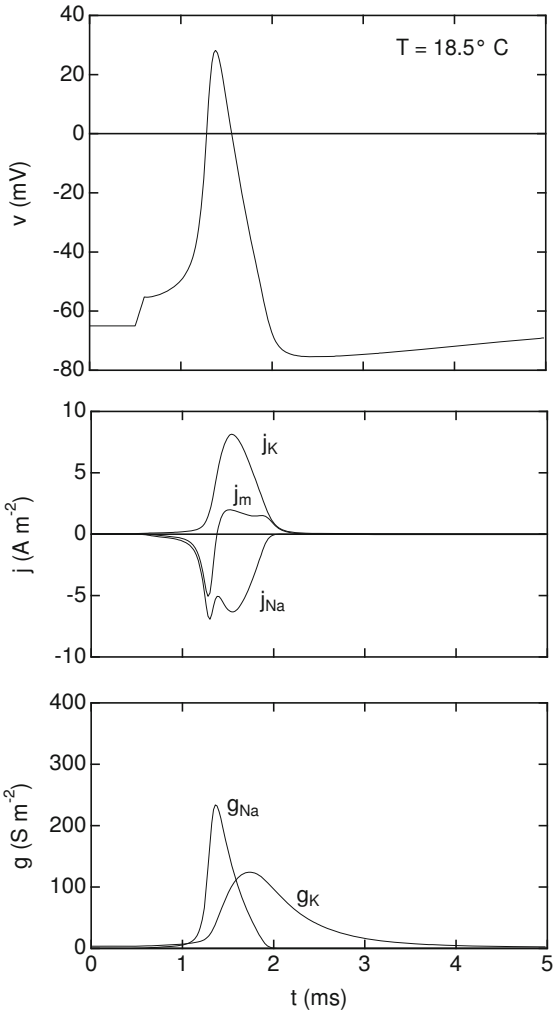


Fig. 6.41 A pulse in a space-clamped axon at 18.5 °C. The pulse lasts about 1 ms

provide an equation for $u(a)$. Rather than review the work that has been done (developing equations for the behavior of the foot of the action potential, for example), we will use a simple dimensional argument. This will not give an exact expression for $u(a)$, but it will indicate the functional form it must have.

In either the myelinated or the unmyelinated fiber the signal travels to neighboring regions by electrotonus, where it initiates HH-type membrane conductance changes. In the myelinated case the signal jumps from node to node; in the unmyelinated case the influence is on adjacent parts of the axon. When the neighboring region begins to depolarize, the HH change is much more rapid than that due to electrotonus. (Another way to say this is that during depolarization ρ_m and therefore τ become much smaller.) Therefore the conduction speed is limited by electrotonus. Regardless of the details of the calculation, the speed is proportional to the characteristic length in the problem divided by the characteristic time. For the unmyelinated case it is plausible to assume that the only characteristic length and time are λ and τ, so the speed is

$$u_{\text{unmyelinated}} \propto \frac{\lambda}{\tau} = \sqrt{\frac{b}{2\rho_i \rho_m}} \frac{1}{\kappa \epsilon_0} \sqrt{a}. \tag{6.76}$$

Since the membrane thickness for an unmyelinated fiber is always about 6 nm, this gives

$$u_{\text{unmyelinated}} \propto 270\sqrt{a} \tag{6.77}$$

as shown in Table 6.2.

For myelinated nerves the myelin thickness is $b \approx 0.67a$. This means that the space constant is proportional to a:

$$\lambda = \sqrt{\frac{ab\,\rho_m}{2\rho_i}} = \sqrt{\frac{0.67a^2\rho_m}{2\rho_i}} = a\sqrt{\frac{0.67\rho_m}{2\rho_i}} = 1750a. \tag{6.78}$$

Table 6.2 Properties of unmyelinated and myelinated axons of the same radius

Quantity	Unmyelinated	Myelinated
Axon inner radius a	5 μm	5 μm
Membrane thickness b'	6 nm	
Myelin thickness b		3.4 μm
$\kappa\epsilon_0$	6.20×10^{-11} s^{-1} Ω^{-1} m^{-1}	6.20×10^{-11} s^{-1} Ω^{-1} m^{-1}
Axoplasm resistivity ρ_i	1.1 Ω m	1.1 Ω m
Membrane (resting) or myelin resistivity ρ_m	10^7 Ω m	10^7 Ω m
Time constant $\tau = \kappa\epsilon_0\rho_m$	6.2×10^{-4} s	6.2×10^{-4} s
Space constant λ	$\lambda = \sqrt{\frac{ab\rho_m}{2\rho_i}}$ $= 0.165\sqrt{a}$ $= 370$ μm	$\lambda = \sqrt{\frac{ab\rho_m}{2\rho_i}} = \sqrt{\frac{0.67a^2\rho_m}{2\rho_i}}$ $= a\sqrt{\frac{0.67\rho_m}{2\rho_i}}$ $= 1750a$ $= 8.8$ mm
Node spacing D		$D = 340a = 1.7$ mm
Conduction speed from model	$u_{\text{unmyelinated}} \propto \lambda/\tau \approx 270\sqrt{a}$	$u_{\text{myelinated}} \propto \lambda/\tau \approx 2.9 \times 10^6 a$ or $D/\tau = 0.55 \times 10^6 a$
Conduction speed, empirical	$u_{\text{unmyelinated}} \approx 1800\sqrt{a}$	$u_{\text{myelinated}} \approx 17 \times 10^6 a$
Ratio of empirical to model conduction speed	6.7	5.9 or 31
Space constant using thick membrane model		$\lambda = a\sqrt{\frac{\ln(1+b/a)\rho_m}{2\rho_i}}$ $= a\sqrt{\frac{\ln(1.67)\rho_m}{2\rho_i}}$ $= 1530a$ $= 7.6$ mm

The spacing between the nodes, D, is about $340a$. There are two characteristic lengths for the myelinated case, both proportional to a because of the way the myelin is arranged. If we assume that the speed is proportional to D/τ, we obtain

$$u_{\text{myelinated}} \propto 0.55 \times 10^6 a. \tag{6.79}$$

If we assume that the speed is proportional to λ/τ, we obtain

$$u_{\text{myelinated}} \propto 2.9 \times 10^6 a. \tag{6.80}$$

Table 6.2 compares the space constants, time constants and conduction speeds for myelinated and unmyelinated fibers. The empirical expressions for the conduction speed are 7 or 8 times greater than what we estimate based on λ/τ. We might expect firing at the next node to occur when the signal has risen to about 10 % of its maximum value. This would reduce the time by about a factor of 10.

The internodal spacing is about 20 % of the space constant. Suppose that a constant current is injected at one node, as in Fig. 6.30. When the voltage has reached its full value at the next node it is given by

$$\frac{v}{v_0} = e^{-D/\lambda} = e^{-1.4/6.2} = 0.8.$$

If for some reason this node does not fire, the signal at the next node will be 0.64 of the original value, and so on. A local anesthetic such as procaine works by preventing permeability changes at the node. It is clear from this discussion that a nerve must be blocked over a distance of several nodes (a centimeter or more) in order for an anesthetic to be effective (Covino 1972).

6.17 Membrane Capacitance

The value of 7 for the dielectric constant, which has been used throughout this chapter, is considerably higher than the value 2.2, which is known for lipids. The inconsistency arises because part of the membrane is very easily polarized and effectively belongs to the conductor rather than to the dielectric; if the thickness of the lipid alone is considered in calculating the capacitance, then a value of 2.2 for κ is reasonable; if the entire membrane thickness is used, then the much higher dielectric constant for water and the polar groups within the membrane contributes, and $\kappa = 7$ is a reasonable value.

The easiest experiments to understand are those done with artificial bimolecular layers of lipid. The architecture of such a film is shown in Fig. 6.45. Each lipid molecule has a polar head and a hydrophobic tail. The molecules are arranged in a double layer with the heads in the aqueous solution. The dimensions in Fig. 6.45 are consistent with both measurements

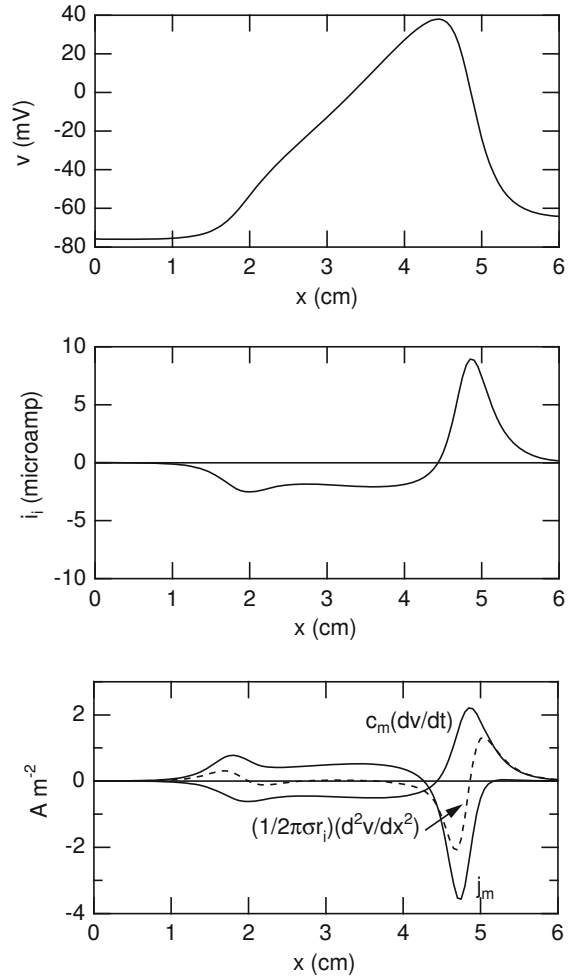


Fig. 6.42 A propagating pulse plotted against position along the axon at an instant of time. The middle graph shows the longitudinal current inside the axon. The bottom curve shows the current charging or discharging the membrane and the two terms comprising the right-hand side of Eq. 6.72

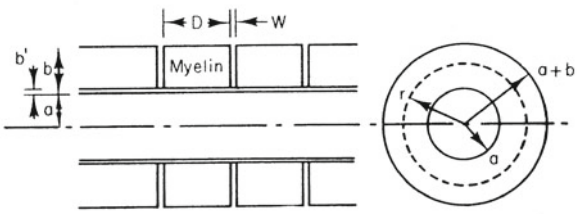


Fig. 6.43 The idealized structure of a myelinated fiber in longitudinal section and in cross section. The internodal spacing D is actually about 100 times the outer diameter of the axon

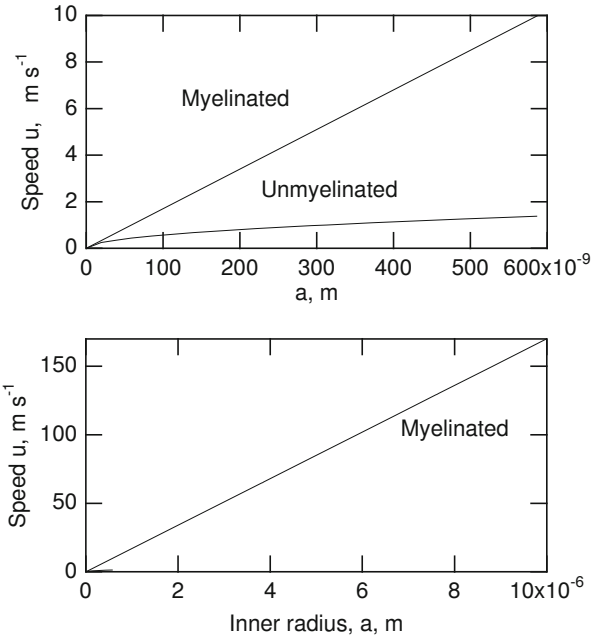


Fig. 6.44 The conduction speed versus the inner axon radius a for myelinated and unmyelinated fibers. Unmyelinated fibers with $a > 0.6$ μm are not found in the body

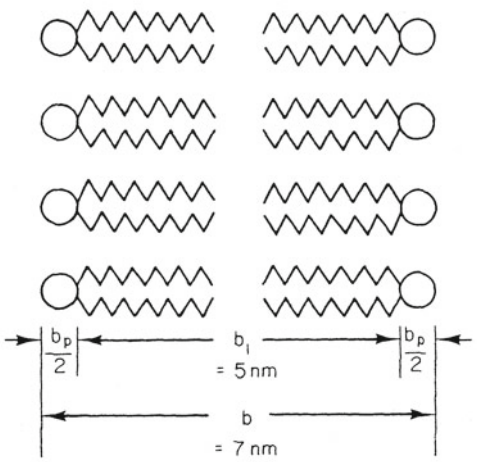


Fig. 6.45 Structure of a bimolecular lipid membrane

of the film thickness and with the known structure of the lipid molecules. Linear aliphatic hydrocarbons have a bulk dielectric constant of about 2. The polar heads have a much higher dielectric constant, probably about 50. Water has a dielectric constant of about 80.

The capacitance per unit area of bimolecular lipid films is about 0.3×10^{-2} F m^{-2} (0.3 μF cm^{-2}). The simplest way to explain this value is to assume that the polar heads are part of the surrounding conductor. The capacitance per unit area is then

$$\frac{C}{S} = \frac{\kappa\epsilon_0}{b_1} = \frac{(2.2)(8.85 \times 10^{-12})}{5 \times 10^{-9}} = 0.4 \times 10^{-2}\ \text{F m}^{-2}.$$

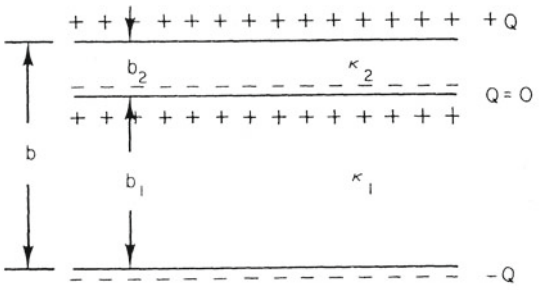


Fig. 6.46 A membrane composed of two phases. The ith phase has thickness b_i and dielectric constant κ_i. The total thickness is b and the effective dielectric constant is κ. The charges shown are external charge; polarization of the dielectric is not shown but determines the value of κ

A more sophisticated approach is to regard the membrane as made up of three layers: polar, lipid, polar. The same effect can be obtained by considering two layers with all the polar component lumped together, as in Fig. 6.46. Suppose that we put charge $+Q$ on one surface and $-Q$ on the other surface of the membrane. We put no charge on the interface between layers 1 and 2. The charge of zero on the interface can be thought of as a superposition of positive and negative charges as shown in Fig. 6.46. We are referring only to external charge which we place on the membrane; the charges induced by polarization of the dielectric are not shown. They are taken into account by the value of κ. The situation is that of two parallel-plate capacitors in series. Each layer has a capacitance C_i: $Q = C_i v_i = \kappa_i \epsilon_0 S v_i / b_i$. The total potential across the membrane is $v = v_1 + v_2 = Q/C$. The total capacitance is

$$C = \frac{Q}{v_1 + v_2} = \frac{Q}{Qb_1/\kappa_1\epsilon_0 S + Qb_2/\kappa_2\epsilon_0 S} = \frac{1}{b_1/\kappa_1\epsilon_0 S + b_2/\kappa_2\epsilon_0 S}. \tag{6.81}$$

The effective dielectric constant is obtained by equating the total capacitance to $\kappa\epsilon_0 S/b$:

$$\kappa = \frac{b}{b_1/\kappa_1 + b_2/\kappa_2}. \tag{6.82}$$

Application of these equations to the bimolecular lipid membrane (with $\kappa_1 = 2.2$, $\kappa_2 = 50$, $b_1 = 5$ nm, $b_2 = 2$ nm) gives

$$\kappa = 3.0, \qquad \frac{C}{S} = 0.38 \times 10^{-2}\ \mathrm{F\,m^{-2}}. \tag{6.83}$$

The capacitance per unit area is nearly that obtained by assuming the polar groups are perfect conductors.

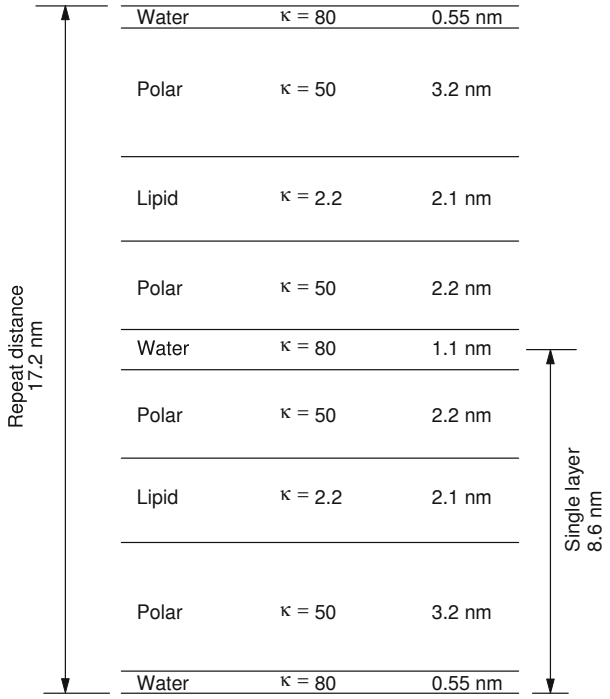


Fig. 6.47 The results of x-ray diffraction measurements of the structure of myelin surrounding frog sciatic nerve. Data are adapted from Worthington (1971, p. 35).

The myelin surrounding a nerve fiber consists of several layers wrapped tightly together. Each repeating layer is made up of two single layers back to back. The best data on the structure of these layers are from x-ray diffraction experiments. The layers repeat every 17 nm. One model for the structure within a repeat distance is shown in Fig. 6.47 (Worthington 1971, p. 35). A single layer of the myelin has a thickness of 8.55 nm. A surprising feature of this model is that the lipid layer is less than half the thickness of that in a bilayer lipid membrane. However, the measured capacitance of a nerve-cell membrane or myelin is greater than for the bilayer lipid membrane; if one is to keep the lipid value for κ, the membrane must be thinner. It is gratifying that the membrane thickness as measured by x-ray diffraction is consistent with the observed membrane capacitance.

To check the consistency, note that Eqs. 6.81 and 6.82 are easily extended to more than two phases. Use the following data:

	κ_i	b_i (nm)
Water	80	2.2
Lipid	2.2	4.2
Polar	50	10.8

With these values, the effective dielectric constant is

$$\kappa = \frac{17.2}{4.2/2.2 + 2.2/80 + 10.8/50} = 7.95.$$

If we assume that the membrane on an unmyelinated axon has the same structure as a half-unit of the myelin, then the thickness is 8.55 nm. With a dielectric constant of 7.95, the capacitance per unit area is calculated to be 0.82×10^{-2} F m^{-2}. The measured value is 1.0×10^{-2} F m^{-2}.

When one begins to look at the detailed structure of the membrane as we have done in this section, there is no justification for using the same membrane thickness b for the capacitance and the conductance of the membrane. The capacitance is determined primarily by the thickness of the lipid portion of the membrane; the conductance includes the effect of ions passing through the polar layers. The product, $\kappa\rho$, of the previous section is meaningful only for a membrane that is homogeneous and has the same thickness for both capacitive and conductive effects.

As long as the membrane structure is not being considered, it is safer to express such things as attenuation along the axon in terms of the directly measured parameters: length and time constants. Nonetheless, a preliminary formulation in terms of a homogeneous membrane model can be useful to start thinking about the problem.

6.18 Rhythmic Electrical Activity

Many cells exhibit rhythmic electrical activity. Various nerve transducers produce impulses with a rate of firing that depends on the input to which the transducer is sensitive. The beating of the heart is controlled by the *sinoatrial node* (SA node) that produces periodic pulses that travel throughout the heart muscle.

The mechanism for such repetitive activity is similar to what we have seen in the Hodgkin–Huxley model, though the details of the ionic conductance variations differ. The computer program of Fig. 6.38 can easily be modified to model rhythmic activity. Figure 6.48 shows a plot of the output of a modified program. The only modification was to make `jStim` be a constant leakage current of 0.2 A m^{-2} (0.2×10^{-4} A cm^{-2}). This provides the essential feature: a small inward current between beats that causes the potential inside the cell to increase slowly. When the voltage exceeds a certain threshold, the membrane channels open and the cell produces another impulse.

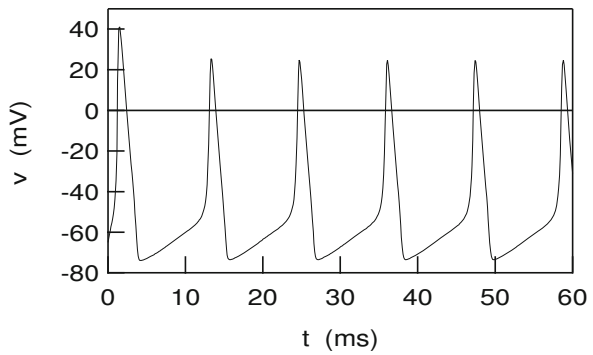


Fig. 6.48 By changing the leakage current, it is possible to make the Hodgkin–Huxley model display periodic electrical activity

While this simple change produces repetitive firing, and in fact the shape of the curve in Fig. 6.48 is very similar to that measured in the SA node, the details of ionic conduction are actually very different. The SA node contains no sodium channels. The rapid depolarization is due to an inward calcium current. There are a number of contributions to the current in the SA node, and detailed ionic models of them have been described (Demir et al. 1994; Noble 1989, 1995; Wilders et al. 1991).The slow leakage is a complicated combination of currents, the details of which are still not completely understood (Anumonwo and Jalife 1995; DiFrancesco et al. 1995).

6.19 The Relationship Between Capacitance, Resistance, and Diffusion

There is a fundamental relationship between the capacitance and resistance between two conductors in a homogeneous conducting dielectric. It is also possible to develop an analogy between capacitance and steady-state diffusion, so that known expressions for the capacitance of conductors in different geometries can be used to solve diffusion problems.

6.19.1 Capacitance and Resistance

Consider two conductors carrying equal and opposite charge and embedded in an insulating medium with dielectric constant κ. The potential difference between the conductors is Δv, and the magnitude of the charge on each is $Q = C\Delta v$. The electric field is $\mathbf{E}(x, y, z)$. In a vacuum Gauss's law applied to a surface surrounding the positively charged conductor gives $\iint \mathbf{E} \cdot d\mathbf{S} = Q/\epsilon_0$. Polarization in a dielectric surrounding the conductor reduces the electric field by a factor of κ. If $\mathbf{E}$ refers to the electric field in the dielectric and Q to the charge on the conductor, Gauss's law becomes

$$\iint \mathbf{E} \cdot d\mathbf{S} = Q/\kappa\epsilon_0. \tag{6.84}$$

For a given charge on the conductor, the presence of the dielectric reduces $\mathbf{E}$ and Δv by $1/\kappa$ and, therefore, increases the capacitance by κ.

Suppose that the dielectric is not a perfect insulator but obeys Ohm's law and has conductivity σ ($\mathbf{j} = \sigma\mathbf{E}$). If

some process maintains the magnitude of the charge on each conductor at Q, the current leaving the positive conductor is

$$i = \iint \mathbf{j} \cdot d\mathbf{S} = \sigma \iint \mathbf{E} \cdot d\mathbf{S} = \sigma\, Q/\kappa\epsilon_0. \tag{6.85}$$

The resistance between the conductors is

$$R = \frac{\Delta v}{i} = \frac{Q/C}{\sigma\, Q/\kappa\epsilon_0} = \frac{\kappa\epsilon_0}{\sigma C}. \tag{6.86}$$

This inverse relationship between the resistance and capacitance is independent of the geometry of the conductors, as long as the dielectric constant and conductivity are uniform throughout the medium.

If the charge on the conductors is not replenished, it leaks off with a time constant $\tau = RC = \kappa\epsilon_0/\sigma$. We have seen this result earlier in several special cases; we now understand that it is quite general.

6.19.2 Capacitance and Diffusion

In Chap. 4, we saw that the transport equations for particles, heat, and electric charge all have the same form. We now develop an analogy between these transport equations and the equations for the electric field. The analogy is useful because it relates the diffusion of particles between different regions to the electrical capacitance between conductors with the same geometry; the electrical capacitance in many cases is worked out and available in tables.

Fick's first law of diffusion was developed in Chap. 4, Eq. 4.20:[13]

$$\mathbf{j}_s = -D\nabla c. \tag{6.87}$$

The relationship between fluence rate (particle current density) and particle flux (current) is

$$\iint \mathbf{j}_s \cdot d\mathbf{S} = i_s, \tag{6.88}$$

where i_s is the current of particles out of the volume enclosed by the surface. This equation is very similar to Gauss's law,

$$\iint_{\text{surface}} \mathbf{E} \cdot d\mathbf{S} = \frac{q}{\kappa\epsilon_0}, \tag{6.89}$$

where q is the electric charge. The electric potential and the electric field are related by the three-dimensional version of Eq. 6.16:

$$\mathbf{E} = -\nabla v. \tag{6.90}$$

[13] In this section, we will use c for concentration of solute particles and C for the electrical capacitance.

The similarity between Eqs. 6.90 and 4.20 and between 6.5 and 6.89 suggests that we make the substitutions

$$\begin{aligned} i_s &\longleftrightarrow \frac{q}{\kappa\epsilon_0}, \\ c &\longleftrightarrow \frac{v}{D}, \\ \mathbf{j}_s &\longleftrightarrow \mathbf{E}. \end{aligned} \tag{6.91}$$

For any electrostatic configuration in which there are two equipotential surfaces containing charge $+q$ and $-q$, there is an analogous diffusion problem in which there is a flow of particles from one surface to another, each surface having a constant concentration on it. In the electrical case, the charge and potential are related by the capacitance, which is a geometric property of the two equipotential surfaces: $q = C\,\Delta v$. An analogous statement can be made for diffusion between two surfaces of fixed concentration:

$$i_s = -\frac{C\,\Delta v}{\kappa\epsilon_0} = -\frac{C}{\kappa\epsilon_0}\, D\,\Delta c. \tag{6.92}$$

We can find the rate of flow of particles if we know the diffusion constant, the concentration difference, and the capacitance for the electrical problem with the same geometry. To see the utility of this method, we will consider some cases of increasing geometrical complexity.

As a first example, suppose that two concentric spheres have radii a and b. You can show (from the work in Problem 16, for example) that the capacitance of this configuration is

$$\frac{C}{\kappa\epsilon_0} = \frac{4\pi}{1/a - 1/b}. \tag{6.93}$$

As $b \to \infty$, this becomes

$$\frac{C}{\kappa\epsilon_0} = 4\pi a. \tag{6.94}$$

This can be applied to diffusion to or from a spherical cell of radius a. If the diffusion is outward, as of waste products, imagine that the outward flow rate is i_s and that the concentration difference between the cell surface and infinity is c_0. Then

$$i_s = -4\pi a\, D\, c_0. \tag{6.95}$$

If, on the other hand, the concentration infinitely far away is greater than that at the cell surface by an amount c_0, the number of particles in the cell will increase at a rate

$$i_s = +4\pi a\, D\, c_0. \tag{6.96}$$

These results were obtained directly in Chap. 4.

As another example, consider a circular disk of radius a with the other electrode infinitely far away. It is more difficult

to calculate the capacitance in this case, but we can look it up (Smythe et al. 1957). It is $C/\kappa\epsilon_0 = 8a$. But this is the capacitance for the charge on *both sides* of the disk; the lines of **E** and **j** go off in both directions. We want only half of this, since we will use the result to calculate the end correction for a pore. (If we were concerned with diffusion to a disk-shaped cell, we would use the whole thing.) For the half-space

$$i_{s\,\mathrm{half}} = -4Da\,\Delta c \tag{6.97}$$

is proportional to the radius of the disk, not its area.

Still another geometrical situation that may be of interest is the diffusion of particles from one sphere of radius a to another sphere of radius a, when the centers of the spheres are separated by a distance b.

The capacitance between two such spherical electrodes is (Smythe et al. 1957, pp. 5–14)

$$\frac{C}{\kappa\epsilon_0} = 2\pi a \sinh\beta \left(\frac{1}{\sinh\beta} + \frac{1}{\sinh 2\beta} + \frac{1}{\sinh 3\beta} + \cdots\right),$$

where $\cosh\beta = b/2a$. This formula is written in terms of the *hyperbolic functions*

$$\begin{aligned}\sinh\beta &= \tfrac{1}{2}\left(e^{\beta} - e^{-\beta}\right),\\ \cosh\beta &= \tfrac{1}{2}\left(e^{\beta} + e^{-\beta}\right).\end{aligned} \tag{6.98}$$

When the spheres are far apart $b/2a \to \infty$, and $\cosh\beta \approx \frac{1}{2}e^{\beta}$, $\sinh\beta \approx \frac{1}{2}e^{\beta}$. In that limit,

$$\begin{aligned}\frac{C}{\kappa\epsilon_0} &= 2\pi a\left(\tfrac{1}{2}e^{\beta}\right)\left(\frac{1}{\frac{1}{2}e^{\beta}} + \frac{1}{\frac{1}{2}e^{2\beta}} + \frac{1}{\frac{1}{2}e^{3\beta}} + \cdots\right)\\ &= 2\pi a e^{\beta}\left(e^{-\beta} + e^{-2\beta} + e^{-3\beta} + \cdots\right)\\ &= 2\pi a\left[1 + (a/b) + (a/b)^2 + \cdots\right].\end{aligned} \tag{6.99}$$

The diffusive flow between two spheres is therefore

$$i_s = -2\pi a\, D\,\Delta c \tag{6.100}$$

if they are sufficiently far apart. Note that this is just one-half of the flow from a sphere of radius a to a concentric sphere infinitely far away. The earlier results in this section show that the electrical resistance between two spherical electrodes sufficiently far apart is $1/2\pi\sigma a$.

Symbols Used in Chap. 6

Symbol	Use	Units	First used page
a	Distance	m	145
a	Axon inner radius	m	157
a	Radius of spherical ion or cell	m	173
a	Radius of disk	m	173
b, c	Distance	m	145
b	Membrane thickness	m	142
b	Myelin thickness	m	167
b'	Membrane thickness at node of Ranvier	m	168
b	Sphere radius	m	173
c	Concentration	m^{-3}	142
c_i, c_o	Ion concentrations	m^{-3}; $\mathrm{mol\,l}^{-1}$	155
c_m	Membrane capacitance per unit area	$\mathrm{F\,m}^{-2}$	157
e	Electronic charge	C	155
g_{Na}, g_K, g_m, g_L	Membrane conductance per unit area	$\mathrm{S\,m}^{-2}$	157
$g_{Na\infty}, g_{K\infty}$	Asymptotic membrane conductance per unit area	$\mathrm{S\,m}^{-2}$	163
h, h_∞	Parameters used to describe sodium conductance		164
i	Electric current	A	151
i_i	Currents along inside of axon	A	158
i_m	Current through a section of membrane	A	156
i_s	Solute current or flux	s^{-1}	173
j, **j**	Current per unit area	$\mathrm{A\,m}^{-2}$	151
j_m	Membrane current per unit area	$\mathrm{A\,m}^{-2}$	157
j_{Na}, j_K, j_L	Membrane current per unit area for that species	$\mathrm{A\,m}^{-2}$	161
k_B	Boltzmann's constant	$\mathrm{J\,K}^{-1}$	155
m, m_∞	Parameters used to describe sodium conductance		164
n, n_∞	Parameters used to describe potassium conductance		163
p, **p**	Dipole moment	C m	175
q	Electric charge	C	143
q_{bound}, q_{free}	Bound and free charge	C	150
r, **r**	Distance	m	143
r_i	Resistance per unit length along inside of axon	$\Omega\,\mathrm{m}^{-1}$	157
t	Time	s	142
u	Propagation velocity of a wave or signal	$\mathrm{m\,s}^{-1}$	167
v	Potential difference	V	142
v	$v_i - v_o$	V	158
v_K, v_{Na}	Equilibrium (Nernst) potential for potassium, sodium	V	162
v_r	Resting membrane potential	V	159
x, y, z	Distance	m	146
z	Valence of ion		155
C	Capacitance	F	149

C_m	Membrane capacitance	F	156
D	Length of myelinated segment	m	167
D	Diffusion constant	$m^2\ s^{-1}$	173
$\mathbf{E}$, E_x, E_y, E_z	Electric field and components	$V\ m^{-1}$	144
E_p	Electric field due to polarization charge	$V\ m^{-1}$	149
E_e, E_{ext}	External electric field	$V\ m^{-1}$	149
$\mathbf{E}_{tot}$	Total electric field	$V\ m^{-1}$	149
$\mathbf{F}$	Force	N	143
$\mathbf{F}_{ext}$	External force	N	147
G	Conductance	Ω^{-1} or S	151
G_m	Conductance of a section of axon membrane	Ω^{-1} or S	157
L	Length of cylinder or axon segment	m	145
$[Na_i]$, $[Na_o]$	Sodium concentrations inside and outside an axon	m^{-3}	161
P	Power	W	153
Q	Electric charge	C	149
Q_{10}	Factor by which the rate of a chemical reaction increases with a temperature rise of 10 K		164
R	Resistance	Ω	151
R_i	Internal resistance along a segment of axon	Ω	157
R_m	Resistance across a segment of membrane	Ω	156
S, ΔS, dS	Surface area	m^2	144
T	Temperature	K	155
U	Potential energy	J	147
W	Work	J	151
α_m, β_m, α_n, β_n, α_h, β_h	Rate parameters for Hodgkin–Huxley model	s^{-1}	163
β	Dimensionless variable		174
ϵ_0	Electrical permittivity of free space	$N^{-1}\ m^{-2}\ C^2$	143
κ	Dielectric constant		150
λ	Charge per unit length	$C\ m^{-1}$	145
λ	Electrotonus spatial decay constant	m	159
ρ	Resistivity	Ω m	152
ρ_i	Resistivity of axoplasm	Ω m	157
ρ_m	Resistivity of membrane	Ω m	156
θ	Angle		146
σ	Charge per unit area	$C\ m^{-2}$	145
σ	Conductivity	$S\ m^{-1}$	152
χ	Electrical susceptibility		149
τ	Time constant	s	156
τ	Electrotonus time constant	s	160
τ_h, τ_m, τ_n	Time constants in Hodgkin-Huxley model	s	164

Problems

Section 6.1

Problem 1. Suppose that an action potential in a 1-μm diameter unmyelinated fiber has a speed of 1.3 $m\,s^{-1}$. Estimate how long it takes a signal to propagate from the brain to a finger. Repeat the calculation for a 10-μm diameter myelinated axon that has a conduction speed of 85 $m\,s^{-1}$. Speculate on the significance of these results for playing the piano.

Problem 2. The median nerve in your arm has a diameter of about 3 mm . If the nerve consists only of 1 μm-diameter unmyelinated axons, how many axons are in the nerve? (Ignore the volume occupied by extracellular space.) Repeat the calculation for 20 μm outer diameter myelinated axons. Repeat the calculation for 0.5 mm diameter unmyelinated axons (about the size of a squid axon). Speculate on why higher animals have myelinated axons instead of larger unmyelinated axons.

Section 6.2

Problem 3. Two equal and opposite charges $\pm q$ separated by a distance a form a dipole. The *dipole moment* $\mathbf{p}$ is a vector pointing in the direction from the negative charge to the positive charge of magnitude $p = qa$. In electrochemistry the dipole moment is often expressed in debyes: 1 debye (D) $= 10^{-18}$ electrostatic units (statcoulomb cm) (1 statcoulomb $= 3.3356 \times 10^{-10}$ C).

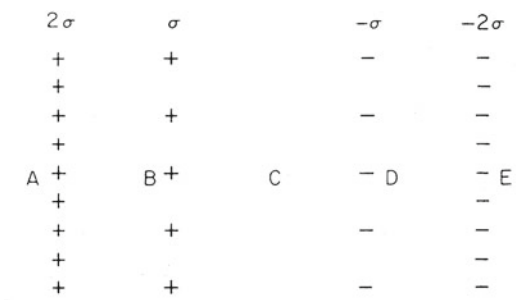
(a) Find the relationship between the debye and the SI unit for the dipole moment.

(b) Express the dipole moment of charges $\pm 1.6 \times 10^{-19}$ C separated by 2×10^{-10} m in debyes and in the appropriate SI unit.

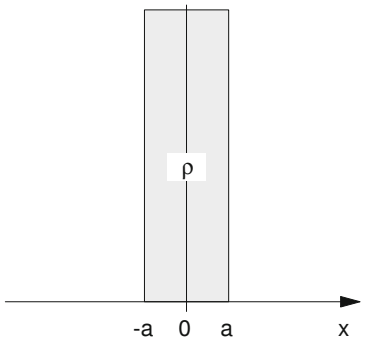
Problem 4. The electric field of a dipole can be calculated by assuming the positive charge q is at $z = a/2$ and the negative charge $-q$ is at $z = -a/2$ ($x = y = 0$). The electric field along the z axis is found by vector addition of the electric field from the individual charges using Eq. 6.3. Find an expression for the electric field. (Hint: $1/(1+x)^2$ is approximately equal to $1-2x$ for small x.) By what power of z does the electric field fall off?

Section 6.3

Problem 5. Use the principle of superposition to calculate the electric field in regions A, B, C, D, and E in the figure.



Problem 6. An infinite sheet of charge has a thickness $2a$ as shown. The charge density is ρ C m^{-3}. Find the electric field for all values of x.



Problem 7. Derive Eq. 6.10 from Eq. 6.9. At some point in your derivation you may need to use the substitution $u = y/\sqrt{c^2+y^2+z^2}$ and the integrals

$$\int \frac{dx}{(x^2+a^2)^{3/2}} = \frac{x}{a^2\sqrt{x^2+a^2}}$$

$$\int \frac{dx}{x^2+a^2} = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right).$$

Problem 8. Show that Eq. 6.10 reduces to Eq. 6.8 when $z \ll b, c$. Show that Eq. 6.10 is consistent with Coulomb's law when $z \gg b, c$.

Section 6.4

Problem 9. Show that N C^{-1} is equivalent to V m^{-1}.

Problem 10. Use Coulomb's law and $v = -\int_\infty^x E_x dx$ to determine the potential along the x axis due to a point charge. Assume that $v(x = \infty) = 0$. Because there is no preferred direction in space, the potential in any radial direction from the charge has the same form.

Problem 11. Try to apply the equation $v(r) = -\int_\infty^r E_r dr$ to the equation for the electric field of a line of charge, Eq. 6.7. Why does it not work?

Section 6.5

Problem 12. A person stands near a high-voltage power line. Assume for this problem that its voltage is not changing with time. Since much of a person's body is an ionic solution, treat the body as a conductor and the surrounding air as an insulator. In a static situation, what is the electric field inside the person's body caused by the power line? (Hint: Think before you calculate.)

Section 6.6

Problem 13. Two plane parallel conducting plates each have area S and are separated by a distance b. One carries a charge $+Q$; the other carries a charge $-Q$. Neglect edge effects.

(a) What is the charge per unit area on each plate? Where does it reside?
(b) What is the electric field between the plates?
(c) What is the capacitance?
(d) As the plate separation is increased what happens to E, v, and C?
(e) If a dielectric is inserted between the plates, what happens to E, v, and C? (See Sect. 6.7.)

Problem 14. It was shown in the text that the electric field from an infinitely long line of charge, of charge density λ C m^{-1}, is $E = \lambda/2\pi\epsilon_0 r$ at a distance r from the line.

(a) Show that if positive charge is distributed with density σ C m^{-2} on the surface of a cylinder of radius a, the electric field is

$$\begin{array}{ll} 0, & r < a \\ \sigma a/\epsilon_0 r, & r > a. \end{array}$$

(b) Find the potential difference between a point a distance a from the center of the cylinder and a point a distance d from the center of the cylinder ($d > a$).
(c) Is a or d at the higher potential?
(d) Suppose that another hollow cylinder of radius $d > a$ is placed concentric with the first. It has a charge $-\sigma'$ per unit area. How will its presence affect the potential difference calculated in part (b)?
(e) Calculate the capacitance between the two cylinders and show that it is $2\pi\epsilon_0 L/\ln(d/a)$, where L is the length of the cylinder.

Problem 15. Problem 14 showed that the capacitance of a pair of concentric cylinders, of radius a and d ($d > a$) is $2\pi\epsilon_0 L/\ln(d/a)$. Suppose now that $d = a + b$, where b is the thickness of the region separating the two cylinders. (It might, for example, be the thickness of the axon membrane.) Use the fact that $\ln(1+x) = x - x^2/2 + x^3/3 + \cdots$ to show that, for small b (that is, $b \ll a$), the formula for the capacitance becomes the same as that for a parallel-plate capacitor.

Problem 16. Find the capacitance of two concentric spherical conducting shells. The inner sphere has radius a and the outer sphere has radius b.

Section 6.7

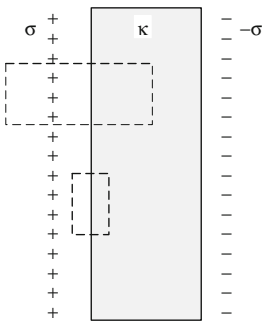
Problem 17. A parallel-plate capacitor has area S and plate separation b. The region between the plates is filled with dielectric of dielectric constant κ. The potential difference between the plates is v.

(a) What is the total electric field in the dielectric?
(b) What is the magnitude of the charge per unit area on the inner surface of the capacitor plates?
(c) What is the magnitude of the polarization charge on the surface of the dielectric?

Problem 18. For the unmyelinated axon of Table 6.1 and Fig. 6.3,
(a) How many sodium, potassium, and miscellaneous anions are there in a 1-mm segment?
(b) How many water molecules are there in a 1-mm segment?
(c) What is the charge per unit area on the inside of the membrane?
(d) What fraction of all the atoms and ions inside the segment are charged and not neutralized by neighboring ions of the opposite charge?

Problem 19. A nerve-cell membrane has a layer of positive charge on the outside and negative charge on the inside. These charged layers attract each other. The potential difference between them is $v = 70$ mV. Assuming a dielectric constant $\kappa = 5.7$ for the membrane, an axon radius of 5 μm, and a membrane thickness $b = 5$ nm, what is the force per unit area that the charges on one side of the membrane exert on the other? Express the answer in terms of b, v, and κ. (Hint: The force is calculated by multiplying the charge in a given layer by the electric field due to the charge in the other layer. Think carefully about factors of 2.)

Problem 20. The drawing represents two infinite plane sheets of charge with an infinite slab of dielectric filling part of the space between them. The dashed lines represent cross sections of two Gaussian surfaces. The sides are parallel to the electric field, and the ends are perpendicular to the electric field. Apply the second form of Gauss's law, Eq. 6.21b, to find the electric field within the dielectric using the upper Gaussian surface. Repeat using the lower Gaussian surface.



Section 6.8

Problem 21. This problem will give you some insight into the resistance of electrodes used in neurophysiology. Consider two concentric spherical electrodes. The region between them is filled with material of conductivity σ. The inner radius is a, the outer radius is b.
(a) Imagine that there is a total charge Q on the inner sphere. Find the electric field between the spheres in terms of the potential difference between them and their radii.
(b) The current density in the conducting material is given by $\mathbf{j} = \sigma\mathbf{E}$. Find the total current.
(c) Find the effective resistance, $R = v/i$. What is the resistance as $b \to \infty$? This is the resistance of a small spherical electrode in an infinite medium; infinite means the other electrode is "far away."

Problem 22. Patients undergoing electrosurgery sometimes suffer burns around the perimeter of the electrode. Wiley and Webster (1982) analyzed the potential produced by a circular disk electrode of radius a and potential v_0 in contact with a medium of conductivity σ. They found that the normal component of current density at the surface of the electrode is given by

$$j_n = \frac{2\sigma v_0}{\pi}\frac{1}{(a^2 - r^2)^{1/2}}, \qquad 0 < r < a.$$

(a) Calculate the total current I coming out of the electrode.
(b) Determine the resistance of the electrode.
(c) Plot j_n vs. r. Use the plot to explain why the patients suffer burns near the edge of the electrode.

Problem 23. The *Coulter counter* or resistive pulse technique is used to count and size particles in a wide variety of applications (Kubitschek 1969; DeBlois and Bean 1970), including the automated counting of blood cells. The cells being counted are assumed to be nonconducting and immersed in a conducting fluid. The fluid is made to flow through a narrow channel. When a suspended particle enters the channel there is a change in resistance. Assume a long channel of radius b with no end effects.
(a) What is the resistance of pure fluid of resistivity $\rho = 1/\sigma$ in a segment of channel of length $2a$?
(b) A cylindrical non-conducting cell of radius a and length $2a$ is in the channel. Its axis and the axis of the channel coincide. What is the resistance of a segment of channel of length $2a$? Ignore end effects.
(c) Show that the resistance change (the difference between these two reults) is proportional to the volume of the cell, $V = 2\pi a^3$, and inversely proportional to b^4.

Section 6.9

Problem 24. Derive the equation for the resistance of a set of resistors connected (a) in series and (b) in parallel.

Section 6.10

Problem 25. The resting concentration of calcium ions, $[Ca^{++}]$, is about 1 mmol l^{-1} in the extracellular space but is very low (10^{-4} mmol l^{-1}) inside muscle cells. Determine the Nernst potential for calcium. Is calcium in equilibrium at a resting potential of -70 mV?

Problem 26. In our analysis of the electric field in the cell membrane, we assume the charge on the membrane can be represented as a continuous distribution of surface charge. For a 6-nm thick membrane this will be a good approximation if the number of discrete charges in a 6-nm square patch of membrane is large.

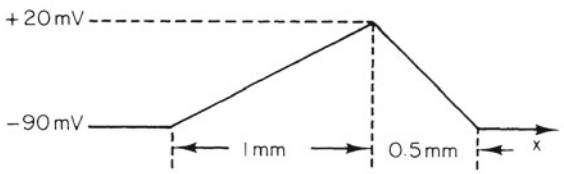
(a) Estimate how many discrete charged ions are present on a 6-nm square patch of membrane in a resting cell. Does the charge distribution appear to be continuous or discrete?

(b) Assume the ion has a diffusion constant in water of 10^{-9} m^2 s^{-1} and calculate the time required for the ion to diffuse 6 nm. If averaged over 1 ms, a time characteristic of neural activity, does the charge distribution appear continuous or discrete?

Section 6.11

Problem 27. The resistivity of the fluid within an axon is 0.5 Ω m. Calculate the resistance along an axon 5 mm long with a radius of 5 μm. Repeat for a radius of 500 μm.

Problem 28. The voltage along an axon is as shown at some instant of time. The axon radius is 10 μm; the resistivity of the axoplasm is 0.5 Ω m. What is the longitudinal current in the axon as a function of position?

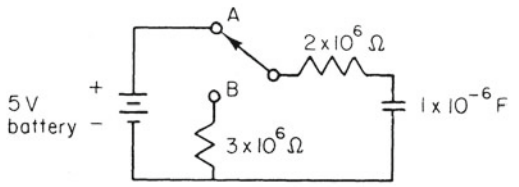


Problem 29. This problem is designed to show you how a capacitance, such as the cell membrane, charges and discharges. To begin, the switch has been in position B for a long time, so that there is no charge on the capacitor. At $t = 0$ the switch is put in position A. It is kept there for 20 s, then thrown back to position B.

(a) Write a differential equation for the voltage on the capacitor as a function of time when the switch is in position A and solve it.

(b) Repeat when the switch is in position B.

(c) Plot your results.



Problem 30. Sometimes an organ is lined with a single layer of flat cells. (One example is the lining of the *jejunum*, the upper portion of the small intestine.) Experimenters can then apply a time-varying voltage across the sheet of cells and measure the resulting current. The cells are packed so tightly together that one model for them is two layers of insulating membrane of dielectric constant κ and thickness b that behave like a capacitor, separated by intracellular fluid of thickness a and resistivity ρ. Find a differential equation or integral equation that relates the total voltage difference across the layer of cells $v(t)$ to the current per unit area through the layer, $j(t)$, in terms of κ, ρ, b, a.

Problem 31. The current that appears to go "into" a section of membrane is made up of two parts: that which charges the membrane capacitance and that which is a leakage current through the membrane: $i = v/R + C(dv/dt)$. Suppose that the total current is sinusoidal: $i = I_0 \cos \omega t$.

(a) Show that the voltage must be of the form $v = I_0 R' \cos \omega t + I_0 X \sin \omega t$ and that the differential equation is satisfied only if

$$R' = \frac{R}{1 + \omega^2 (RC)^2},$$

$$X = R \frac{\omega (RC)}{1 + \omega^2 (RC)^2}.$$

(b) What happens to R' and X as $\omega \to 0$? $\omega \to \infty$? For what value of ω is X a maximum? What is the corresponding value of R'? Plot these points.

(c) Your plot in part b should suggest that the locus of X vs R' is a semicircle, centered at $X = 0$, $R' = R/2$. Prove that this is so. [Remember that the equation of a circle is $(x - a)^2 + (y - b)^2 = r^2$.]

Section 6.12

Problem 32. Consider the myelinated and unmyelinated axons of Tables 6.1 and 6.2. Compare the decay distance for electrotonus in both cases. Neglect attenuation due to the leakage at the node of Ranvier.

Problem 33. Show by direct substitution that $v(x) = v_0 e^{-x/\lambda} + v_r$ satisfies the equation

$$\frac{d^2 v}{dx^2} = 2\pi a \, g_m r_i (v - v_r)$$

if v_r is constant.

Problem 34. In an electrotonus experiment a microelectrode is inserted in an axon at $x = 0$, and a constant current i_0 is injected. After the membrane capacitance has charged, the voltage outside is zero everywhere and the voltage inside is given by Eq. 6.58:

$$v - v_r = \begin{cases} v_0 e^{-x/\lambda}, & x > 0 \\ v_0 e^{x/\lambda}, & x < 0. \end{cases}$$

(a) Find $i_i(x)$ in terms of v_0, λ, and r_i.
(b) Find $j_m(x)$ in terms of g_m, v_0, and λ.
(c) Find the current i_0 injected at $x = 0$ in terms of v_0, λ, and r_i.
(d) Find the input resistance v_0/i_0.

Problem 35. The cable equation is $\lambda^2(\partial^2 v/\partial x^2) - v - \tau(\partial v/\partial t) = 0$. Let $v(x,t) = w(x,t)\exp(-t/\tau)$. Substitute this expression into the cable equation and determine a new differential equation for $w(x,t)$. You should find that $w(x,t)$ obeys the diffusion equation (Chap. 4). Find the diffusion constant in terms of the axon parameters and evaluate it for a typical case.

Problem 36. The voltage along an axon when a constant current is injected at $x = 0$ for all times $t > 0$ is given by Hodgkin and Rushton (1946)

$$v(x,t) - v_r = \frac{v_0}{2}\left\{e^{-|x|/\lambda}\left[1 - \operatorname{erf}\left(\frac{|x|}{2\lambda}\sqrt{\frac{\tau}{t}} - \sqrt{\frac{t}{\tau}}\right)\right] - e^{|x|/\lambda}\left[1 - \operatorname{erf}\left(\frac{|x|}{2\lambda}\sqrt{\frac{\tau}{t}} + \sqrt{\frac{t}{\tau}}\right)\right]\right\}$$

where the error function $\operatorname{erf}(y)$ and its derivatives are

$$\operatorname{erf}(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-z^2}dz$$

$$\frac{d}{dy}\operatorname{erf}(y) = \frac{2}{\sqrt{\pi}}e^{-y^2}.$$

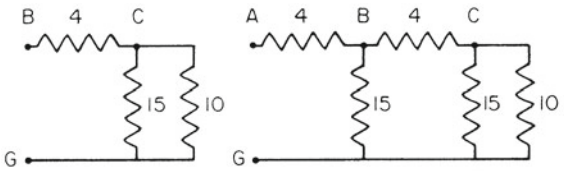
(a) Show that the expression for $v(x,t)$ obeys the cable equation, Eq. 6.55.
(b) Use $\operatorname{erf}(0) = 0$, $\operatorname{erf}(-\infty) = -1$, and $\operatorname{erf}(\infty) = 1$ to show that as $t \to \infty$, the expression for $v(x,t)$ approaches the solution in Eq. 6.58 and Fig. 6.30.
(c) Find a simple expression for $v(x,t)$ when $x = 0$. Use $\operatorname{erf}(1) = 0.843$ and $\operatorname{erf}(0.5) = 0.520$ to check that this expression is consistent with the plots in Fig. 6.31.

Problem 37. Consider a space-clamped axon with a membrane time constant τ. Initially ($t \le 0$), $v' = 0$. From $t = 0$ until a time $t = d$ a stimulus is applied to the membrane. Assume that when $v' < V'$ the membrane behaves passively (V' is called the threshold potential), and when $v' > V'$, an action potential will fire. v' obeys the equation $dv'/dt = -v'/\tau + s$.

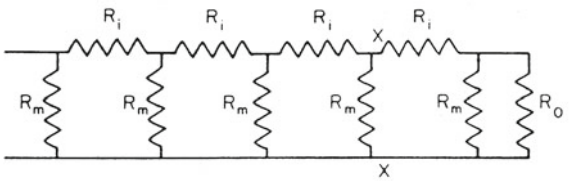
(a) Find v' for $0 < t < d$ and for $t > d$. Note that v' is maximum for $t = d$.
(b) Find an expression for $v(t = d)$, and then solve it for s.
(c) Plot s as a function of the pulse duration d. This plot is called the *strength–duration curve*.
(d) Find the value of s that corresponds to threshold stimulation for very long durations, in terms of V' and τ. This value of s is called the *rheobase* stimulus.
(e) Find the value of d corresponding to threshold stimulation using a stimulus strength of twice rheobase. This duration is called *chronaxie*.
(f) Find an expression for τ in terms of chronaxie. Measuring the strength–duration curve is one way to determine the membrane time constant.

Problem 38. An alternative model to the cable equation is an attenuating network of resistors and capacitors. This problem is designed to show you how a "ladder" of resistances can attenuate a signal.

(a) Show that the resistance between points B and G in the circuit on the left is 10 Ω.
(b) Show that the resistance between points A and G in the circuit on the right is also 10 Ω. What will be the result if an infinite number of ladder elements are added to the left of AG?
(c) Assume that v_C (measured with respect to point G) is 6 V. Calculate v_B and v_A. Note that the ratios are the same: $v_B/v_A = v_C/v_B$.



Problem 39. This is a more general version of the previous problem, which can be applied directly to electrotonus when capacitance is neglected. Consider the ladder shown, which represents an axon. R_0 is the effective resistance between the inside and outside of the axon to the right of the section under consideration. The axon has been divided into small slices; R_i is the resistance along the inside of the axon in the small slice, and R_m is the resistance across the membrane in the slice. The resistance outside the axon is neglected. Note that the resistance looking into the axon to the right of points XX is also R_0.



(a) Show that R_0 is given by a quadratic equation: $R_0^2 - R_i R_0 - R_i R_m = 0$ and that the solution is

$$R_0 = \frac{1}{2}\left[R_i + (R_i^2 + 4R_i R_m)^{1/2}\right].$$

(b) Show that the ratio of the voltage across one ladder rung to the voltage across the immediately preceding rung is

$$\frac{R_m R_0}{R_m R_0 + R_m R_i + R_i R_0}.$$

(c) Now assume that $R_i = r_i dx$ and $R_m = 1/(2\pi a g_m dx)$. Calculate R_0 and the voltage ratio. Show that the voltage ratio (as $dx \to 0$) is

$$\frac{1}{1 + (2\pi a r_i g_m)^{1/2} dx}.$$

(d) The preceding expression is of the form $1/(1 + x)$. For sufficiently small x, this is approximately $1 - x$. Therefore, show that the voltage change from one rung to the next is $dv = -\left[(2\pi a\, r_i g_m)^{1/2} dx\right] v$ so that v obeys the differential equation

$$\frac{dv}{dx} = -(2\pi a r_i g_m)^{1/2} v.$$

Section 6.13

Problem 40. Use the Hodgkin–Huxley parameters to answer the following questions.

(a) When $v = v_r$, what are α_n and β_n?

(b) Show that $dn/dt = 0$ when $n = 0.318$. What is the resting value of g_K?

(c) At $t = 0$ the voltage is changed to -25 mV and held constant. Find the new values of α_n, β_n, n_∞, τ_n and the asymptotic value of g_K.

(d) Find an analytic solution for $n(t)$. Plot n and n^4 for $0 < t < 10$ ms.

Problem 41. Calculate the values of the gates m, n, and h for the resting membrane ($v = -65$ mV), using the Hodgkin and Huxley model. Recall that at rest, $m = m_\infty(v = -65$ mV), etc.

Problem 42. If α_n and β_n depend on temperature according to Eq. 6.67, how do n_∞ and τ_n depend on temperature?

Problem 43. Calculate the resting membrane conductance per unit area for the resting membrane, using the Hodgkin and Huxley model. Hint: $j_m = 0$ at rest. Let $v = v_r + dv$, where dv is small. Determine the steady-state j_m as a function of dv. To keep things simple, ignore any changes to m, n, and h resulting from dv.

Problem 44. In a voltage-clamp experiment, a wire of radius b is threaded along the interior of an axon of radius a. Assume the axoplasm displaced by the wire is pushed out the end so that the cross-sectional area of the axon containing the wire remains πa^2. The resistivities of wire and axoplasm are ρ_w and ρ_a. Find the wire radius needed so that voltage changes along the axon are reduced by a factor of 100 from what they would be without the wire. Ignore the electrode surface impedance.

Problem 45. A wire of resistivity $\rho_w = 1.6 \times 10^{-8}\ \Omega$ m and radius $w = 0.1$ mm is threaded along the exact center of an axon segment of radius $a = 1$ mm, length $L = 1$ cm, and resistivity $\rho_i = 0.5\ \Omega$ m. The axon membrane has conductance $g_m = 10$ S m^{-2}. Find numerical values for

(a) the resistance along the wire,

(b) the resistance of the axoplasm from the wire to the membrane, and

(c) the resistance of the membrane.

Problem 46. If the voltage across an axon membrane is changed by 25 mV as in Fig. 6.34, how long will it take for all the potassium to leak out if it continues to move at the constant rate at which it first leaks out? Use the asymptotic value for the potassium conductance from Fig. 6.34. Use Table 6.1, and Fig. 6.3 for any other values you need.

Section 6.14

Problem 47. Use the data of Fig. 6.40 to answer the following questions about a nerve impulse in a squid axon of radius $a = 0.1$ mm.

(a) Estimate the peak sodium ion flux (ions m^{-2} s^{-1}) and the total number of sodium ions per unit area that pass through the membrane in one pulse.

(b) By what fraction does the sodium concentration in the cell increase during one nerve pulse?

(c) Estimate the peak potassium flux and total potassium transport.

Problem 48. Show by direct substitution that Eq. 6.64c satisfies the equation $dn/dt = \alpha_n(1 - n) - \beta_n n$ if α_n and β_n are functions of v, but not of time.

Problem 49. The Hodgkin–Huxley equation for the potassium parameter n is $dn/dt = \alpha_n(1 - n) - \beta_n n$. What is the asymptotic value of n when $t \to \infty$?

Problem 50. For $t < 0$ a squid axon has a resting membrane potential of -65 mV. The sodium Nernst potential is $+50$ mV. The Hodgkin–Huxley parameters are $m = 0.05$, $h = 0.60$, and $g_{Na} = 1200$ S m^{-2}.

(a) What is j_{Na}?

(b) For $t > 0$ a voltage clamp is applied so that $v = -30$ mV. Suppose that $m = 0.72 - 0.67e^{-2.2t}$ and $h = 0.6e^{-0.63t}$ (where t is in milliseconds). What is the total charge transported across unit area of the membrane by sodium ions?

Problem 51. Consider a 1-mm long segment of a squid nerve axon, with a diameter of 1 mm.

(a) Let the intracellular sodium ion concentration be 15 mmol l^{-1}. Calculate the number of sodium ions in this segment of the axon.
(b) Use the plot of j_{Na} versus time in Fig. 6.41 to estimate the total number of sodium ions that enter the axon during the action potential (if you have to determine the area under the j_{Na} curve, just estimate it).
(c) Find the ratio of the number of sodium ions entering the axon in one action potential to the number present in the resting axon. Does a single action potential change the intracellular concentration of sodium ions significantly?
(d) What diameter axon is needed in order for the intracellular sodium ion concentration to change by 10 % during one action potential?

Problem 52. A stimulating current of 1 A m^{-2} is applied for 100 μs. How much does it change the potential across the membrane?

Problem 53. Using the resting value of j_K from Fig. 6.39, calculate how long it would take for the concentration of potassium inside an axon of radius 100 μm to decrease by 1 %.

Problem 54. Modify the program in Fig. 6.38 to calculate the values of m, h, and n as functions of time during an action potential. Plot $m(t)$, $h(t)$, $n(t)$, and $v(t)$.

Problem 55. Modify the program in Fig. 6.38 so it uses different stimulus strengths other than $j_{\text{stim}} = 1$ A m^{-2}. Find the minimum value of j_{stim} that results in an action potential. This value is known as the *threshold stimulus strength*.

Problem 56. Modify the program in Fig. 6.38 so it applies two stimulus pulses. The first is of strength $j_1 = 1$ A m^{-2}, duration 0.5 ms, and starts at $t_1 = 0$. The second is of strength j_2, duration 0.5 ms, and starts at time t_2. For a given t_2 value, determine the threshold stimulus strength j_2. Plot the threshold j_2 as a function of the interval $t_2 - t_1$, for 1 ms $< t_2 - t_1 <$10 ms. This plot is called a *strength–interval curve*. The increase of threshold j_2 for small intervals reflects the refractoriness of the membrane.

Problem 57. When a squid nerve axon is *hyperpolarized* by a stimulus (the transmembrane potential is more negative than resting potential) for a long time and then released, the transmembrane potential drifts back towards resting potential, overshoots v_r and becomes more positive than v_r, and eventually reaches threshold and fires an action potential. This process is called *anode-break excitation*: anode because the membrane is hyperpolarized, and break because the excitation does not occur until after the stimulus ends. Modify the program in Figure 6.38, so that the stimulus lasts 3 ms, and the stimulus strength is -0.15 A m^{-2}. Show that the program predicts anode break stimulation. Determine the mechanism responsible for anode break stimulation. Hint: pay particular attention of the sodium inactivation gate (the h gate). You may want to plot h versus time to see how it behaves.

Problem 58. Consider a space-clamped axon for which the resting potential is v_r. Assume that the membrane current density follows a very strange behavior:

$$j_m = \begin{cases} B(v - v_r)^2, & v > v_r \\ 0, & v < v_r. \end{cases}$$

(a) Write a differential equation for $v(t)$.
(b) What are the units of B?
(c) What sign would B have for depolarization to take place after a small positive change of v?
(d) Integrate the equation obtained in (a).

Problem 59. A comment was made in the text that the potassium current is not required to generate an action potential. Modify the program of Fig. 6.38 to eliminate the potassium current. (First make sure that you have an unmodified program that reproduces Fig. 6.39 correctly.) Comment on the shape of the resulting pulse. After the pulse there is a new value of the resting potential. Why? Is it significant?

Section 6.15

Problem 60. A pulse which propagates along the axon with speed u is of the form $v(x, t) = f(x - ut)$.
(a) Use the chain rule to show that this means

$$\frac{\partial v}{\partial t} = -u\frac{\partial v}{\partial x}, \quad \frac{\partial^2 v}{\partial t^2} = u^2\frac{\partial^2 v}{\partial x^2}.$$

(b) Find an expression for the membrane current per unit area in terms of c_m, ρ_i, ρ_m, a, and the various partial derivatives of f with respect to x.

Problem 61. Consider an action potential propagating along an axon. The "foot" of the action potential is that part of the initial rise of the transmembrane potential that occurs before the sodium channels open. Starting from Eq. 6.72, set the j_m equal to zero and assume that the action potential propagates with a uniform speed u. As in Problem 60, replace the spatial derivatives with temporal derivatives and show that the transmembrane potential during the foot of the action potential rises exponentially. Find an expression for the time constant of this exponential rise in terms of r_i, c_m, a, and u.

Problem 62. An unmyelinated axon has the following properties: radius of 0.25 mm, membrane capacitance of 0.01 F m^{-2}, resistance per unit length along the axon of 2×10^6 Ω m^{-1}, and propagation velocity of 20 m s^{-1}. The propagating pulse passes an observer at $x = 0$. The peak of the pulse can be approximated by a parabola, $v(t) = 20(1 - 10t^2)$, where v is in millivolts and t is in milliseconds.
(a) Find the current along the axon at $x = 0$, $t = 0$.
(b) Find the membrane current per unit area j_m at $x = 0$, $t = 0$.

Problem 63. A space-clamped axon (v independent of distance along axon) has a pulse of the form

$$v(t) - v_r = \begin{cases} 0, & t < -t_1 \\ v_0\left[1 - (t/t_1)^2\right], & -t_1 < t < t_1 \\ 0, & t > t_1, \end{cases}$$

as shown in Problem 62. The axon has radius a, length L, resistivity ρ_i, and membrane capacitance c_m per unit area.

(a) What is the total change in charge on the membrane from $t = -t_1$ to $t = 0$?
(b) What is the total change in charge on the membrane from $t = -t_1$ to $t = +t_1$?
(c) What is $j_m(t)$?
(d) If j_m is given by $g_m(v - v_r)$, what is $g_m(t)$? Comment on its behavior.

Problem 64. Modify the program in Fig. 6.38 to include x-dependence as outlined in the text. Reproduce Fig. 6.42 and determine the propagation speed. Use $r_i = 19.89 \times 10^5\ \Omega\ \mathrm{m}^{-1}$ and $a = 0.238$ mm.

Section 6.16

Problem 65. Saltatory conduction is often described as the action potential jumping from node to node. In one sense this is correct: the nodes of Ranvier are active patches connected by passive myelinated segments. However, one should realize that the upstroke of the action potential is spread out over many nodes. Use Table 6.2, an action potential upstroke lasting 0.5 ms, and an outer diameter of 20 μm to estimate the number of nodes over which the action potential extends. If many nodes contribute simultaneously to the excitation, should propagation be considered continuous or discrete?

Problem 66. Consider a myelinated fiber in which the nodes of Ranvier are spaced every 2 mm. The resistance of the axoplasm per unit length is $r_i = 1.4 \times 10^{10}\ \Omega\ \mathrm{m}^{-1}$. The nodal capacitance is about 1.5×10^{-12} F.

(a) If the voltage difference between nodes is 10 mV, what is the current along the axon? Assume that the voltages are not changing with time, so that the membrane charge does not change. Also neglect leakage current through the membrane.
(b) If the nerve impulse rises from −70 mV to +30 mV in 0.5 ms, what is the average current required to charge the nodal capacitance?

Problem 67. A myelinated cylindrical axon has inner radius a and outer radius b. The potential inside is v. Outside it is 0. The myelin is too thick to be treated as a plane sheet of dielectric. Express all answers in terms of a, b, and v.

(a) Give an expression for E for $r < a$.
(b) Give an expression for E for $a < r < b$.
(c) Give an expression for E for $r > b$.
(d) Assuming $\kappa = 1$, what is the charge density on the inner surface? The outer surface?

Problem 68. Develop equations for the resistance and capacitance of a cylindrical membrane whose thickness is appreciable compared to its inner radius. Use Gauss's law for cylindrical symmetry to determine the electric field. Consider total charge Q distributed uniformly over the inner surface of a section of the membrane of length D and inner radius a. The membrane has dielectric constant κ.

(a) Any charge on the outer surface of the membrane has no effect on the calculation of the electric field between $r = a$ and $r = a + b$ as long as the charge is distributed uniformly on the outer cylindrical surface at $r = a + b$. Show that the electric field within the membrane is $E = \left(\frac{1}{4\pi\epsilon_0\kappa}\right) 2Q/Dr$.
(b) Show that the potential difference is $v = v(a) - v(a + b) = \frac{Q}{2\pi\epsilon_0\kappa D} \ln(1 + b/a)$, and that the capacitance is

$$C = \frac{2\pi\kappa\epsilon_0 D}{\ln(1 + b/a)} \quad \text{(cylinder)}.$$

(c) Now place a conducting medium with resistivity $\rho = 1/\sigma$ in the region of the membrane. Charge will move. It will be necessary to provide a battery to replenish it. Show that the resistance of the membrane segment of length D is given by $R = \frac{\rho}{2\pi D} \ln(1 + b/a)$, so that

$$\rho = \frac{2\pi R D}{\ln(1 + b/a)} \quad \text{(cylinder)}.$$

(d) Show that the resistivity of a plane resistor of cross sectional area $2\pi a D$ and thickness b is

$$\rho = \frac{2\pi R D}{b/a} \quad \text{(plane)},$$

and that the capacitance of this plane section of membrane is

$$C = \frac{2\pi\kappa\epsilon_0 D}{b/a} \quad \text{(plane)}.$$

(e) How large is this correction for a myelinated axon in which $b/a = 0.4$?

Problem 69. Suppose that the outer radius of a myelinated axon, $d = a + b$, is fixed. Determine the value of a that maximizes the length constant of the axon (Eq. 6.75b). Ignore the Nodes of Ranvier. Your result should be expressed as $a = \gamma d$, where γ is a dimensionless constant.

Problem 70. Use the empirical relationships between axon radius and conduction speed in Table 6.2 to determine the radius and speed at which the speed along a myelinated and unmyelinated fiber is equal. For radii less than this radius, is propagation faster in myelinated or unmyelinated fibers? For speeds greater than this speed, in what type of fibers is propagation fastest?

Section 6.18

Problem 71. Modify the computer program of Fig. 6.38 to have a constant value of `jStim` and run it.

References

Anumonwo JB, Jalife J (1995) Cellular and subcellular mechanisms of pacemaker activity initiation and synchronization in the heart. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd edn. Saunders, Philadelphia

Covino BG (1972). Local anesthesia. N Engl J Med 286:975–983

Davis L Jr, Lorente de Nó R (1947) Contribution to the mathematical theory of the electrotonus. Stud Rockefeller Inst Med Res Repr 131(Part 1):442–496

DeBlois RW, Bean CP (1970) Counting and sizing of submicron particles by the resistive pulse technique. Rev Sci Inst 41(7):909–916

Demir SS, Clark JW, Murphy CR, Giles WR (1994) A mathematical model of a rabbit sinoatrial node cell. Am J Physiol 266:C832–C852

DiFrancesco D, Mangoni M, Maccaferri G (1995) The pacemaker current in cardiac cells. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd edn. Saunders, Philadelphia

Geddes LA (2000) Historical perspectives 3: recording of action potentials. In Bronzino JD (ed) The biomedical engineering handbook, 2nd edn, vol I. CRC, Boca Raton, pp HP3-1–HP3-11

Hall JE (2011) Guyton and Hall textbook of medical physiology, 12th edn. Saunders/Elsevier, Philadelphia

Hille B (2001) Ion channels of excitable membranes, 3rd edn. Sinauer, Sunderland

Hodgkin AL (1964) The conduction of the nervous impulse. Thomas, Springfield

Hodgkin AL, Huxley AF (1939) Action potentials recorded from inside a nerve fibre. Nature 144:710–711

Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117:500–544

Hodgkin AL, WAH Rushton (1946) The electrical constants of a crustacean nerve fiber. Proc R Soc B 133:444–479

Horowitz P, Hill W (2015) The art of electronics, 3rd edn. Cambridge University Press, London

Jeffreys H, Jeffreys BS (1956) Methods of mathematical physics. Cambridge University Press, London, p 602

Kane BJ, Storment CW, Crowder SW, Tanelian DL, Kovacs GTA (1995) Force-sensing microprobe for precise stimulation of mechanoreceptive tissues. IEEE Trans Biomed Eng 42(8):745–750

Katz B (1966) Nerve, muscle and synapse. McGraw-Hill, New York

Kubitschek HE (1969) Counting and sizing micro-organisms with the coulter counter. In: Ribbons RW, Norris JR (eds) Methods in microbiology, vol 1. Academic, London, pp 593–610. doi:10.1016/S0580-9517(08)70148-X

Laüger P (1991) Electrogenic ion pumps. Sinauer, Sunderland

Luo CH, Rudy Y (1994) A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes. Circ Res 74(6):1071–1096

Martin AR (1966) Quantal nature of synaptic transmission. Physiol Rev 46:51–66

Nicholls JG, Martin AR, Fuchs PA, Brown DA, Diamond ME, Weisblat D (2011). From neuron to brain, 5th edn. Sinauer, Sunderland

Noble D, DiFrancesco D, Denyer JC (1989) Ionic mechanisms in normal and abnormal cardiac pacemaker activity. In Jacklet JW, ed. Cellular and neuronal oscillators. Dekker, New York, pp 59–85

Noble D (1995) Ionic mechanisms in cardiac electrical activity. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd edn. Saunders, Philadelphia

Nolte J (2002) The human brain: an introduction to its functional anatomy, 5th edn. Mosby, St. Louis

Patton HD, Fuchs AF, Hille B, Scher AM, Steiner RF (1989) Textbook of physiology. Saunders, Philadelphia

Plonsey R (1969) Bioelectric phenomena. McGraw-Hill, New York

Plonsey R, Barr RC (2007) Bioelectricity: a quantitative approach, 3rd edn. Springer, New York

Rushton WAH (1951) A theory of the effects of fibre size in medullated nerve. J Physiol, 115:101–122

Schey HM (2004) Div, grad, curl and all that: an informal text on vector calculus, 4th edn. Norton, New York

Scott AC (1975) The electrophysics of a nerve fiber. Rev Mod Phys 47:487–533

Serway RA, Jewett JW (2013) Principles of physics, 5th edn. Brooks/Cole, Boston

Smythe WR, Silver S, Whinnery JR, Angelakos JD (1957). Formulas. In: Gray DE (ed) American Institute of Physics handbook, Chap. 5b. McGraw-Hill, New York

Wilders R, Jongsma HJ, van Ginneken ACG (1991) Pacemaker activity of the rabbit sinoatrial node: a comparison of models. Biophys J 60:1202–1216

Wiley JD, Webster JG (1982) Analysis and control of the current distribution under circular dispersive electrodes. IEEE Trans Biomed Eng 29:381–385

Worthington CR (1971) X-ray analysis of nerve myelin. In: Adelman JW Jr (ed) Biophysics and physiology of excitable membranes. Van Nostrand Reinhold, New York, pp 1–46

The Exterior Potential and the Electrocardiogram 7

In Chap. 6 we assumed that the potential outside a nerve cell is zero. This is only approximately true. There is a small potential that can be measured and has clinical relevance. Before a muscle cell contracts, a wave of depolarization sweeps along the cell. This wave is quite similar to the wave along the axon. Measurement of these exterior signals gives us the electrocardiogram, the electromyogram, and the electroencephalogram.

In Sect. 7.1 we calculate the potential outside a long cylindrical axon bathed in a uniform conducting medium. Section 7.2 shows that the exterior potential is small compared to the potential inside the cell if there is enough extracellular fluid so that the outside resistance is low. Section 7.3 uses a model in which the action potential is approximated by a triangular pulse to calculate the potential far from the cell. Section 7.4 generalizes this calculation to the case of a pulse of arbitrary shape.

An unusual feature of heart muscle is that the myocardial cells remain depolarized for 100 ms or so, as described in Sects. 7.5 and 7.6. This means that the potential difference outside the cell is much larger than for other cells, giving rise to the electrocardiogram described in Sects. 7.7 and 7.8. These two sections discuss electrocardiography and some factors that contribute to the signal. They make no attempt to consider advanced techniques such as orthogonal leads that are used to reconstruct the electrical activity of the heart from potential difference measurements on the surface of the body. Rather, they are much closer to the way clinicians think about the electrocardiogram, and they can provide a basis from which to learn more complicated techniques.

Section 7.9 talks about improved models that take into account the interaction between the inside and outside of cells and the anisotropies that exist in tissue resistance. Section 7.10 discusses the problem of stimulation: for measurement of evoked responses, for pacing, and for defibrillation.

Section 7.11 discusses the electroencephalogram.

7.1 The Potential Outside a Long Cylindrical Axon

When studying the action potential in Chap. 6 we assumed that the potential outside the axon is zero. Now we calculate the exterior potential distribution if the axon is in an infinite uniform conducting medium.[1] We will discover that for the case studied here the exterior potential changes are less than 0.1 % of those inside. If the exterior medium is not infinite, the exterior potential changes are larger, as is discussed in Sect. 7.9. This model also applies to a muscle cell that is depolarizing before contraction. We will adapt these results to a group of heart (myocardial) cells that depolarize together, leading to a wave of depolarization propagating through the tissue.

Consider a single axon stretched along the x axis. Divide space into three regions as shown on the left in Fig. 7.1: the interior of the axon (the axoplasm), the axon membrane, and the surrounding medium. Imagine that the current inside the axon is constant to the left of a certain point and zero to the right of that point, as shown on the right in Fig. 7.1b Since the axoplasm obeys Ohm's Law, the interior potential decreases linearly with x as shown in Fig. 7.1a. Where the current is zero, the interior potential does not change. At the point where the interior current falls to zero, conservation of charge requires that the current passes through the membrane and flows in the exterior conducting medium, as stated in Eq. 6.47b. Figure 7.1c shows the axon with current flowing in the left part of the axon and then flowing into the surrounding medium.

Now consider how the current flows in the surrounding three-dimensional medium. Suppose that the surrounding or "outside" medium is infinite, homogeneous and isotropic and has conductivity σ_o. Suppose also that the axon stretched

[1] Other textbooks examine this problem in greater detail (Gulrajani 1998; Malmivuo and Plonsey 1995).

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_7, © Springer International Publishing Switzerland 2015

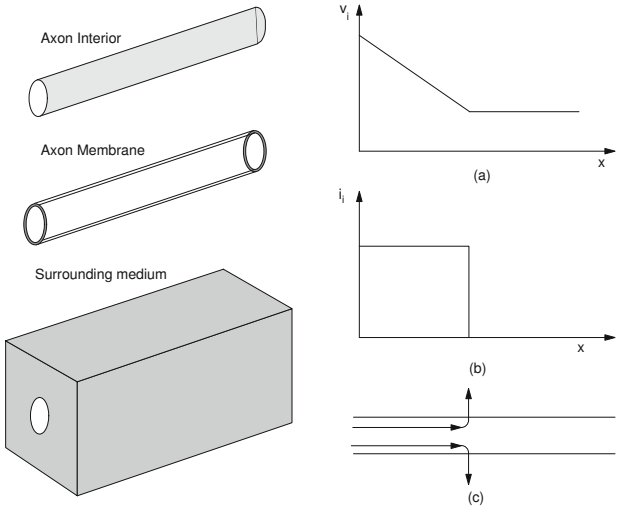


Fig. 7.1 An axon is stretched along the x axis. **a** A plot of a portion of the interior action potential at one instant of time. **b** A plot of the interior current, proportional to the slope of the interior potential because of Ohm's law. **c** Schematic representation of the axon, showing current flowing along the axon and into the exterior conducting medium at the point where the interior current falls to zero

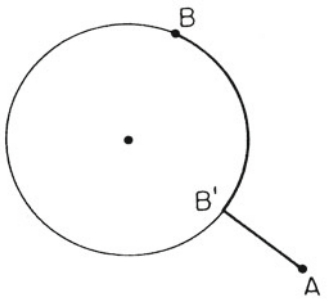


Fig. 7.2 A point current source is at the *center* of a sphere. The path of integration to calculate the potential difference between points A and B goes first from A to B' and then from B' to B

along the x axis is very thin and does not appreciably change the homogeneous and isotropic nature of the extracellular medium, except very close to the x axis. If a current i_o enters the surrounding medium at the origin, the current density $\mathbf{j}$ is directed radially outward and has spherical symmetry. The current density at distance r has magnitude $j = i_o/4\pi r^2$. The magnitude of the electric field is $E = j/\sigma_o = i_o/4\pi\sigma_o r^2$. This has the same form as the electric field from a point charge, for which the electric field is $E = q/4\pi\epsilon_0 r^2$. We speak of i_o as a *point current source*.

We can use the expression for the electric field to calculate the exterior potential. The point current source is shown as the dot in the center of the sphere in Fig. 7.2. To calculate the potential difference between points A and B, it is easier to integrate Eq. 6.16 along a path from A to B' parallel to the direction of $\mathbf{E}$, and then along $B'B$ where the displacement is

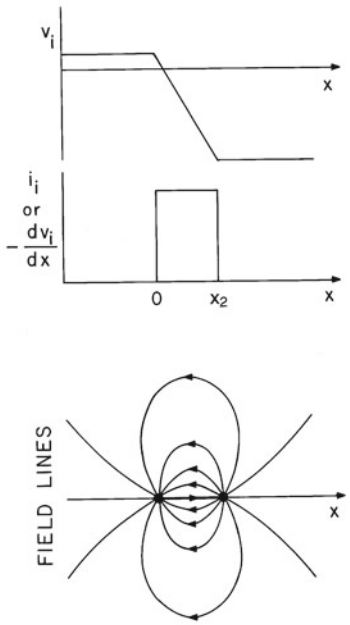


Fig. 7.3 The potential of Fig. 7.1 is extended to the left in a region of constant (depolarized) potential. The interior current is plotted below the potential. The electric field or current-density lines are plotted at the bottom. The current to the right on the axis is current within the axon; the other lines represent current in the exterior conducting medium

always perpendicular to $\mathbf{E}$. The potential change along $B'B$ is zero. Therefore,

$$v(B) - v(A) = -\int_{r_A}^{r_B} E_r\, dr = -\int_{r_A}^{r_B} \frac{i_o}{4\pi\sigma_o r^2} dr$$
$$= \frac{i_o}{4\pi\sigma_o}\left(\frac{1}{r_B} - \frac{1}{r_A}\right).$$

Only a difference of potential between two points has meaning. However, it is customary to define the potential to be 0 at $r_A = \infty$ and speak of the potential as a function of position. Then the potential at distance r from a point current source i_o is

$$v(r) = \frac{i_o}{4\pi\sigma_o r}. \tag{7.1}$$

The analogous expression for the potential due to a point charge q is $v(r) = q/4\pi\epsilon_0 r$.

We do not yet have a useful model, because the potential cannot rise forever as we go along the axon to the left. Let us assume that the potential levels off at some point on the left, as shown in Fig. 7.3. (This will turn out to be a very good model for the electrocardiogram, because the repolarization of myocardial cells does not take place for about 100 ms, so the cells are completely depolarized before repolarization begins.) Define the location of the origin so that the depolarization takes place between $x = 0$ and $x = x_2$. The potential

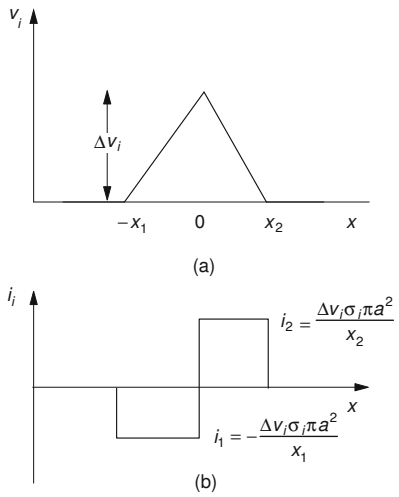


Fig. 7.4 The action potential is approximated by a triangular wave form. In this piecewise-linear approximation, the depolarization and repolarization are both linear. **a** The interior potential. **b** The interior current

change is shown at the top, and the current along the inside of the axon in the middle. The current exists only where there is a voltage gradient between $x = 0$ and $x = x_2$. Its magnitude is $i_i = \Delta v_i / R = \Delta v_i \sigma_i \pi a^2 / x_2$. This current flows out into the surrounding medium at $x = x_2$ and back into the axon at $x = 0$. Such a combination of source and sink of equal magnitude is called a *current dipole*. (A pair of equal and opposite electric charges is called an electric dipole.) The lowest part of the figure shows lines of **j** or **E**. The current is to the right inside the axon (along the axis) and returns outside the axon. The potential at any exterior point is due to two terms: one from the source i_i at $x = x_2$ and the other from the sink $-i_i$ at $x = 0$. If r_2 is the distance from the observation point to x_2 and r_0 is the distance to the origin, then

$$v = \frac{\Delta v_i \sigma_i \, \pi a^2}{4\pi \, \sigma_o x_2} \left(\frac{1}{r_2} - \frac{1}{r_0} \right) = \frac{\Delta v_i \sigma_i a^2}{4\sigma_o x_2} \left(\frac{1}{r_2} - \frac{1}{r_0} \right). \tag{7.2}$$

To estimate the exterior potential from a nerve impulse, we can approximate the action potential by a triangular potential as shown in Fig. 7.4a. The potential is zero far to the left. It rises by an amount Δv_i between $x = -x_1$ and $x = 0$. It falls linearly to zero at $x = x_2$. The current is plotted in Fig. 7.4b. In the region just to the left of the origin it is

$$i_1 = -\frac{\Delta v_i \sigma_i \, \pi a^2}{x_1}. \tag{7.3a}$$

(It is negative because it flows to the left.) To the right of the origin it is

$$i_2 = \frac{\Delta v_i \sigma_i \, \pi a^2}{x_2}. \tag{7.3b}$$

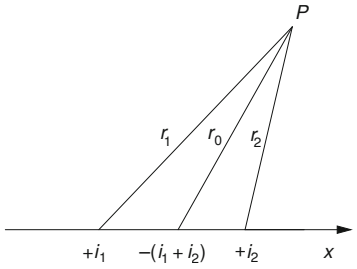


Fig. 7.5 The axon of Fig. 7.4 is stretched along the x axis. There are current sources at $x = -x_1$ and $x = x_2$, and a current sink at the origin. The distances from each source or sink to the observation point P are shown

Figure 7.5 shows the surrounding medium. There is a source of current i_1 at $x = -x_1$, a source i_2 at $x = x_2$, and a sink $-(i_1 + i_2)$ at the origin. The potential at observation point P is calculated by repeated application of Eq. 7.1:

$$v = \frac{1}{4\pi\sigma_o} \left(\frac{i_1}{r_1} - \frac{i_1 + i_2}{r_0} + \frac{i_2}{r_2} \right). \tag{7.3c}$$

Equations 7.3a–7.3c can be combined to give

$$v = \frac{\Delta v_i \sigma_i a^2}{4\sigma_o} \left(\frac{1/x_1}{r_1} - \frac{1/x_1 + 1/x_2}{r_0} + \frac{1/x_2}{r_2} \right). \tag{7.4}$$

Equations 7.3 and 7.4 are valid at any distance from the axon, as long as we can make the piecewise approximation of the action potential shown in Fig. 7.4.

7.2 The Exterior Potential is Small

Let us use Eq. 7.2 for the rising edge of the action potential to estimate the potential outside the axon when it is in an infinite conducting medium. We evaluate Eq. 7.2 close to the surface of the axon where the potential will be largest, say at $x = 0$. In that case r_2 is approximately x_2. However, r_0 is not zero. It can never become smaller than $r_0 = a$, the radius of the axon. (The potential would diverge if the model were extended to $r = 0$.) We will use an approximate value, $r_0 = a$, and call the height of the action potential Δv_i. Then

$$v(0) = \frac{\Delta v_i \sigma_i a^2}{4\sigma_o x_2} \left(\frac{1}{x_2} - \frac{1}{a} \right). \tag{7.5}$$

Since $1/x_2 \ll 1/a$, this becomes

$$v(0) \approx -\frac{\Delta v_i \sigma_i a}{4\sigma_o x_2}. \tag{7.6}$$

Close to $x = x_2$ the potential is

$$v(x_2) = \frac{\Delta v_i \sigma_i a^2}{4\sigma_o x_2} \left(\frac{1}{a} - \frac{1}{x_2} \right) \approx \frac{\Delta v_i \sigma_i a}{4\sigma_o x_2}. \tag{7.7}$$

The potential difference between these two exterior points is

$$\Delta v_o = v(x_2) - v(0) = \frac{\sigma_i}{\sigma_o}\frac{a}{2x_2}\Delta v_i. \tag{7.8}$$

If the conductivities were the same inside and outside, the ratio would be $\Delta v_o/\Delta v_i = a/2x_2$.

The ratio of exterior to interior potential change is proportional to the ratio of the axon radius to the distance along the axon over which the potential changes. From Fig. 6.42 we see that the rising part of the squid action potential has a length $x_2 \approx 1$ cm. If $a = 0.5$ mm (a quite large axon), then the ratio is $1/40$. For a smaller axon, the ratio is even less.

The same result can be obtained by another argument. The resistance between two points is the ratio of the potential difference between the points to the current flowing between them. Inside the axon, **j** and **E** are large because the current is confined to a small region of area πa^2. The resistance inside is $R_i = x_2/\pi a^2\sigma_i$. The same current flows outside, but it is spread out so that **j** and **E** are much less. The resistance between two electrodes in a conducting medium is related to their capacitance (Sect. 6.19). Equations 6.86 and 6.99 can be used to show that two spherical electrodes of radius a spaced distance x_2 apart ($x_2 \gg a$) have a resistance $R_o = 1/2\pi\sigma_o a$. The voltage ratio is

$$\frac{\Delta v_o}{\Delta v_i} = \frac{R_o}{R_i} = \frac{1}{2\pi\sigma_o a}\frac{\pi a^2\sigma_i}{x_2} = \frac{\sigma_i}{\sigma_o}\frac{a}{2x_2},$$

the same result as Eq. 7.8.

7.3 The Potential Far from the Axon

In most cases measurements of the potential are made far from the axon—far compared to the distance the action potential spreads out along the axon. If point P is moved far away, Fig. 7.5 looks like Fig. 7.6. The lines r_1, r_0, and r_2 are nearly parallel. If point P is located at distance r_0 from the origin at angle θ with the x axis, then

$$r_2 \approx r_0 - x_2\cos\theta, \quad r_1 \approx r_0 + x_1\cos\theta. \tag{7.9}$$

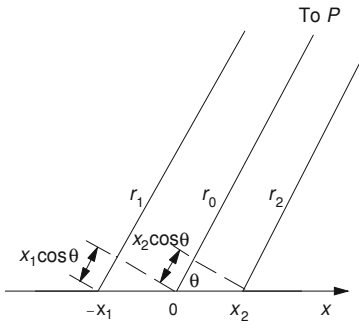


Fig. 7.6 The observation point P is far away compared to distances x_1 or x_2. The *lines* to P are nearly parallel

Consider the potential in Eq. 7.2 due to the leading edge of the action potential. (We will argue later that this is a useful model for the electrocardiogram.) Substituting Eqs. 7.9 in Eq. 7.2 gives

$$v = \frac{\Delta v_i\sigma_i a^2}{4\sigma_o x_2}\left(\frac{1}{r_0[1-(x_2/r_0)\cos\theta]} - \frac{1}{r_0}\right).$$

You can verify by a Taylor's-series expansion or long division that

$$\frac{1}{1-x} = 1 + x + \cdots, \tag{7.10}$$

so that

$$v = \frac{\Delta v_i\sigma_i a^2}{4\sigma_o r_0^2}\cos\theta. \tag{7.11}$$

This is a very important result that will form the basis for our model of the electrocardiogram:

1. The exterior potential v depends on Δv_i but not on x_2, the length of the depolarization region. This is because increasing x_2 decreases the strength of the current at the same time that it increases v because the source and sink are further apart.
2. The potential falls off as $1/r^2$ instead of $1/r$ as it would from a point source.
3. The potential varies with angle, being positive to the right of the transition region and negative to the left.

It is convenient to define a vector **p** that points along the axon in the direction of the advancing depolarization wave front (the region along the axon where the potential rises). It is called the *activity vector* or *current-dipole moment* for reasons discussed shortly. Its magnitude is

$$p = \pi a^2\sigma_i\,\Delta v_i. \tag{7.12}$$

The exterior potential is then (dropping the subscript on **r**)

$$v = \frac{\mathbf{p}\cdot\mathbf{r}}{4\pi\sigma_o r^3}. \tag{7.13}$$

Vector **p** has units of A m. Its magnitude (apart from the conductivity) is the product of the cross-sectional area of the axon and the difference in potential along the axon between the resting and completely depolarized regions. It is called the current-dipole moment because it is the product of the current and the separation of the source and sink. (The electric-dipole moment is the product of the magnitude of the charges and their separation, with units C m.)

Equation 7.11 can also be written in the form

$$v(r) = \frac{\pi a^2\cos\theta}{r^2}\frac{1}{4\pi}\frac{\sigma_i}{\sigma_o}\Delta v_i. \tag{7.14}$$

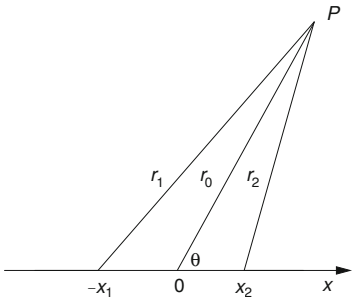


Fig. 7.7 When the observation point is not so far away, or when a complete nerve impulse is being considered, the law of cosines must be used to relate r_1 and r_2 to r_0

The quantity $\pi a^2 \cos\theta / r^2$ is $\Delta\Omega$, the *solid angle*[2] subtended at the observation point by a cross section of the axon where the potential changes. The quantity 4π is the maximum possible solid angle. In terms of the solid angle

$$v = \frac{\Delta\Omega}{4\pi}\frac{\sigma_i}{\sigma_o}\Delta v_i. \tag{7.15}$$

Now consider an entire pulse, one where the potential rises and then returns to the resting value. If the approximation of Eq. 7.10 is applied to Eq. 7.4, the result vanishes. It is necessary to make a more accurate approximation, one that takes into account the fact that the vectors $\mathbf{r}_1$, $\mathbf{r}_0$, and $\mathbf{r}_2$ are not exactly parallel. Figure 7.7 shows the geometry. We use the law of cosines to write [remember that $\cos(\pi - \theta) = -\cos\theta$]

$$r_1 = r_0\left[1 + (2x_1/r_0)\cos\theta + x_1^2/r_0^2\right]^{1/2},$$

$$r_2 = r_0\left[1 - (2x_2/r_0)\cos\theta + x_2^2/r_0^2\right]^{1/2}.$$

When these are inserted in Eq. 7.4 and a Taylor's-series expansion is done to second order in both x_1/r_0 and x_2/r_0, the result is

$$v = \frac{2\pi a^2}{4\pi r^3}\frac{\sigma_i}{\sigma_o}\frac{\Delta v_i(x_1 + x_2)}{2}\frac{3\cos^2\theta - 1}{2}. \tag{7.16}$$

The constants have been arranged to show that the term $\Delta v_i(x_1 + x_2)/2$ is the area under the impulse when v is plotted as a function of x. The angular factor as written with its factor of 2 in the denominator is tabulated in many places as the *Legendre polynomial* $P_2(\cos\theta)$.[3] The exterior potential

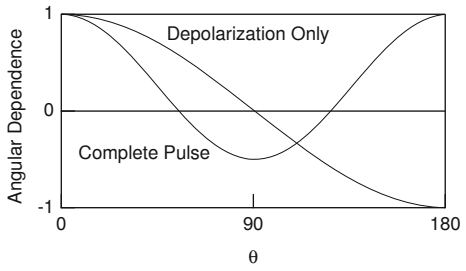


Fig. 7.8 Plot of the angular dependence of the potential from the entire impulse, Eq. 7.16

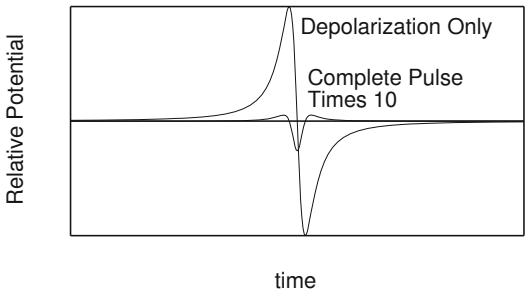


Fig. 7.9 The potential far from the axon as a function of time as an impulse travels from left to right along the axis. The potential from the complete pulse has been multiplied by a factor of 10 in order to show it

now falls off more rapidly with distance, as $1/r^3$. The angular dependence, shown in Fig. 7.8, is symmetric about $\pi/2$. This shows the angular dependence as one moves around the impulse at a constant distance from it.

This is a very different situation and a very different curve from the potential measured at a fixed point outside the axon as an impulse travels past. In the latter case r as well as θ is changing. This behavior is discussed in Problems 8 and 9. The results are shown in Fig. 7.9. The potential from the depolarization is biphasic; that from the complete pulse is triphasic, being positive, then negative, then positive again.

For a single axon in an ionic solution the exterior conductivity is usually higher than in the axon, so $\sigma_i/\sigma_o = 0.2$. The conductivity of tissue is considerably less than the conductivity of an ionic solution, and the ratio becomes greater than one. For the electrocardiogram it will be more appropriate to use $\sigma_o = 0.33\,\mathrm{S\ m^{-1}}$ (muscle) or $0.08\,\mathrm{S\ m^{-1}}$ (lung), in which case σ_i/σ_o is 6 or 25. We will use an approximate value of 10.

7.4 The Exterior Potential for an Arbitrary Pulse

We have derived the results of the previous sections for an action potential that varies linearly during depolarization and repolarization, a piecewise-linear approximation. In general

[2] The solid angle is defined in Appendix A.

[3] You can learn more about Legendre polynomials in texts on differential equations or, for example, in Harris and Stocker (1998). See also Eq. 7.29.

the action potential does not have sharp changes in slope. We will now consider the general case and find that the results are very similar. For depolarization alone, we will again have a potential depending on the dipole moment. For a complete pulse the potential will depend on the area under the pulse curve.

Again, the axon is stretched along the x axis in an infinite, homogeneous conducting medium. Consider a small segment of axon between x and $x + dx$. If the current entering this segment at x is greater than the current leaving at $x + dx$, the difference must flow into the exterior medium. From Eq. 6.47b,

$$di_o = -di_i = -\frac{\partial i_i(x,t)}{\partial x}dx.$$

We can write Ohm's law for the axoplasm as

$$i_i(x,t) = -\pi a^2 \sigma_i \frac{\partial v_i}{\partial x}. \tag{7.17}$$

The current into the exterior medium from length dx of the axon is

$$di_o = \pi a^2 \sigma_i \frac{\partial^2 v_i}{\partial x^2}\, dx. \tag{7.18}$$

It is proportional to the derivative of the current along the axon with respect to x and therefore to the second derivative of the interior potential with respect to x. A small current source di_o generates a potential dv at some point in the exterior medium given by

$$dv = \frac{di_o}{4\pi\sigma_o r}. \tag{7.19}$$

If the radius of the axon stretched along the x axis is very small, the axon's influence can be replaced by a current distribution $di_o(x)$ along the x axis. The potential at any point $\mathbf{R}$ is obtained by integrating Eq. 7.19:

$$v(\mathbf{R}) = \int \frac{di_o}{4\pi\sigma_o r}. \tag{7.20}$$

Vector $\mathbf{R}$ specifies the point at which the potential is measured, and r is the distance from the measuring point to the point on the x axis where di_o is injected, as shown in Fig. 7.10. Combining this with Eq. 7.18 gives

$$v(\mathbf{R}) = \int \frac{\pi a^2 \sigma_i}{4\pi\sigma_o} \frac{\partial^2 v_i}{\partial x^2} \frac{1}{r}\, dx. \tag{7.21}$$

Although it is difficult to integrate Eq. 7.21 analytically, the integration can be done numerically. Figure 7.11 shows a computer program to carry out this integration for a crayfish lateral giant axon immersed in sea water. The axon radius is 60 μm. The conductivity ratio is $\sigma_i/\sigma_o = 0.2$. The action potential was measured by Watanabe and Grundfest (1961). Clark and Plonsey (1968) showed that it could be well represented by a sum of three Gaussians, with $v_i = 0$ taken to be the resting value. Since only $\partial^2 v_i/\partial x^2$ enters into Eq. 7.21, the reference level does not matter. The representation (with v in mV and x in mm) is

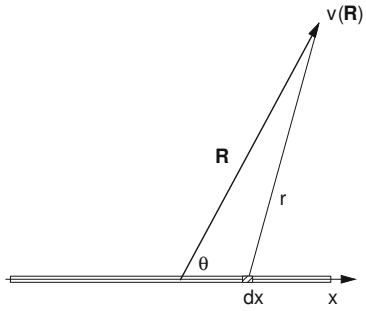


Fig. 7.10 The potential $v(\mathbf{R})$ is obtained by integrating the potential due to current di_o from each element dx of the cell

$$v_i(x) = 51e^{-[(x-5.4)/1.25]^2} + 72e^{-[(x-6.6)/1.876]^2} + 18e^{-[(x-8.6)/3.003]^2}. \tag{7.22}$$

This function corresponds to an impulse traveling to the left. It can be differentiated to obtain an analytic expression for $\partial^2 v_i/\partial x^2$. If the potential is being measured at exterior point (x_0, y_0), the value of r which is used in Eq. 7.21 is $r = \left[(x - x_0)^2 + y_0^2\right]^{1/2}$. The program allows four values of y_0 to be used. The smallest is taken to be a, the radius of the axon.

The results of calculating the exterior potential at $y_0 = a$ are shown in Fig. 7.12. The interior potential, shown in (a), has a peak value of 114 mV. The potential on the surface of the axon (b) ranges from $+0.04$ to -0.07 mV. In general the exterior potential is less than 0.1 % of the interior potential. (This would be different if the extracellular fluid were not infinite.) The original calculation by Clark and Plonsey used much different mathematical techniques (see Problem 30); however, the results are very similar. The results of their more accurate calculation are plotted in Fig. 7.13.

The approximation that the observer is far from the axon can also be applied to the general case. The physics is exactly the same as in the previous section for the triangular pulse, except that now the pulse has an arbitrary shape so current passes through the membrane at all points along the axon where the second derivative is nonzero. The calculation requires making the same type of approximations in order to evaluate the integral (Eq. 7.21). Referring to Fig. 7.10, we again use the law of cosines to write

$$r(x) = R\left[1 - 2(x/R)\cos\theta + x^2/R^2\right]^{1/2}.$$

We need to use this in Eq. 7.21. As in the previous section, we make a Taylor's-series expansion of $1/r$. To second order

```
//program ClarkAndPlonsey;
/*This program integrates our approximate
equation for the extracellular potential.
Original calc. by J. Clark and R. Plonsey,
The extracellular potential field of the
single active nerve fiber in a volume
conductor. Biophys. J. 8:842-864 (1968)*/
/*The nerve pulse is a series of Gaussians
used by Clark and Plonsey to fit data of
Watanabe and Grundfest, J. Gen. Physiol.
45:267 (1961)*/
/*This program uses Romberg integration
function qromb, from the Numerical Recipes
in C, 2nd ed.*/

#include<stdio.h>
#include<stdlib.h>
#include<math.h>
#include "nr.h" //N.R. in C header file

/*Global Variables*/
double x0,y0; //coordinates of obs point

float f(float x)
{
/*Calculates integrand d2v/dx2 div by r
   Uses common variables x0 , y0--
Observation point in meter*/
double   xx, r, d2v, temp;
d2v = 0;
xx = (x-0.0054)/1.25e-3;
temp = (2*51/(1.25e-3*1.25e-3))
   *exp(-(xx*xx));
d2v = d2v+temp*(2*xx*xx-1);
xx = (x-0.0066)/1.876e-3;
temp = (2*72/(1.876e-3*1.876e-3))
   *exp(-(xx*xx));
d2v = d2v+temp*(2*xx*xx-1);
xx = (x-0.0086)/3.003e-3;
temp = (2*18/(3.003e-3*3.003e-3))
   *exp(-(xx*xx));
d2v = d2v+temp*(2*xx*xx-1);
r = sqrt((x0-x)*(x0-x)+y0*y0);
return d2v/r;
}

void main()
{
const
   double SigRatio = 0.2,
   //Interior/exterior conductivity
   a = 6.0e-5;  //axon radius in m
float
   xstart, xfinish,//limits of int.;
double
   y[4],//Calculate at four distances
   potential, xx  //Exterior potential
  Transmemb;   //Transmembrane potential
int i;
FILE *ofp;        //Outputfile Pointer
if (!(ofp = fopen("Plonsey.out","w")))
//Open output file
{
   printf("cannnot open output file\n");
   exit(1);
}
xstart = 0.0;
xfinish = 0.02;
fprintf(ofp,"    x   \t    v   ");
printf("     x          v    ");
for (i=0; i<4; i++)
{
   y[i] = pow(2,i)*a;
   fprintf(ofp,"\t%9.3e",y[i]);
   printf(" %9.3e ",y[i]);
}
fprintf(ofp,"\n\n");
printf("\n\n");
for(x0 = 0.001; x0 < 0.012; x0 += 0.00025)
{
   xx = (x0-0.0054)/1.25e-3;
   Transmemb = 51*exp(-xx*xx);
   xx = (x0-0.0066)/1.876e-3;
   Transmemb = Transmemb+72*exp(-xx*xx);
   xx = (x0-0.0086)/3.003e-3;
   Transmemb = Transmemb+18*exp (-xx*xx);
   fprintf(ofp,"%9.3e\t%10.3e",x0,Transmem
b);
   printf("%9.3e %10.3e ",x0,Transmemb);
   for (i=0; i<4 ; i++)
   {
      y0 = y[i];
      Integral = qromb(f,xstart,xfinish);
//N.R. in C function Call
      potential=Integral*a*a*SigRatio/4.0;
      fprintf(ofp,"\t%10.3e",potential);
      printf("%10.3e ",potential);
   }
   fprintf(ofp,"\n");
   printf("\n");
   }
fclose(ofp);
}
```

Fig. 7.11 The computer program used to calculate the exterior potential by integrating Eq. 7.21 for the problem first solved by Clark and Plonsey (1968). The program uses Romberg integration procedure `qromb` from Press et al. (1992)

the result is

$$\frac{1}{r} \approx \frac{1}{R}\left(1 + \frac{x}{R}\cos\theta + \frac{1}{2}\frac{x^2}{R^2}(3\cos^2\theta - 1)\right). \tag{7.23}$$

The expression for $v(\mathbf{R})$ becomes

$$v(R) = \frac{\pi a^2 \sigma_i}{4\pi\sigma_o}\left[\frac{1}{R}\int_{x_1}^{x_2}\frac{\partial^2 v_i}{\partial x^2}\,dx + \frac{\cos\theta}{R^2}\int_{x_1}^{x_2}\frac{\partial^2 v_i}{\partial x^2}x\,dx + \frac{3\cos^2\theta - 1}{2R^3}\int_{x_1}^{x_2}\frac{\partial^2 v_i}{\partial x^2}x^2\,dx\right]. \tag{7.24}$$

There are three integrals that we must evaluate. Take limits of integration x_1 and x_2 to be points where $\partial v_i/\partial x = 0$. The first integral is $\partial v_i/\partial x$ which vanishes at the end points. The second integral can be integrated by parts. Since $\partial v_i/\partial x = 0$ at the end points the second integral is $v_i(x_1) - v_i(x_2)$. The third integral is integrated by parts twice and is

$$\int_{x_1}^{x_2}\frac{\partial^2 v_i}{\partial x^2}x^2\,dx = \left[x^2\frac{\partial v_i}{\partial x}\right]_{x_1}^{x_2} - 2\left[x v(x)\right]_{x_1}^{x_2} + 2\int_{x_1}^{x_2} v_i(x)\,dx. \tag{7.25}$$

The first term of this vanishes because of the way the end points were chosen.

We now apply these results to Eq. 7.24 in two cases. The first is the case of depolarization only, which is useful in considering the electrocardiogram. Set up the coordinate system so the origin is someplace in the impulse where $\partial v_i/\partial x = 0$. The total change in v_i is Δv_i. Then $x_1 = 0$, $v_i(x_1) = \Delta v_i$, $v_i(x_2) = 0$. The first nonvanishing term of Eq. 7.24 requires only the second integral:

$$v(\mathbf{R}) = \frac{\pi a^2 \sigma_i}{4\pi\sigma_o}\frac{\cos\theta}{R^2}\Delta v_i. \tag{7.26}$$

We obtained this result in a special case as Eq. 7.11.

In the second case we consider the complete pulse, and we take x_1 to the left of the pulse and x_2 to the right. The first integral in Eq. 7.24 still vanishes. Now the second integral also vanishes because $v_i(x_1) = v_i(x_2) = v_{\text{rest}}$ and $\Delta v_i = 0$. It is necessary to use the third integral, Eq. 7.25. The first term in Eq. 7.25 vanishes. The second and third terms must be considered together. Rewrite the potential in terms of departures from the resting potential: $v_i(x) = v_{\text{rest}} + v_{\text{depol}}(x)$.

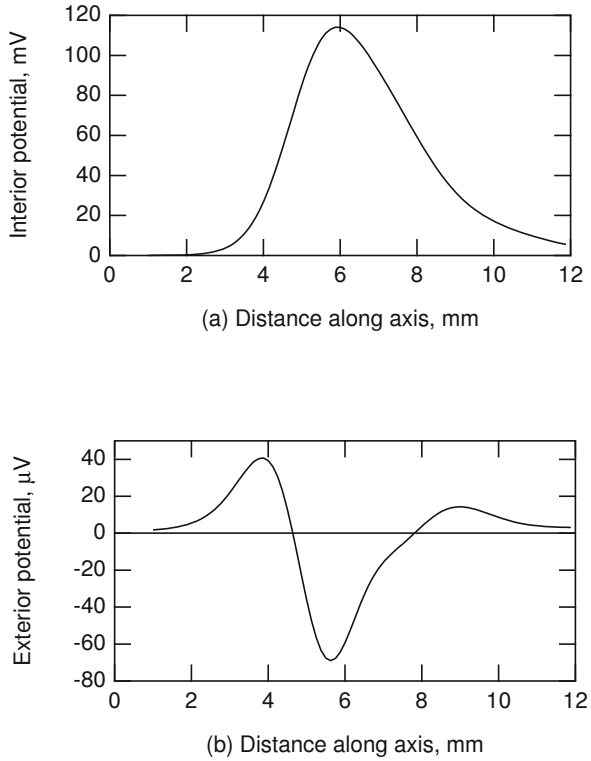


Fig. 7.12 **a** The transmembrane potential used for the calculation in the program of Fig. 7.11. The impulse is traveling to the *left*. **b** The exterior potential along the axon calculated by the program for $y_0 = a$

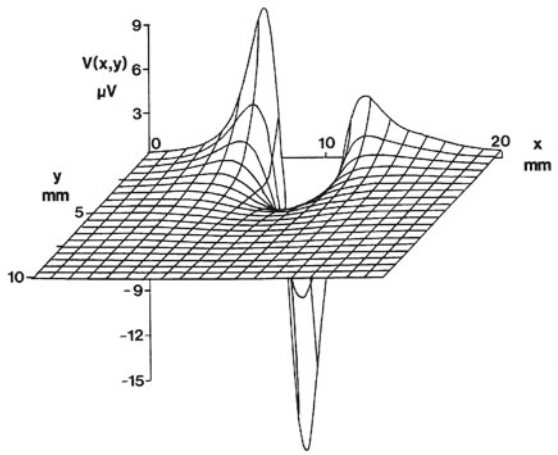


Fig. 7.13 The exterior potential for the same problem calculated using the more accurate method of Clark and Plonsey (1968). The smallest distance from the axon is $y = 0.5$ mm

The second term in Eq. 7.25 is $-2v_{\text{rest}}(x_2 - x_1)$. The third term of Eq. 7.25 is

$$2\int_{x_1}^{x_2} v_{\text{depol}}(x)\,dx + 2(x_2 - x_1)v_{\text{rest}}.$$

Adding these gives

$$v(\mathbf{R}) = \frac{2\pi a^2 \sigma_i}{4\pi\sigma_o}\frac{1}{R^3}\frac{3\cos^2\theta - 1}{2}\int_{x_1}^{x_2}[v_i(x) - v_{\text{rest}}]\,dx. \tag{7.27}$$

Again, we saw a special case of this as Eq. 7.16.

Note the progression in these results. When we are looking at one corner of a depolarization pulse, we have a current source or sink, and the potential is proportional to $1/R$ (Eq. 7.1). We do not find this situation in physiology because the potential would have to keep rising forever. When we consider the entire depolarization portion of the wave form, the potential is proportional to $1/R^2$, as in Eqs. 7.11 or 7.26 (We will find that this is a good model for the electrocardiogram because the repolarization does not commence until the entire heart is depolarized[4]). When the entire pulse is considered, the potential is proportional to $1/R^3$ as in Eq. 7.16 or 7.27. This is a good model for nerve conduction. The potential is considerably less in this case because of the $1/R^3$ dependence.

This is an example of a technique called a *multipole expansion*. Generally, defining $\xi = x/R$, one can make the expansion

$$\frac{1}{(1 - 2\xi\cos\theta + \xi^2)^{1/2}} = P_0 + \xi P_1 + \xi^2 P_2 + \xi^3 P_3 + \cdots, \tag{7.28}$$

where the P_n are functions of $\cos\theta$ and are called Legendre polynomials. The first few Legendre polynomials are

$$\begin{aligned} P_0 &= 1, \\ P_1 &= \cos\theta, \\ P_2 &= \tfrac{1}{2}(3\cos^2\theta - 1), \\ P_3 &= \tfrac{1}{2}(5\cos^3\theta - 3\cos\theta). \end{aligned} \tag{7.29}$$

All of these calculations are based on a model in which the current flows parallel to the axis of the axon, passes through the membrane, and then returns in the extracellular conducting medium. This model is called the *line approximation*. It is, of course, impossible for current inside the axon to pass out through the membrane if it always flows parallel to the axis of the axon. It is possible to do an exact calculation in which **j** has a radial component as well as one parallel to the axis of the axon. (See Sect. 7.9 for a description of how this is done.) Trayanova et al. (1990) have compared the exact solution with two approximations, one of which is the line approximation. The line approximation is quite good if the radius of the axon is much smaller than the distance along the axon over which the depolarization takes place.

[4] This is not strictly true. Atrial repolarization begins before the ventricular depolarization is complete.

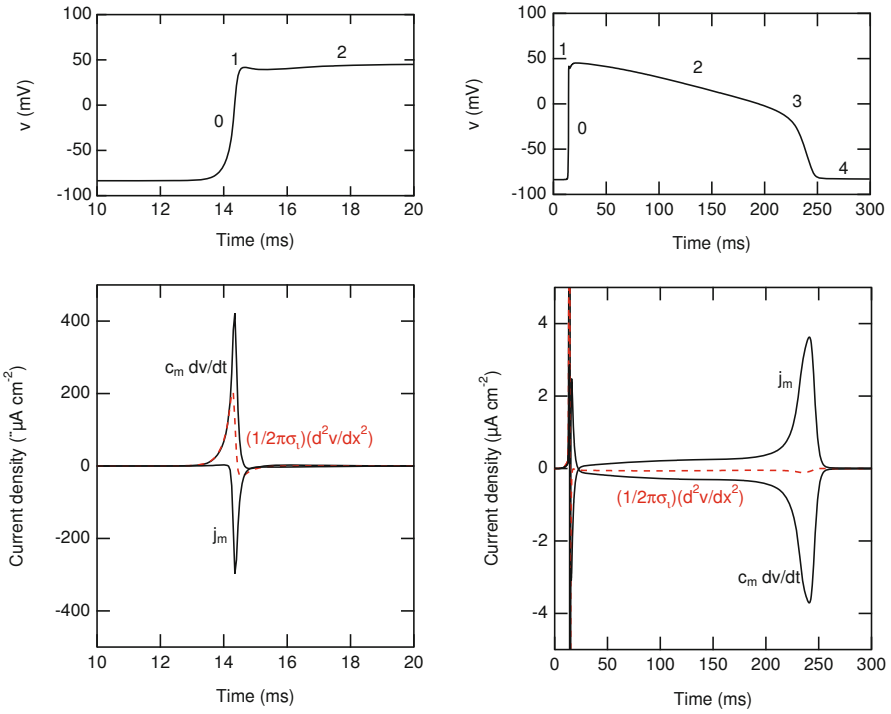


Fig. 7.14 Depolarization and repolarization of cardiac cells based on a model by Luo and Rudy (1994). Panels on the left show the depolarization. Panels on the right show the entire action potential. Note the factor of 100 change in scale of the current density on the lower two panels. Electrotonus (the dashed line) is important during depolarization but is almost nonexistent during repolarization. Compare this with Fig. 6.42. The calculations were performed by Sunil Kandel

7.5 Electrical Properties of the Heart

We saw in Chap. 1 that the heart is divided into a right side and a left side (see Fig. 1.34). Each side consists of an *atrium* and a *ventricle*. The ventricles are primarily responsible for pumping the blood. They are relatively large chambers with thick walls. The smaller atria contract first. The atria fill the ventricles with blood before the ventricles contract. The two chambers on each side are connected by one-way valves: the *tricuspid valve* on the right and the *mitral valve* on the left, so that blood cannot regurgitate back into the atrium when a ventricle contracts.

The right and left atria are electrically connected: if the right atrium contracts, so does the left atrium. The right and left ventricles are similarly connected. The electrical connection between the atria and the ventricles occurs only at the AV node, as discussed below.

There are many similarities between myocardial cells and nerve cells: a membrane separates extracellular and intracellular fluids; the concentrations of the principal ions are about the same; except for a small amount of charge on the membrane, the extracellular and intracellular fluids are electrically neutral; and selective ion channels are responsible for the initiation and propagation of the action potentials. There are also major differences: myocardial cells in mammals are about 100 μm long and 10 μm in diameter. The interiors of neighboring cells are connected through *gap junctions*, so current and ions flow directly from one cell to another (Delmar and Sorgen 2009). This continuum of cells is called a *syncytium*. There are also important differences in the details of the ion currents. We continue for now to use the simple model of long, one-dimensional cells. Refinements to this model are discussed in Sect. 7.9.

In the resting state, the potential inside an atrial cell is about −70 mV, while that in a ventricular cell is about −85 mV. When a cell depolarizes, the action potential lasts for 100–300 ms, depending on the species. A "typical" action potential is shown in Fig. 7.14. There are variations in pulse shape between species and also in different parts of the heart. The initial rapid depolarization is caused by an inward sodium current (phase 0 on the curve) and takes about 1 ms. This is sometimes followed by a rapid fall (phase 1) not prominent in Fig. 7.14, caused by a transient outward potassium current. This current is small in *endocardium* (near the inside of the heart) but is prominent enough in the *epicardium* (outer layers of the heart) so that there can be a "spike and dome" shape to the potential (George 2009). This is followed by a Ca^{2+} influx that maintains the plateau (phase 2) of the action potential. The "slow" potassium channels

finally open (Oudit and Backx 2004), and potassium efflux causes repolarization (phase 3). During phase 4 the original ion concentrations are restored.

The heart can beat in isolation. If it is removed from an animal and bathed in nutrient solution, it continues to beat spontaneously. With each beat, a wave of depolarization sweeps over the heart, and it contracts. The wave is initiated by some specialized fibers located in the right atrium called the *sinoatrial node* (SA node). As was mentioned in Sect. 6.18, the SA node does not have the usual sodium channels, and the depolarization is due to calcium. The shape of the SA node potential is much more like Fig. 6.48 than Fig. 7.14. In humans the SA node fires about 60–100 times per minute; this rate is increased by the *sympathetic* nerves to the heart (which release *norepinephrine*) and decreased by the *parasympathetic* nerves (which release *acetylcholine*). Devices that produce such periodic firing are common in physics and engineering. They are called free-running relaxation oscillators.

Figure 7.15 shows how the depolarization progresses through the heart. Once the SA node has fired, the depolarization sweeps across both atria (a, b). When the atria are completely depolarized (c) there is no depolarization wavefront. The atria are separated from the ventricles by fibrous connective tissue that does not transmit the impulse. The only electrical connection between the atria and the ventricles is some conduction tissue called the *atrioventricular node* (AV node). After passing through the AV node, the depolarization spreads rapidly over the ventricles through the *conduction system*—a set of specialized muscle cells on the inner walls of the ventricles—(d, e), and finally through the myocardium of each ventricle to the outer wall (e, f, g). The conduction system consists of the *common bundle* (or *bundle of His*), the *left and right bundles,* and the fine network of *Purkinje fibers*. The AV node will spontaneously depolarize at a rate of about 50 beats per minute; it usually does not because it is triggered by the more rapid beating of the atria. In well-trained athletes, the resting pulse rate can be so low that the AV node fires spontaneously, giving rise to what are called nodal escape beats. These are physiologic and no cause for concern.

There is a difference between depolarization, which propagates as a wave, and repolarization, which is a local phenomenon. Sodium conductance increases as the transmembrane potential rises during depolarization. As the potential rises at some point on the advancing wave front, electrotonus increases the potential further along the cell, as can be seen in the left panels of Fig. 7.14, where the contribution from electrotonus is shown by the dashed line. This causes the sodium conductance to rise at that point, resulting in the propagation of the signal at speeds of about $0.2 - 0.5$ m s^{-1}. During repolarization electrotonus contributes almost nothing to the repolarization, as can be seen in the panels on the right (note the factor of 100 difference in the current density).

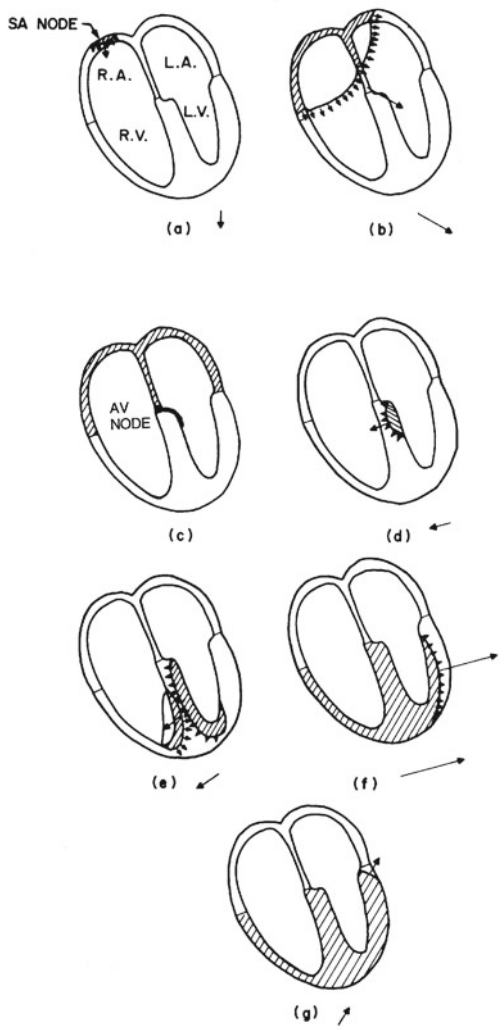


Fig. 7.15 The wave of depolarization sweeping over the heart. Atrial and ventricular muscle are not connected except through the AV node. **a** Depolarization beginning at the SA node. **b** Atria nearly depolarized. **c** The AV node is conducting. **d** Beginning of depolarization of the left ventricle. **e**, **f** Continuing ventricular depolarization. **g** Ventricular depolarization nearly complete. (Reprinted with permission from Hobbie 1973, Copyright © 1973, American Association of Physics Teachers)

Normally, depolarization progresses through the myocardium in an orderly fashion (Fig. 7.15). It is followed by repolarization, and after a brief *refractory period* the heart is ready to beat again. During the refractory period the cells do not respond to a stimulus. It is possible in abnormal situations for a wave of depolarization to travel in a closed path through the myocardium. This closed path, called a *reentrant circuit*, can surround an obstacle such as scar tissue, the aorta, or the pulmonary artery. It can also surround an area that simply has different conduction properties. If the time to travel around the reentrant circuit is greater than the refractory period, the

wave can continue to travel on the closed path. Reentrant excitation is thought to be responsible for several kinds of heart disease, including most life-threatening *ventricular tachycardias* (rapid heart rate). Another type of reentrant excitation is spiral waves that occur because of the nonlinear nature of the myocardium (Gray 2009; Such nonlinear behavior will be discussed in Chap. 10). It is also possible for a reentrant wave to leave behind a refractory state that blocks normal conduction.

7.6 The Current-Dipole Vector of the Heart as a Function of Time

Each myocardial cell depolarizes and repolarizes during the cardiac cycle. These cells are short—about 100 μm in length—but are interconnected. We apply our axon model by noting that a current i_i flows within each cell during depolarization and a return current, which could be ignored in our axon model, flows in the surrounding tissue. We assume that each cell as it depolarizes has a current dipole moment, and that these can be summed.

The total current-dipole vector at any instant is then the sum of the vectors for all the cells in the heart. This section considers how the total current-dipole vector changes with time as the myocardium depolarizes and then repolarizes. Initially, all the cells are completely polarized (resting) and there is no net dipole moment. The cells begin to depolarize near the SA node, and a wave of depolarization sweeps across the atria. For each myocardial cell, the dipole vector points in the direction that the wave of depolarization is traveling[5] and moves along the cell with the depolarization wave. These vectors for all the cells that are depolarizing constitute an advancing wave that moves across the heart.

The potential at the point of observation can be calculated by applying Eq. 7.13 for each cell. Vector **r** is the vector from the cell to the point of observation and is different for each cell. However, we will assume for now that the observation point is so far from the heart that all points in the myocardium are nearly equidistant from it. This is a terrible assumption; later we will be more realistic. It allows us to speak of the *instantaneous total current dipole moment*, which is the sum of the dipole moments of all depolarizing cells at that instant.

The locus of the tip of the total dipole moment during the cardiac cycle is shown in Fig. 7.16 for a typical case. The x axis points to the patient's feet, the y axis to the patient's left, and the z axis from back to front. The small loop labeled P occurs during atrial depolarization. The loop labeled QRS is the result of ventricular depolarization. Ventricular repolarization gives rise to the "T wave." Atrial repolarization is masked by ventricular depolarization. A plot of the x, y, and z components of **p** is shown in Fig. 7.17. These components are typical; there can be considerable variation in the directions of the loops in Fig. 7.16.

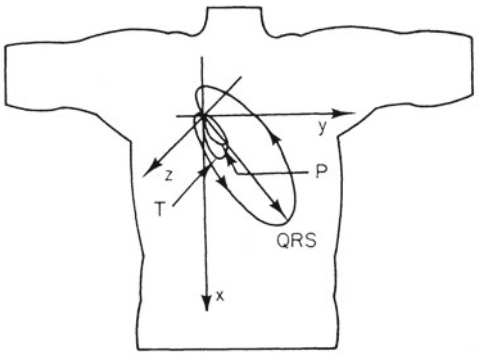


Fig. 7.16 The locus of the tip of the total current-dipole vector during the cardiac cycle. The z axis is perpendicular to the x and y axes and the subject's chest and comes out of the page

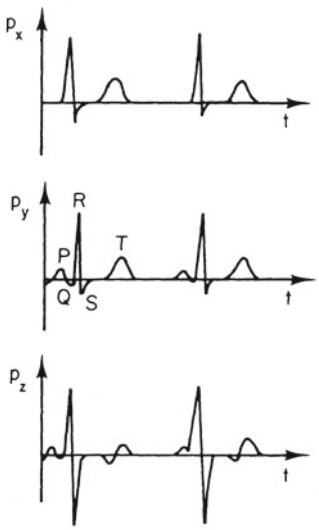


Fig. 7.17 The three components of the total current dipole vector **p** as a function of time

7.7 The Electrocardiographic Leads

We turn next to how the electrocardiographic measurements are made. We model the torso as an infinite homogeneous conductor and continue to assume that every myocardial cell

[5] If one takes into account the anisotropies in the conductivities of myocardial tissue discussed in Sect. 7.9, the depolarization does not travel in the direction that **p** points. We ignore this for now.

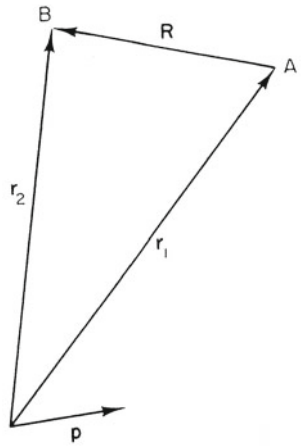


Fig. 7.18 Geometry for calculating the potential difference due to **p** between points *A* and *B*

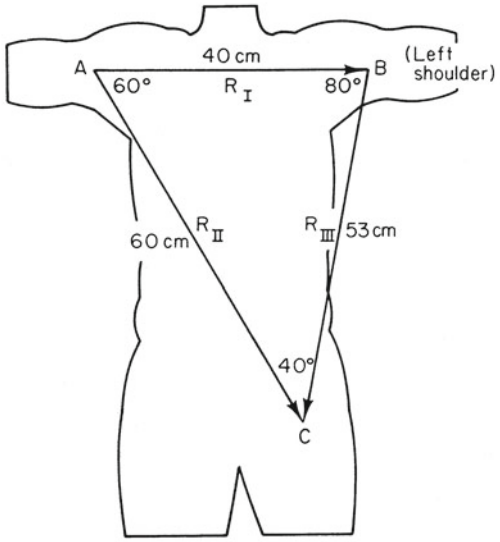


Fig. 7.19 Vectors connecting the three electrodes for a typical patient. The limbs are extensions of the leads of the electrocardiograph machine

is the same distance from each electrode. Both assumptions are wrong, of course, and later we will improve upon them.

The potential at **r** from a current dipole **p** is given by Eq. 7.13. The potential difference between two points at positions $\mathbf{r}_1$ and $\mathbf{r}_2$, each at a distance r from the dipole, is therefore (see Fig. 7.18)

$$v(\mathbf{r}_2, \mathbf{r}_1) = \frac{\mathbf{p} \cdot (\mathbf{r}_2 - \mathbf{r}_1)}{4\pi\sigma_o r^3}.$$

Denoting $\mathbf{r}_2 - \mathbf{r}_1$ by **R**, we have

$$v = \frac{\mathbf{p} \cdot \mathbf{R}}{4\pi\sigma_o r^3}. \tag{7.30}$$

The potential difference between two electrodes separated by a displacement **R** *and equidistant from the current-dipole vector* **p** *measures the instantaneous projection of vector* **p** *on* **R**.

If the depolarization can be described by a single current-dipole vector, only three measurements are needed in principle, corresponding to the projections on three perpendicular axes. The standard electrocardiogram (ECG) records 12 potential differences using nine electrodes. There are many reasons for this. The body is not an infinite, homogeneous conductor, and the relationship between cellular dipole moments and the potential is more complicated than our model; to convert the three perpendicular components to the instantaneous values of **p** would require a mathematical reconstruction; and the electrodes are not far away compared to the size of the heart. With 12 recorded potential differences, it is fairly easy to interpret the electrocardiogram by inspection.

The first three electrodes are placed on each wrist and the left leg. The limbs serve as extensions of the wires, so that the potential is measured where the limbs join the body. This is a major correction to our crude model that the heart is in an infinite conducting medium. If the subject were immersed in a conducting medium such as sea water, movement of the arms would change the size of the ECG signal because it would change **R**. In air, however, movement of the arms does not change the size of the signal. The simplest correction to explain this is to say that **R** for the two arm electrodes goes from shoulder to shoulder. These three electrodes measure potential differences between three points located approximately as shown in Fig. 7.19. The dimensions are for a typical adult. The three potential differences are called *limb leads* I, II, and III:

$$\begin{aligned} \mathrm{I} &= v_B - v_A, \\ \mathrm{II} &= v_C - v_A, \\ \mathrm{III} &= v_C - v_B. \end{aligned} \tag{7.31}$$

In the approximation used here, the voltage difference I is proportional to the projection of **p** on $\mathbf{R}_\mathrm{I}$, and so forth. These leads measure the projections of **p** on the three vectors $\mathbf{R}_\mathrm{I}$, $\mathbf{R}_\mathrm{II}$, and $\mathbf{R}_\mathrm{III}$ of Fig. 7.19.

It is customary also to combine these three potentials in a slightly different way to obtain projections of **p** on three other directions. These combinations are called the *augmented limb leads*. They contain no information that was not already present in the limb leads, but the six signals are easier to interpret by inspection. The combinations are

$$aVR = v_A - \frac{1}{2}(v_B + v_C) = -\frac{1}{2}(\mathrm{I} + \mathrm{II}),$$

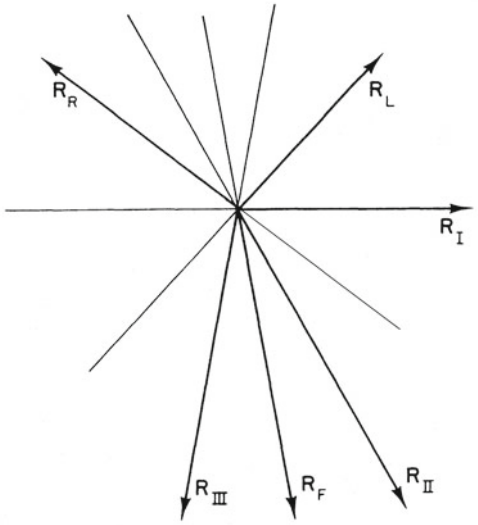


Fig. 7.20 The six directions in the frontal plane defined by the limb leads and the augmented limb leads. The angles are for the same subject as in Fig. 7.19

$$aVL = v_B - \frac{1}{2}(v_A + v_C) = \frac{1}{2}(\mathrm{I} - \mathrm{III}), \text{ and} \qquad (7.32)$$
$$aVF = v_C - \frac{1}{2}(v_A + v_B) = \frac{1}{2}(\mathrm{II} + \mathrm{III}).$$

These are proportional to the projections of $\mathbf{p}$ on vectors $\mathbf{R}_L$, $\mathbf{R}_R$, and $\mathbf{R}_F$ of Fig. 7.20. The subscripts refer to the fact that the vectors point toward the left shoulder, right shoulder, and foot, respectively.

The six lines in Fig. 7.20 are spaced approximately every 30° in the frontal plane. Many texts argue that the leads are spaced exactly every 30° and that the triangle of Fig. 7.19 is an equilateral triangle (Einthoven's triangle). While the directions are not far from 30°, this assumption is not really necessary. Physicians often want to know the direction of $\mathbf{p}$ at some point during the cardiac cycle, or the average direction of $\mathbf{p}$ during the QRS wave (ventricular depolarization). With six directions measured, this can be determined by inspection.

These six leads measure projections in the frontal plane. It is also necessary to have at least one projection in a plane perpendicular to the frontal plane. It is customary to place six leads across the chest wall in front of the heart; they are called the *precordial leads*. Their locations are shown in Fig. 7.21. The potential difference is measured between each precordial electrode and the average of v_A, v_B, and v_C. A lead therefore measures the projection of $\mathbf{p}$ on a vector from the center of triangle ABC to the electrode for that lead. This fact is not obvious, and in fact is true only if differences in $1/r^2$ are neglected. To see that it is true with the appropriate approximation, pick an arbitrary point O and from it construct vectors $\mathbf{R}_A$, $\mathbf{R}_B$, $\mathbf{R}_C$, and $\mathbf{R}_D$ to the points A, B, and C

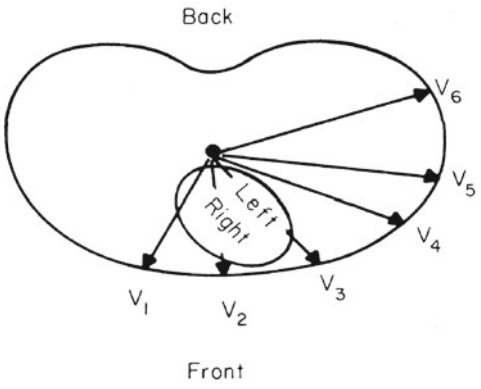


Fig. 7.21 The location of the precordial leads and the directions of the components of $\mathbf{p}$ which they measure. Reprinted with permission from Hobbie 1973. Copyright 1973, American Association of Physics Teachers

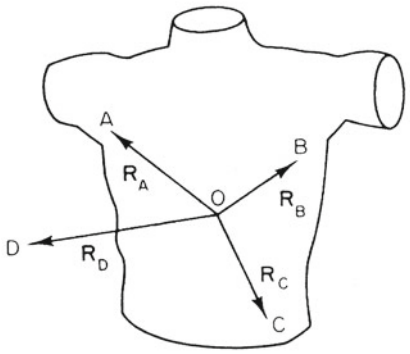


Fig. 7.22 A perspective drawing of the vectors used to calculate the potential in a precordial lead. Reprinted with permission from Hobbie 1973. Copyright 1973, American Association of Physics Teachers

of Fig. 7.22 and to the precordial electrode at D. The desired potential is

$$v = v_D - \frac{v_A + v_B + v_C}{3}.$$

It can be calculated using Eq. 7.30 for each term:

$$v = \frac{1}{4\pi\sigma_o}\left[\frac{\mathbf{p}\cdot\mathbf{R}_D}{R_D^3} - \frac{1}{3}\left(\frac{\mathbf{p}\cdot\mathbf{R}_A}{R_A^3} + \frac{\mathbf{p}\cdot\mathbf{R}_B}{R_B^3} + \frac{\mathbf{p}\cdot\mathbf{R}_C}{R_C^3}\right)\right].$$

So far, the location of O is arbitrary. If it is picked to be at the center of the triangle, then $\mathbf{R}_A + \mathbf{R}_B + \mathbf{R}_C = 0$. (This is the definition of center.) Since $R_A \approx R_B \approx R_C$, the term in large parentheses vanishes. The desired potential difference is then

$$v = \frac{1}{4\pi\sigma_o}\frac{\mathbf{p}\cdot\mathbf{R}_D}{R_D^3}.$$

In this approximation, each precordial lead measures the projection of $\mathbf{p}$ on a vector from the center of the triangle ABC to the electrode. The amplitude of the signal will be larger

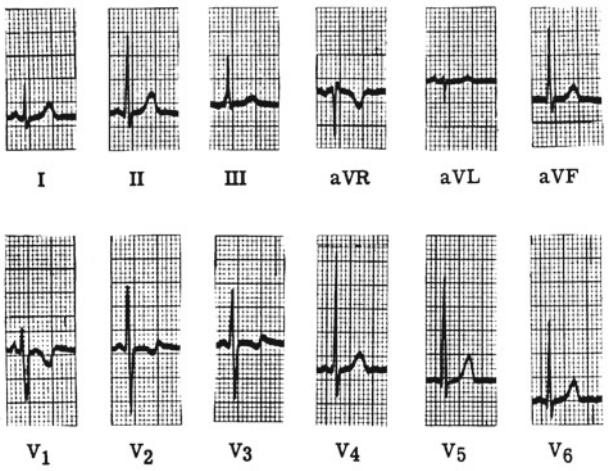


Fig. 7.23 A normal electrocardiogram. The large divisions are 0.5 mV vertically and 0.2 s horizontally. Reprinted with permission from Hobbie 1973. Copyright 1973, American Association of Physics Teachers. The electrocardiogram was supplied by Prof. James H. Moller, MD

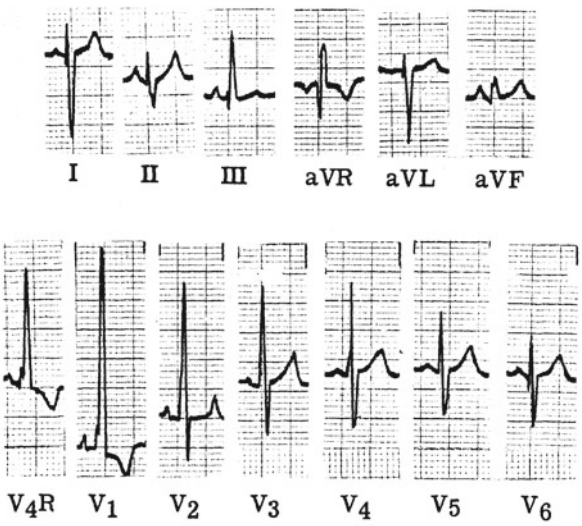


Fig. 7.24 The electrocardiogram of a patient with right ventricular hypertrophy. Reprinted with permission from Hobbie 1973. Copyright 1973, American Association of Physics Teachers. The electrocardiogram was supplied by Prof. James H. Moller, MD

than for the limb leads, because $R_D < R_A$. Some of the precordial leads are quite close to the heart. The assumption that r is the same for all parts of the myocardium is not valid. Because of the factor $1/r^2$, the greatest contribution to the potential comes from the closest regions of myocardium. A lead is said to "look at" the myocardium closest to it.

7.8 Some Electrocardiograms

A normal electrocardiogram is shown in Fig. 7.23. When **p** has its greatest magnitude during the QRS wave, it is nearly parallel to $\mathbf{R}_{\text{II}}$. There is almost no signal in aVL, which is perpendicular to $\mathbf{R}_{\text{II}}$.

Compare this to Fig. 7.24, which shows the electrocardiogram for a patient with *right ventricular hypertrophy*, an enlargement and thickening of the right ventricle. Because of the greater right ventricular muscle volume, **p** points to the right during the QRS wave, so that the QRS signal is negative in lead I. Lead aVF shows that there is very little vertical component of **p** during the QRS wave. The precordial leads V_1 and V_2 show the strongest signals, because the right ventricle faces the front of the body. In this case an extra lead V_4R has been used, which is symmetrical with V_4 but on the right side of the body.

The electrocardiogram in Fig. 7.25 is from a patient with left ventricular hypertrophy. The thicker left ventricular wall causes the QRS dipole to point to the left. As a result, lead I has an abnormally high peak, aVL is large and positive,

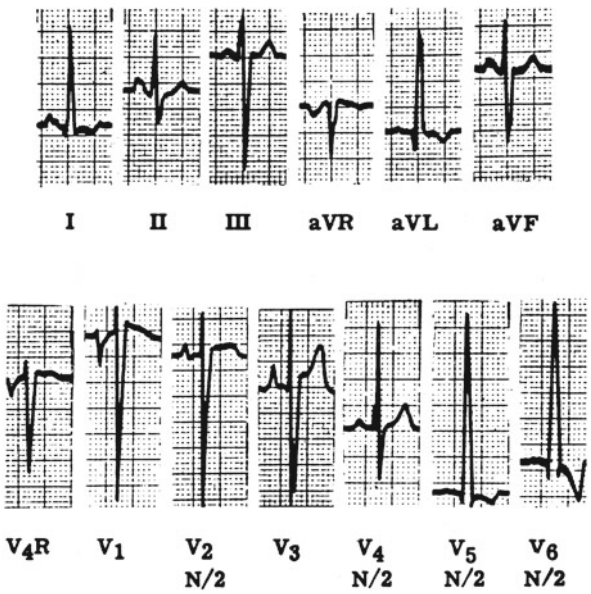


Fig. 7.25 The electrocardiogram for a patient with left ventricular hypertrophy. Reprinted with permission from Hobbie 1973. Copyright 1973, American Association of Physics Teachers. The electrocardiogram was supplied by Prof. James H. Moller, MD

V_2 is negative, and V_4, V_5, and V_6 have very large positive peaks. These last four leads are shown at half scale.

A fault in the conduction system known as a *bundle branch block* causes the depolarization wave to travel

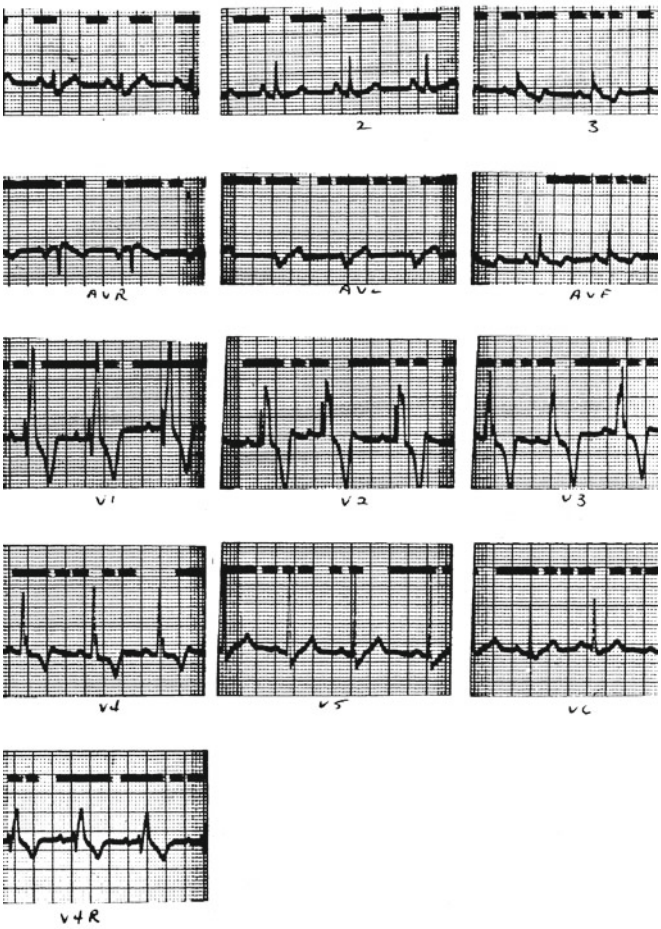


Fig. 7.26 The electrocardiogram for a patient with right bundle branch block. The electrocardiogram was supplied by Prof. James H. Moller, MD

through the myocardium rather than over the conduction system. Since the speed of propagation in myocardium is slower than that in the conduction system, the depolarization takes longer than usual. An electrocardiogram for a patient with *right bundle branch block* (a block in the bundle for the right ventricle) is shown in Fig. 7.26. The effect is most striking in leads that are most sensitive to the right ventricle: precordial leads 1 and 2. In V_1 the early part of the QRS wave has the usual biphasic, up–down pattern as the left ventricle depolarizes. This is followed by a large and prolonged vector pointing to the right, as the right ventricle slowly depolarizes. Lead V_2 shows a strong and prolonged bipolar signal as the right ventricle depolarizes.

7.9 Refinements to the Model

Our model for the potential outside a nerve or muscle cell has been a long single conducting fiber in an infinite, homogeneous medium. We will consider four ways to extend and improve the model. The first is to recognize that current must also flow radially inside the cell. If it did not, it could never leave the cell. At the same time we will abandon the assumption that the presence of the cell along the x axis does not perturb the current outside the cell. The third improvement is to recognize that the conductivity may depend on position. This is particularly important outside the cell, where there are muscle, fat, lungs, etc. Finally, the conductivity at a given point may depend on which direction the current flows–for example, parallel or perpendicular to the cells.

In order to make these refinements to the model, we must develop a different formulation of the problem. Consider some region of space containing a conducting material described by Ohm's law. The electric field is related to the potential by Eq. 6.16b: $\mathbf{E} = -\text{grad}\, v = -\nabla v$. If the material is isotropic and obeys Ohm's law, then from Eq. 6.26

$$\mathbf{j} = \sigma \mathbf{E} = -\sigma \nabla v. \tag{7.33}$$

We now apply the equation of continuity or conservation of charge, casting Eq. 4.8 in terms of the electric current density $\mathbf{j}$ and the electric charge per unit volume, ρ:

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot \mathbf{j}. \tag{7.34}$$

Combining these two equations gives

$$\frac{\partial \rho}{\partial t} = \text{div}(\sigma\ \text{grad}\ v) = \nabla \cdot (\sigma \nabla v). \tag{7.35}$$

Leaving the conductivity inside the divergence term allows the conductivity to depend on position. If the conductivity is the same everywhere it can be taken outside the divergence operator to give

$$\frac{\partial \rho}{\partial t} = \sigma \nabla^2 v = \sigma \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right). \tag{7.36a}$$

We can write this in cylindrical coordinates, which are more useful for modeling a cylindrical cell stretched along the z axis. From Appendix L, assuming that the potential does not depend on the angle ϕ, we have

$$\frac{\partial \rho}{\partial t} = \sigma \nabla^2 v = \sigma \left[\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial v}{\partial r} \right) + \frac{\partial^2 v}{\partial z^2} \right]. \tag{7.36b}$$

These are very general equations, applicable to any volume of space where the material is homogeneous and isotropic and obeys Ohm's law. They were derived using Ohm's law and the conservation of charge. Equation 7.36a is actually the same result we had in Eq. 6.51. This is demonstrated in Problem 29.

7.9.1 The Fiber Has a Finite Radius

Now we can make the first two improvements: we relax the assumption that the fiber radius is very small. Except at the cell membrane, where charge on the membrane capacitance is changing as the membrane potential changes, $\partial\rho/\partial t = 0$. If we assume that the transmembrane potential v_m is known, then Eq. 7.36b can be applied separately to the extracellular and the intracellular fluid for a long straight fiber to determine the potential everywhere outside (or inside). This was first done by Clark and Plonsey (1968). In the extracellular and intracellular fluids, Eq. 7.36b becomes

$$\begin{aligned} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_o(r,z)}{\partial r}\right)+\frac{\partial^2 v_o(r,z)}{\partial z^2}=0, \quad r>a \\ \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_i(r,z)}{\partial r}\right)+\frac{\partial^2 v_i(r,z)}{\partial z^2}=0, \quad r<a \\ v_m(z)=v_i(a,z)-v_o(a,z). \end{aligned} \tag{7.37}$$

With v_m known, these equations were solved for the potential distribution inside and outside the cell. This is the calculation that was done to obtain Fig. 7.13. The result of this type of calculation has been compared to the line-source model by Trayanova et al. (1990).

7.9.2 Nonuniform Exterior Conductivity

To make the next improvement, consider an extracellular region in which the conductivity is not uniform. In a region without sources, the potential obeys

$$\nabla\cdot(\sigma_o\nabla v_o)=0. \tag{7.38}$$

Often, the conductivity is assumed to be "piecewise" homogenous, with a different value assigned to each kind of tissue. Within each tissue the potential then obeys Laplace's equation, $\nabla^2 v_o = 0$. At the boundary between tissues, the potential and the normal component of the current are continuous.

When the different tissues have realistic and irregular boundaries, special techniques are needed to solve Laplace's equation. One important technique is the *finite-element method* (Miller and Henriquez 1990); another is the *boundary-element method* (Gulrajani 1998).

A typical application, which serves as the basis for *noninvasive electrocardiographic imaging*, is to measure the potential at the body surface and then calculate the potential on the epicardium (the outer surface of the heart; Rudy and Burnes 1999; Stanley et al. 1986). One cannot calculate the potential inside the heart unless the sources are known, but finding the potential on the epicardial surface is possible.

7.9.3 Anisotropic Conductivity: The Bidomain Model

The final improvement recognizes that the cardiac tissue is generally not isotropic. If it is still described by Ohm's law, then we can write $\mathbf{j}=\widetilde{\sigma}\cdot\mathbf{E}$ where $\widetilde{\sigma}$ is a matrix or tensor. In Cartesian coordinates

$$\begin{pmatrix} j_x \\ j_y \\ j_z \end{pmatrix} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}. \tag{7.39}$$

This is a compact notation for

$$j_x=\sigma_{xx}E_x+\sigma_{xy}E_y+\sigma_{xz}E_z,$$

with similar equations for j_y and j_z. It can be shown that the conductivity matrix must be symmetric, so there are actually six conductivity coefficients, not nine. It is often possible to make some of the matrix elements zero by suitable choice of a coordinate system and suitable orientation of the axes.

Problem 29 shows that for a small cylindrical region of isotropic axoplasm of length h and radius a, the cylindrical surface of which is surrounded by cell membrane, the total charge Q within the axoplasm changes according to

$$\frac{\partial Q}{\partial t}=\pi a^2 h\frac{\partial\rho_i}{\partial t}=C\frac{\partial v_m}{\partial t}+i_m=2\pi a h\left(c_m\frac{\partial v_m}{\partial t}+j_m\right),$$

or

$$c_m\frac{\partial v_m}{\partial t}+j_m=\frac{\pi a^2 h}{2\pi a h}\sigma_i\frac{\partial^2 v_i}{\partial x^2}=\frac{\sigma_i a}{2}\frac{\partial^2 v_i}{\partial x^2}.$$

This can also be written as

$$\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right)=\sigma_i\frac{\partial^2 v_i}{\partial x^2},$$

where $\beta = 2\pi a h/\pi a^2 h = 2/a$ is the ratio of surface area to volume of the cell. Our cell was cylindrical. With other geometrical configurations, such as a cubic or a spherical cell, β would have a different value, but it always has the dimensions of $(\text{length})^{-1}$. In the general three-dimensional anisotropic case, the equivalent equation is

$$\underbrace{\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right)}_{\text{zero, except at the cell membrane}} = \operatorname{div}(\widetilde{\sigma}_i\ \operatorname{grad} v_i)=\nabla\cdot(\widetilde{\sigma}_i\nabla v_i). \tag{7.40}$$

Both σ_i and v_i are functions of position. The left-hand side is zero except at the cell membrane. The main theme of this chapter has been that current that stops flowing inside the cell must flow outside the cell. We can write an analogous equation for the region outside the cell:

$$\underbrace{-\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right)}_{\text{zero, except at the cell membrane}} = \nabla\cdot(\widetilde{\sigma}_o\nabla v_o). \tag{7.41}$$

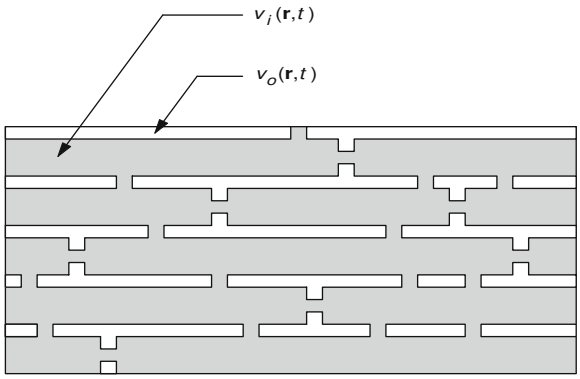


Fig. 7.27 The interior of myocardial cells (*shaded*) is connected to adjoining cells by gap junctions. The bidomain model assumes that in a small region of space (large compared to a cell) there are two potentials: the interior potential and outside potential that are functions of position and time

Myocardial cells are typically about 10 μm in diameter and 100 μm long. They have the added complication that they are connected to one another by gap junctions, as shown schematically in Fig. 7.27. This allows currents to flow directly from one cell to another without flowing in the extracellular medium. The *bidomain* (two-domain) model is often used to model this situation (Henriquez 1993; Henriquez and Ying 2009). It considers a region, small compared to the size of the heart, that contains many cells and their surrounding extracellular fluid. It simplifies the problem by assuming that each small volume element contains two domains, intracellular and extracellular. Think of the volume element as the entire region shown in Fig. 7.27. There are two potentials in each small volume element: $v_i(\mathbf{r},t)$ and $v_o(\mathbf{r},t)$. These potentials are averages over the intracellular and extracellular domains contained in the volume element. The transmembrane potential is the difference between these two potentials: $v_m(\mathbf{r},t) = v_i(\mathbf{r},t) - v_o(\mathbf{r},t)$. Charge can pass between the two domains, but the total charge within a volume element is conserved. If the current densities in each domain are $\mathbf{j}_i$ and $\mathbf{j}_o$, then the divergence of the sum is zero: $\nabla\cdot(\mathbf{j}_i+\mathbf{j}_o)=0$. The divergence of each current individually passes through the membrane or charges the membrane capacitance. The anisotropic analogs of Eqs. 7.40 and 7.41 are now

$$\begin{aligned}\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right) &= \operatorname{div}(\widetilde{\sigma}_i\cdot\operatorname{grad}v_i)=\nabla\cdot(\widetilde{\sigma}_i\cdot\nabla v_i),\\ -\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right) &= \operatorname{div}(\widetilde{\sigma}_o\cdot\operatorname{grad}v_o)\\ &= \nabla\cdot(\widetilde{\sigma}_o\cdot\nabla v_o).\end{aligned} \tag{7.42}$$

The quantity β is the membrane surface area per unit volume of the entire bidomain—both intracellular and extracellular volumes. For example, if we consider that the cells are all cylindrical of length h and radius a, then the surface area of a cell is $2\pi ah$. If the fraction of the total volume occupied by cells is f, then the total volume associated with this cell is $\pi a^2 h/f$, so

$$\beta=\frac{2f}{a}. \tag{7.43}$$

The membrane current j_m can be modeled by either a passive membrane (Ohm's law—electrotonus) or with one of the models for an active membrane.

Anisotropy plays an important role in the bidomain model. To see why, consider a solution to Laplace's equation in a monodomain—a two-dimensional sheet of homogeneous, anisotropic tissue with straight fibers. If the x direction is chosen to be along the fiber direction (the direction of greatest conductivity), then Laplace's equation becomes

$$\sigma_{ox}\frac{\partial^2 v_o}{\partial x^2}+\sigma_{oy}\frac{\partial^2 v_o}{\partial y^2}=0.$$

Now define a new set of coordinates $x'=x$ and $y'=\sqrt{\sigma_{ox}/\sigma_{oy}}\,y$. You can show that in these new coordinates Laplace's equation becomes

$$\frac{\partial^2 v_o}{\partial x'^2}+\frac{\partial^2 v_o}{\partial y'^2}=0.$$

We have removed the effect of anisotropy by rescaling distance in the direction perpendicular to the fibers. If you try a similar trick with the bidomain model

$$\sigma_{ix}\frac{\partial^2 v_i}{\partial x^2}+\sigma_{iy}\frac{\partial^2 v_i}{\partial y^2}=\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right) \tag{7.44a}$$

$$\sigma_{ox}\frac{\partial^2 v_o}{\partial x^2}+\sigma_{oy}\frac{\partial^2 v_o}{\partial y^2}=-\beta\left(c_m\frac{\partial v_m}{\partial t}+j_m\right), \tag{7.44b}$$

you can find a new coordinate system that removes the effect of anisotropy in either the intracellular space or the extracellular space, but in general you cannot find a coordinate system that removes the anisotropy in both spaces simultaneously (Roth 1992). Only in the special case of equal anisotropy ratios ($\sigma_{ix}/\sigma_{iy}=\sigma_{ox}/\sigma_{oy}$) will the equations simplify dramatically. But the anisotropy ratios in the heart are not equal. In the intracellular space the ratio of conductivities parallel and perpendicular to the fibers is about 10:1, while in the extracellular space this ratio is about 4:1 (Roth 1997). Anisotropy plays an essential role in the electrical behavior of the heart, especially during electrical stimulation.

7.10 Electrical Stimulation

The information that has been developed in this chapter can also be used to understand some of the features of stimulating electrodes. These may be used for electromyographic

studies; for stimulating muscles to contract called *functional electrical stimulation* (Peckham and Knutson 2005); for a *cochlear implant* to partially restore hearing (Zeng et al. 2008); deep brain stimulation for Parkinson's disease (Perlmutter and Mink 2006); for cardiac pacing (Moses and Mullin 2007); and even for defibrillation (Dosdall et al. 2009). The electrodes may be inserted in cells, placed in or on a muscle, or placed on the skin.

A pulse of current is sent to the stimulating electrode. The current required to produce a response depends on the shape and size of the electrode, its placement, the kind of cell being stimulated, and the duration of the pulse. For a given electrode geometry the shorter the pulse, the larger the current required for a tissue response. For very long pulses there is a minimum current required to stimulate that is called *rheobase*. The *strength-duration curve* was first discovered by Weiss in 1901. He expressed it in terms of total charge in the stimulating pulse. A description of the strength-duration curve and its history has been given by Geddes and Bourland (1985). They also describe some techniques for making accurate measurements. The strength-duration curve for current was first described by Lapicque (1909) as

$$i = i_R \left(1 + \frac{t_C}{t}\right), \tag{7.45}$$

where i is the current required for stimulation, i_R is the rheobase, t is the duration of the pulse, and t_C is chronaxie, the duration of the pulse that requires twice the rheobase current.

Equation 7.45 provides an empirical fit to the experimental data. We can develop a model to explain it using information from Chap. 6. A nerve fires after a certain departure from the resting potential. Subthreshold behavior can be modeled by electrotonus. Suppose that we inject a stimulating current into a cell at the origin. Equation 6.58 gave the voltage along the axon for a current injected in the cell at the origin after an infinitely long time: $v - v_r = v_0 e^{-|x|/\lambda}$. The solution to Problem 6.34 shows that the current injected is

$$i_0 = 2v_0/\lambda r_i. \tag{7.46}$$

The quantities λ and r_i are defined in Chap. 6. The factor of 2 arises because current flows both ways along the cell. The rheobase current is

$$i_R = 2\frac{v_{\text{threshold}}}{\lambda r_i}. \tag{7.47}$$

If we assume that the threshold voltage is independent of pulse duration, we can use the curve for $x = 0$ in Fig. 6.31c to relate the minimum current to the pulse duration. As long as the pulse is applied, the voltage will rise along this curve. When the current is turned off, the voltage will start to fall. If the voltage reaches threshold, the cell will fire. This

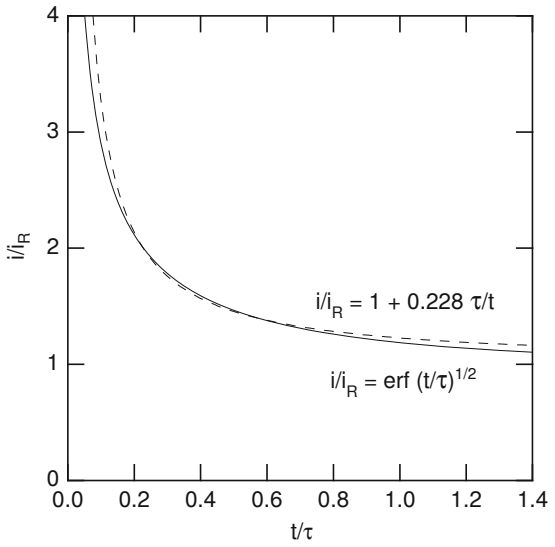


Fig. 7.28 The stimulus strength-duration curve plotted for the chronaxie–rheobase model, Eq. 7.45 and for electrotonus, Eq. 7.49

curve is the solution of Eq. 6.55. The solution is (Chap. 6, Problem 36; Plonsey 1969, p. 132))

$$v(0,t) - v_r = v_0 \operatorname{erf}\left(\sqrt{\frac{t}{\tau}}\right), \tag{7.48}$$

where τ is the membrane time constant, $\kappa\epsilon_0\rho_m$. The error function is defined in Eq. 4.74 and is plotted in Fig. 4.21. The current required for stimulation with an intracellular electrode at the origin is therefore

$$i = \frac{2v_{\text{threshold}}}{\lambda r_i \operatorname{erf}(\sqrt{t/\tau})} = \frac{i_R}{\operatorname{erf}(\sqrt{t/\tau})}. \tag{7.49}$$

Chronaxie can be related to the time constant τ by setting $i = 2i_R$:

$$2i_R = \frac{i_R}{\operatorname{erf}\left(\sqrt{t_C/\tau}\right)}. \tag{7.50}$$

From a table of values of the error function, we find

$$t_C = 0.228\tau. \tag{7.51}$$

Figure 7.28 compares the standard empirical curve, Eq. 7.45, with this model. The curves are experimentally indistinguishable.

Equation 7.45 is also used for surface electrodes. Table 7.1 shows some experimental values for rheobase and chronaxie. The further the electrode from the tissue being stimulated, the greater the rheobase current that is required.

An electrode that is transferring positive charge to the medium is called an *anode*. One that is collecting positive charge is called a *cathode*. If the stimulating electrode is

Table 7.1 Comparison of values for rheobase and chronaxie for different stimulations

Stimulation	Rheobase (mA)	Chronaxie (ms)
Intracellular, from Table 6.1, $v_{\text{threshold}} = 15\,\text{mV}$	6.7×10^{-6}	0.23
Myocardium, from good pacing electrodes	0.1	
Motor nerves for inspiration, from stimulation of chest wall (Voorhees et al. 1992)	49	0.17
Myocardium, from stimulation of chest wall (Voorhees et al. 1992)	204	1.82

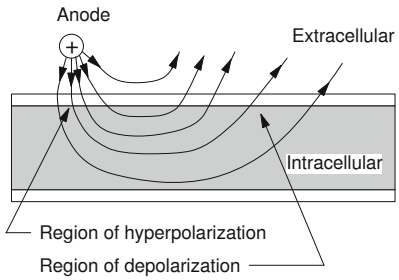


Fig. 7.29 A schematic drawing showing why there is a region of hyperpolarization near a stimulating anode (positive electrode) with a region of weaker depolarization further away

inside the cell, a positive current leaving the electrode will increase the positive charge within the cell and depolarize it. Another way to say it is that current from the electrode flows out through the membrane, so the inside of the membrane will be made more positive than the outside. On the other hand, an anodic electrode just outside the cell will send positive current in through the membrane near the electrode, as shown in Fig. 7.29. This lowers the potential inside and hyperpolarizes the membrane near the electrode. Further away from the stimulation point will be a region where current flows out through the membrane, thus depolarizing the cell. However, the outward current is in general spread out over more membrane, so the current density and hence the depolarization is less than the hyperpolarization near the anode. The situation is, of course, reversed for a cathodic electrode. Figure 7.29 is conceptual; to draw the field lines accurately would require taking into account the conductivities of the extracellular and intracellular fluid as well as the membrane.

The electrotonus model also helps us understand another effect that is observed: the *virtual cathode*. The point of origin for a stimulus can be measured by placing sensing electrodes in or on the heart at different distances from the stimulating electrode and plotting the time required for the depolarization wave front to reach the electrode vs. its position. Extrapolation to the time of stimulus gives the size of the region of initial depolarization. Imagine a stimulating electrode inside a one-dimensional cell. When the stimulus current is just above rheobase, the region of depolarization is very small and surrounds the electrode. As the stimulating current is increased, the size of the initial depolarized region grows. From Eqs. 6.58 and 7.50 we obtain

$$v_{\text{threshold}} = \frac{i_0 \lambda r_i}{2} e^{-x_{\text{vc}}/\lambda}$$

or

$$x_{\text{vc}} = \lambda \ln\left(\frac{i_0 \lambda r_i}{2 v_{\text{threshold}}}\right) = \lambda \ln\left(\frac{i_0}{i_R}\right), \tag{7.52}$$

where x_{vc} is the size of the virtual cathode.

Cardiac pacemakers are a useful treatment for certain heart diseases (Jeffrey 2001; Moses and Mullin 2007; Barold 1985). The most frequent are an abnormally slow pulse rate (*bradycardia*) associated with symptoms such as dizziness, fainting (*syncope*), or heart failure. These may arise from a problem with the SA node (*sick sinus syndrome*) or with the conduction system (*heart block*). One of the first uses of pacemakers was to treat complete or *third degree* heart block. The SA node and the atria fire at a normal rate but the wave front cannot pass through the conduction system. The AV node or some other part of the conduction system then begins firing and driving the ventricles at its own, pathologically slower rate. Such behavior is evident in the ECG in Fig. 7.30, in which the timing of the QRS complex from the ventricles is unrelated to the P wave from the atria. A pacemaker stimulating the ventricles can be used to restore a normal ventricular rate.

A pacemaker can be used temporarily or permanently. The pacing electrode can be threaded through a vein from the shoulder to the right ventricle (*transvenous pacing*, Fig. 7.31) or placed directly in the myocardium during heart surgery. Sometimes two pacing electrodes are used, one in the atrium and one in the ventricle. The pacing electrode can be unipolar or bipolar. With a unipolar electrode, the stimulation current flows into the myocardium and returns to the case of the pacemaker, which is often placed in a pocket in the muscle of the chest wall near the shoulder. The return current in a bipolar electrode goes to a ring electrode a few centimeters back along the pacing lead from the electrode at the tip. The surface area of a typical tip is about 10 mm^2 ($10^{-5}\,\text{m}^2$). The current density required to initiate depolarization depends on the spatial distribution of the current and is approximately 100 A m^{-2}. Thus, in this model the current is about 1 mA.[6] The resistance of the tissue is typically 500 Ω, so the voltage is 0.5 V. After the pacing electrode is implanted, the size of the voltage pulse required to initiate ventricular activity rises because inflammatory tissue grows around the electrode. It is conducting, but the myocardium is further away, and the

[6] Acute implants of smaller electrodes where the electrode resistance is low, as well as computer simulations, have shown stimulation with currents as small as 18 μA (Lindemans and Denier van der Gon 1978).

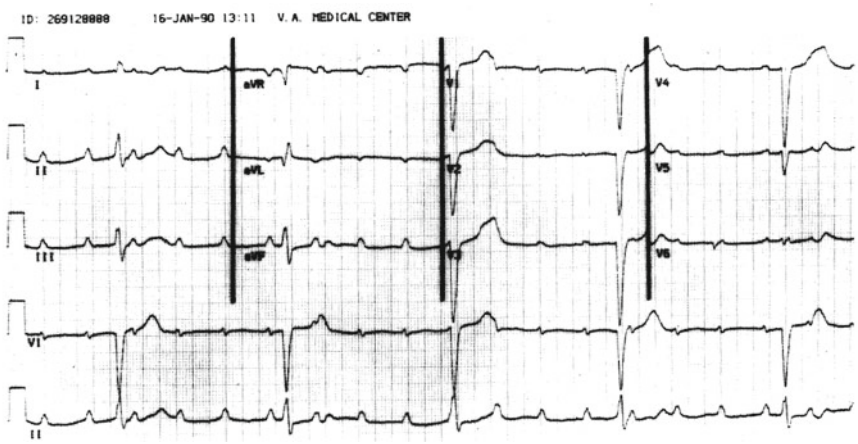


Fig. 7.30 A patient with 3rd degree AV heart block. (From Rardon et al. 2000. Used by permission)

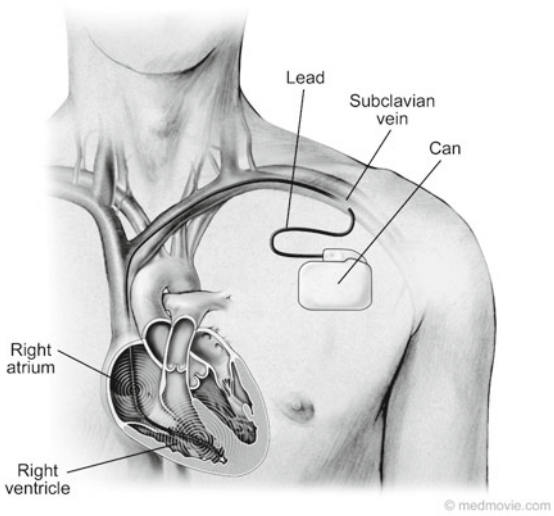


Fig. 7.31 An implanted pacemaker or defibrillator. The battery and electronics are in a sealed container (the "can") placed under the skin near the left shoulder. The electrode or "lead" is threaded through the subclavian vein and right atrium into the right ventricle. (Image © Copyright by medmovie.com. Used by permission of medmovie.com)

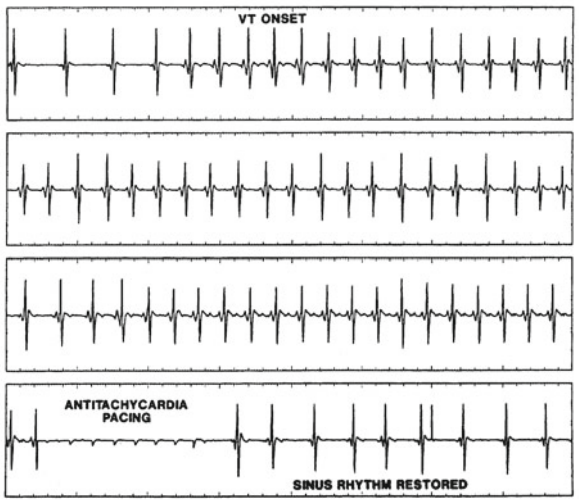


Fig. 7.32 The top strip shows the onset of ventricular tachycardia, which persists in the next two strips. Very rapid pacing in the fourth strip restores a normal sinus rhythm. (Source: Mitrani et al. 1995. Used by permission)

inflammatory tissue effectively increases the size of the electrode, thereby reducing the current density. After 6 months or so, the inflammation has been replaced by a small fibrous capsule, resulting in an effective electrode size larger than the bare electrode but smaller than the region of inflammation. Electrodes that elute steroids have been used to reduce the inflammation.

Pacemakers can also be designed to detect an abnormal rhythm and apply an electrical stimulus to reverse it. Fig. 7.32 shows a patient with ventricular tachycardia due to a reentrant circuit (p. 194, 290) which has been corrected by pacing very rapidly so that the refractory period prevents the propagation of the reentrant wave. *Ventricular fibrillation* occurs when the ventricles contain many interacting reentrant wavefronts that propagate chaotically. Fibrillation is discussed in greater detail in Chap. 10. During fibrillation the ventricles no longer contract properly, blood is no longer pumped through the body, and the patient dies in a few minutes. Implantable defibrillators are similar to pacemakers, but are slightly larger. An implanted defibrillator continually measures the ECG. When a signal indicating fibrillation is sensed, it delivers a much stronger shock that can eliminate the reentrant wavefronts and restore normal heart rhythm (Fig. 7.33).

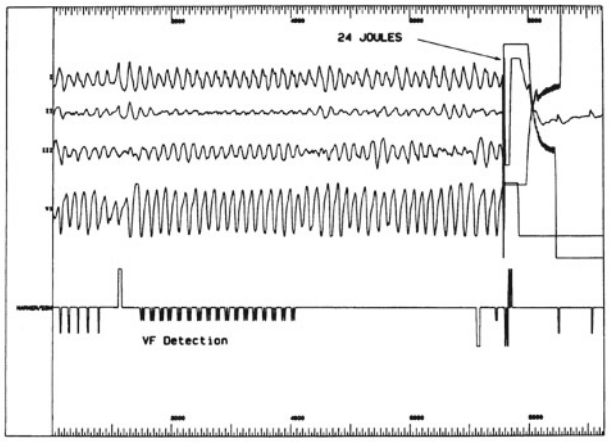


Fig. 7.33 Ventricular fibrillation has been induced in the electrophysiology laboratory. A pacemaker cardioverter-defibrillator detects the ventricular fibrillation. A capacitor is then charged and applies a 24-joule defibrillation pulse that restores normal rhythm. (Source: Mitrani et al. 1995)

The bidomain model has been used to understand the response of cardiac tissue to stimulation (Janks and Roth 2009; Trayanova and Plank 2009). This model explains a remarkable experimental observation. Although the speed of the wave front is greater along the fibers than perpendicular to them, if the stimulation is well above threshold, the wavefront originates farther from the cathode in the direction perpendicular to the fibers—the direction in which the speed of propagation is slower. The simulations show that this is due to the anisotropy in conductivity. This is called the *dog-bone* shape of the virtual cathode. It can rotate with depth in the myocardium because the myocardial fibers change orientation. The difference in anisotropy accentuates the effect of a region of hyperpolarization (a virtual anode) adjacent to the depolarization region produced by a cathodic electrode. This hyperpolarization can shorten the refractory period of the tissue, thereby creating new excitable paths through which reentrant wave fronts can propagate (Wikswo and Roth 2009; Ripplinger and Efimov 2009).

One of the fundamental problems with research in this area can be seen in equations like Eq. 7.41. The variable on the left is the transmembrane potential v_m. The variable on the right is the potential inside or outside the cell. Measurement of v_m requires measurement or calculation of the difference $v_i - v_o$. Experimental measurements of the transmembrane potential often rely on the use of a voltage-sensitive dye whose fluorescence changes with the transmembrane potential (*optical mapping*) (Rosenbaum and Jalife 2001).

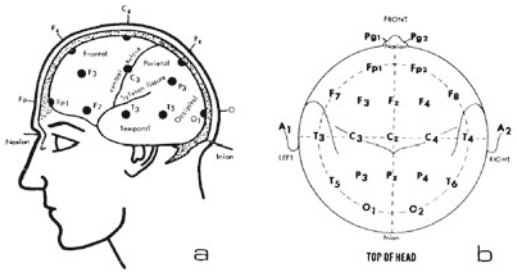


Fig. 7.34 The standard "10–20" arrangement of electrodes on the scalp for the EEG. (Courtesy of Natus Neurology Grass brand products)

7.11 The Electroencephalogram

Much can be learned about the brain by measuring the electric potential on the scalp surface. Such data are called the *electroencephalogram* (EEG). Nunez and Srinivasan (2005) have written an excellent book about the physics of the EEG. We briefly examine the topic here. The EEG is used to diagnose brain disorders, to localize the source of electrical activity in the brain in patients who have epilepsy (Lopes da Silva 2008), and as a research tool to learn more about how the brain responds to stimuli (*evoked responses*) and how it changes with time (*plasticity*). Typically, the EEG is measured from 21 electrodes attached to the scalp according to the *10–20 system* (Fig. 7.34). A typical signal from an electroencephalographic electrode is shown in the top panel of Fig. 11.39. One difficulty in interpreting the EEG is the lack of a suitable reference electrode. None of the 21 electrodes in Fig. 7.34 qualifies as a distant ground against which all other potential recordings can be measured. One way around this difficulty is to subtract from each measured potential the average of all the measured potentials. In the problems, you are asked to prove that this *average reference recording* does not depend on the choice of reference electrode; it is a reference-independent method.

Symbols Used in Chapter 7

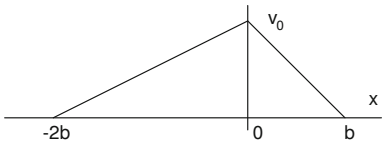
Symbol	Use	Units	First used page
a	Axon radius	m	187
c_m	Membrane capacitance per unit area	F m^{-2}	200
f	Intracellular volume fraction		201
h	Length of segment	m	200
i	Current	A	186
i_i, i_o	Current inside, outside axon	A	186
i_R	Rheobase current	A	202
j, $\mathbf{j}$	Current density	A m^{-2}	186
j_m	Current density through membrane	A m^{-2}	200
p, $\mathbf{p}$	Activity vector or current dipole moment	A m	188
q	Charge	C	186
r_i	Resistance per unit length inside axon	Ω m^{-1}	202
r, $\mathbf{r}$	Distance	m	186
t	Time	s	199
t_C	Chronaxie	s	202
v	Potential	V	186
v_i, v_o	Potential inside, outside axon	V	200
v_m	Potential across membrane	V	200
x, y, z, x_0, x_1, x_2, y_0	Distance or position	m	187
x_{vc}	Size of virtual cathode	m	203
C	Capacitance	F	200
E, $\mathbf{E}$	Electric field	V m^{-1}	186
P_n	Legendre Polynomial		189
Q	Electric charge	C	200
R	Resistance	Ω	187
R, $\mathbf{R}$	Distance or position	m	190
β	Ratio of surface area to volume	m^{-1}	200
ϵ_0	Permittivity of free space	N^{-1} m^{-2} C^2	186
λ	Space constant	m	202
σ, σ_i, σ_o	Electrical conductivity	S m^{-1}	186
ρ	Charge density	C m^{-3}	199
τ	Time constant	s	202
θ	Angle		188
ξ	Ratio of x to R		192
Ω	Solid angle		189

Problems

Section 7.1

Problem 1. A single nerve or muscle cell is stretched along the x axis and embedded in an infinite homogeneous medium of conductivity σ_o. Current i_0 leaves the cell at $x = b$ and enters the cell again at $x = -b$. Find the current density $\mathbf{j}$ at distance r from the axis in the $x = 0$ plane.

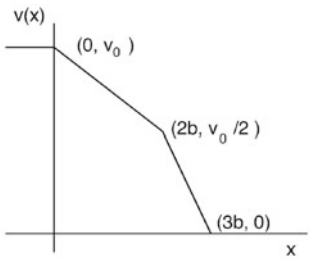
Problem 2. An axon is stretched along the x axis. At one instant of time an impulse traveling along the axon has the form shown in the graph. The electrical conductivity inside the axon of radius a is σ_i. In the infinite external medium it is σ_o. Find an expression for the potential at point (x_0, y_0).



Problem 3. The interior potential of a cylindrical cell is plotted at one instant of time. Distances along the cell are given in terms of length b. The cell has radius a and electrical conductivity σ_i. The resting potential is 0 and the depolarized potential is v_0. The conductivity of the external medium is σ_o.

(a) Find expressions for, and plot, the current along the cell in the four regions $(x < 0, 0 < x < 2b, 2b < x < 3b, 3b < x)$.

(b) Find the potential at a point (x, y) outside the cell in terms of the parameters given in the problem. The point is not necessarily far from the cell.



Section 7.2

Problem 4. Modify the closing argument of Sect. 7.2 by considering electrodes that are disks rather than spheres. (Hint: The capacitance you will need is given in Sect. 6.19.)

Problem 5. Suppose an axon is surrounded by a thin layer of extracellular fluid of thickness d. Use arguments based on the intracellular and extracellular resistances to estimate the ratio $\Delta v_o / \Delta v_i$ in this case.

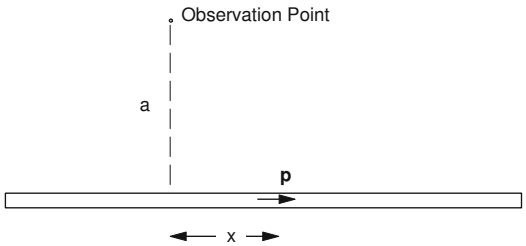
Section 7.3

Problem 6. Starting with Eq. 7.4, make the Taylor's series expansions described in the text, and use them to derive Eq. 7.16.

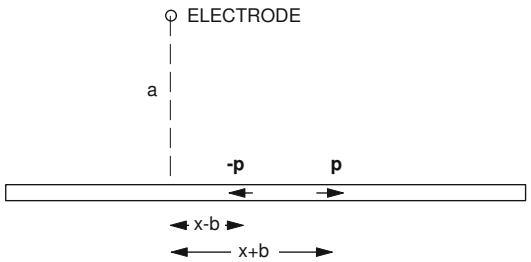
Problem 7. What would be the current-dipole moment of a nerve cell of radius 2 μm when it depolarizes? Would myelination make any difference? Does the result depend on the rise time of the depolarization? If the impulse lasts 1 ms and the conduction speed is 5 m s^{-1}, how far apart are the rising and falling edges of the pulse?

Problem 8. An axon or muscle cell is stretched along the x axis on either side of the origin. As it depolarizes, a constant

current dipole **p** pointing to the right sweeps along the axis with velocity u. An electrode at $(x = 0, y = a)$ measures the potential with respect to $v = 0$ at infinity. Ignore repolarization. Find an expression for v at the electrode as a function of time and sketch it. Assume that at $t = 0$, **p** is directly under the electrode at $x = 0$.



Problem 9. An electrode at $(x = 0, y = a)$ measures the potential outside an axon with respect to $v = 0$ at infinity. A nerve impulse is at point x along the axon, measured from the perpendicular from the electrode to the axon. At $x + b$ a current dipole points to the right, representing the depolarization wave front. At $x - b$ a vector of the same magnitude points to the left, representing repolarization. Obtain an expression for v as a function of x, b, p, and a. Plot it in the case $a = 1$, $b = 0.05$.



Problem 10. A dipole **p** located at the origin (0, 0, 0) is oriented in the x direction. The potential $v_o(x)$ produced by this dipole is measured along the line $y = 0$, $z = d$.

(a) Find an equation for $v_o(x)$ in terms of x, d, σ_o (the conductivity of the medium) and the dipole strength, p.

(b) Find an expression for the depth d of the dipole in terms of the distance Δx, defined as the distance between the minimum and maximum of $v_o(x, y)$. This is an example of an "inverse problem," in which you try to learn about the source (in this case, the depth of **p**) from measurements of v_o.

Problem 11. The *solid angle theorem* is often used to interpret electrocardiograms. The relationship between the exterior potential and the solid angle in Eq. 7.15 is a general result: the potential is proportional to the solid angle subtended by the wave front. Use this result to explain (a) why a closed wave front produces no exterior potential, and (b) why an open wave front produces a potential that depends only on the geometry of its opening or rim.

Section 7.4

Problem 12. Run the program of Fig. 7.11 and plot the potential for different distances from the axon.

Problem 13. Modify the program of Fig. 7.11 to calculate the potential from a single Gaussian action potential and plot the potential.

Problem 14. Let the intracellular potential be zero except in the range $-a < x < a$, where it is given by

$$v_i = \begin{cases} 2\left(\dfrac{a+x}{a}\right)^2, & -a < x < -a/2 \\ 1 - 2\left(\dfrac{x}{a}\right)^2, & -a/2 < x < a/2 \\ 2\left(\dfrac{a-x}{a}\right)^2, & a/2 < x < a. \end{cases}$$

Plot v_i vs x. Use Eq. 7.21 to calculate the exterior potential at (x_0, y_0). You may need the integral

$$\int \frac{dx}{\sqrt{x^2 + b^2}} = \sinh^{-1}\left(\frac{x}{b}\right).$$

Section 7.5

Problem 15. Suppose a wave front propagates at a speed of 0.25 m s^{-1} and its refractory period lasts 250 ms. Calculate the minimum path length of its reentrant circuit. Most reentrant wave fronts are somewhat slower and briefer than this, so their paths may be shorter.

Section 7.7

Problem 16. Two electrodes are placed in a uniform conducting medium 10 cm from a cell of radius 5 μm and 10 cm from each other, so that the two electrodes and the cell form an equilateral triangle. When the cell depolarizes the potential rises 90 mV. What will be the potential difference between the two electrodes when the cell orientation is optimum? How many cells would be needed to give a potential difference of 1 mV between the electrodes? Assume $\sigma_i/\sigma_o = 10$.

Problem 17. Guess whatever parameters you need to predict the voltage at the peak of the QRS wave in lead II. Compare your results to the electrocardiogram of Fig. 7.23.

Problem 18. At a particular instant of the cardiac cycle, **p** is located at the midpoint of a line connecting two electrodes that are 50 cm apart, and **p** is parallel to that line. At that instant the magnitude of the potential difference between the electrodes is 1.5 mV. Upon depolarization, the potential change within the cells has magnitude 90 mV.

(a) What is the magnitude of **p**?

(b) If $\sigma_i/\sigma_o = 10$, what is the cross-sectional area of the advancing region of depolarization?

Problem 19. A semi-infinite slab of myocardium occupies the region $z > 0$. A hemispherical wave of depolarization moves radially away from the origin through the slab. At some instant of time the radius of the hemispherical depolarizing wavefront is R. Assume that $\mathbf{p} = \int d\mathbf{p}$, that $d\mathbf{p}$ is everywhere perpendicular to the advancing wavefront, and that the magnitude of $d\mathbf{p}$ is proportional to the local area of the wavefront. Find $\mathbf{p}$. Assume that the observation point is very far away compared to R.

Problem 20. Make measurements on yourself and construct Fig. 7.19.

Problem 21. Experiments have been done in which a dog heart was stimulated by an electrode deep within the myocardium. No exterior potential difference was detected until the spherical wave of depolarization grew large enough so that part of it intercepted one wall of the heart. Why?

Problem 22. Prove directly from Eq. 7.32 that $\mathrm{I} - \mathrm{II} + \mathrm{III} = 0$. (It is sometimes said that the equilateral nature of Einthoven's triangle is necessary to prove this.)

Problem 23. Derive Eqs. 7.32.

Section 7.8

Problem 24. Estimate the lower limit for the duration of the QRS complex by calculating the time required for a wave front to propagate across the heart wall. Assume the wall thickness is 10 mm and the propagation speed is 0.2 m s^{-1}.

Problem 25. In an ECG recording, the width of one large square corresponds to 200 ms. A normal heart rate is between 60 and 100 beats min^{-1}. The heart rate is usually measured by counting the number of large squares between adjacent QRS complexes.

(a) How many large squares are there for a normal heart rate?

(b) In Fig. 7.30 determine the rate of the atria and of the ventricles.

Problem 26. Consider Lead II of the normal ECG in Fig. 7.23. The QRS wave and the T wave are both positive. Use a 1-dimensional model to convince yourself that the QRS complex and the T wave should have opposite polarities. Why then is the T wave inverted? Find a way to explain the inverted T wave by letting the action potential duration vary between epicardium (outside) and endocardium (inside). On which surface should the duration be longest?

Section 7.9

Problem 27. Ohm's law says that $\mathbf{j} = \sigma\mathbf{E}$. Draw what $\mathbf{j}$ and $\mathbf{E}$ look like (a) in a circuit consisting of a battery and a resistor; (b) for the current flowing when a nerve cell depolarizes.

Problem 28. Obtain the values for β for a cube of length a on a side, for a cylinder of radius a and length h, and for a sphere of radius a.

Problem 29. Show that Eq. 7.36a is the same as Eq. 6.51 by considering the interior of a single cell stretched along the x axis as in Fig. 6.28. Consider the charge in a small cylindrical region of axoplasm of length h and radius a, the cylindrical surface of which is surrounded by cell membrane. Show that the total charge Q within the axoplasm changes according to

$$\begin{aligned}\frac{\partial Q}{\partial t} &= \pi a^2 h \frac{\partial \rho_i}{\partial t} = C\frac{\partial v_m}{\partial t} + i_m \\ &= 2\pi a h\left(c_m \frac{\partial v_m}{\partial t} + j_m\right),\end{aligned}$$

and that this can be combined with Eq. 7.36a to give

$$\begin{aligned}c_m \frac{\partial v_m}{\partial t} + j_m &= \frac{\pi a^2 h}{2\pi a h}\sigma_i \frac{\partial^2 v_i}{\partial x^2} \\ &= \frac{\sigma_i a}{2}\frac{\partial^2 v_i}{\partial x^2},\end{aligned}$$

which is the same as Eq. 6.51, except that it is written in terms of σ_i, a, and h instead of a and r_i.

Problem 30. Clark and Plonsey (1968) solved Eq. 7.34 for a cylindrical axon of radius a using the following method. Assume that the potentials all vary in the z direction sinusoidally, for instance $v_m(z) = V\sin(kz)$, where V is a constant.

(a) Show that the intracellular and extracellular potentials can be written as

$$\begin{aligned}v_i &= A I_0(kr)\sin(kz) \\ v_o &= B K_0(kr)\sin(kz),\end{aligned} \tag{7.53}$$

where I_n and K_n are *modified Bessel functions* obeying the equation

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right) - \left(k^2 + \frac{n^2}{r^2}\right)v = 0.$$

(b) Determine the constants A and B in terms of V, using the following two boundary conditions: $v_m = v_i - v_o$, and $\sigma_i(\partial v_i/\partial r) = \sigma_o(\partial v_o/\partial r)$, both evaluated at $r = a$. You will need to use the Bessel function identities $dI_0(kr)/dr = kI_1(kr)$ and $dK_0(kr)/dr = -kK_1(kr)$. Clark and Plonsey used this result and Fourier analysis (Chap. 11) to determine v_i and v_o when they are not sinusoidal in z.

Problem 31. Starting with the bidomain equations, divide Eq. 7.44a by σ_{ix} and Eq. 7.44b by σ_{ox}. Now subtract one equation from the other. Under what conditions do the equations contain $v_m = v_i - v_o$ but not v_i and v_o individually?

Section 7.10

Problem 32. Verify Eq. 7.47.

Problem 33. Verify the values given for rheobase and chronaxie in Table 7.1 that are based on Table 6.1.

Problem 34. An approximation to the error function is given by Abramowitz and Stegun (1972)

$$\operatorname{erf}(x) \approx 1 - \Big(1 + 0.278393x + 0.230389x^2 + 0.000972x^3 + 0.078108x^4\Big)^{-4}, \quad x > 0.$$

Calculate $\operatorname{erf}(x)$ using this approximation for $x = 0, 0.5, 1.0, 2.0$ and ∞. Using trial and error, determine the value of x for which $\operatorname{erf}(x) = 0.5$. (See Eq. 7.50.)

Problem 35. Find the equivalent of Eq. 7.45 in terms of the charge required for the stimulation.

Problem 36. If the medium has a constant resistance, find the energy required for stimulation as a function of pulse duration.

Problem 37. A typical pacemaker electrode has a surface area of 10 mm^2. What is its resistance into an infinite medium if it is modeled as a sphere? If it is modeled as a disk? (You will have to use results from Chap. 6 and assign a value for σ_o.)

Problem 38. Equation 6.51 is the cable equation for a nerve axon. Assume that the axon membrane is passive ($j_m = g_m(v_i - v_o)$, where g_m is a constant).

(a) Express the equation in terms of v_m and v_o instead of v_i and v_o, where $v_m = v_i - v_o$.

(b) Divide the resulting equation by g_m, and then write the cable equation in terms of the time constant c_m/g_m and the length constant $1/\sqrt{2\pi a r_i g_m}$.

(c) Put all the terms containing v_m on the left side, and terms containing v_o on the right side. The resulting equation should look like Eq. 6.55, except for a new source term on the right side equal to $-\lambda^2 \partial^2 v_o/\partial x^2$. (Measure v_m with respect to resting potential so $v_r = 0$ in Eq. 6.55). The negative of this new term has been called the *activating function* (Rattay 1987). It is useful when studying electrical stimulation of nerves.

Problem 39. For this problem, use the activating function derived in Problem 38. Assume that λ and τ are negligibly small, so that v_m simply equals the activating function. Consider a point electrode in an infinite, homogeneous volume conductor at distance d from the axon. The extracellular potential is $v_o = (1/4\pi\sigma_o)\, I/r$.

(a) Calculate v_m as a function of position x along the axon ($x = 0$ is the closest position to the electrode).

(b) Assume that the axon will fire an action potential if v_m somewhere along the axon is greater than $V_{\text{threshold}}$. Calculate the ratio of the stimulation current I needed to excite the axon for a cathode (negative electrode) and an anode (positive electrode).

Problem 40. For this problem, use the activating function derived in Problem 38. An action potential can be excited if a stimulus depolarizes an axon to a value greater than $V_{\text{threshold}}$, and a propagating action potential can be blocked if a stimulus hyperpolarizes to a value of v_m less than $-V_{\text{block}}$ ($V_{\text{block}} > V_{\text{threshold}}$).

(a) For a cathodal electrode [$v_o = (1/4\pi\sigma_o)\, I/r$] calculate the ratio of the threshold current to the current needed to block propagation.

(b) Use two electrodes (one cathodal and one anodal) to design a stimulator that will result in one-way propagation along the axon (say, propagation only in the positive x direction, but blocked in the negative x direction). For an application of such electrodes during functional electrical stimulation, see Ungar et al. (1986).

Problem 41. For this problem, use the activating function derived in Problem 38, and block by hyperpolarization derived in Problem 40. The factor of λ^2 in the activating function implies that larger diameter axons are easier to stimulate than smaller diameter axons. Sometimes you want to excite the smaller fibers without the larger fibers (*physiological recruitment*). Describe qualitatively how you can use a single electrode and block in the hyperpolarized region to obtain physiological recruitment. For a more complete discussion, see Tai and Jiang (1994).

Problem 42. In *second degree heart block*, the wave front sometimes passes through the conduction system and sometimes does not. Qualitatively sketch the ECG for a heart with second degree block for at least five beats. Specifically include the case where every third wave is blocked. Include the P wave, the QRS wave, and the T wave.

Problem 43. During *sinus exit block* the SA node functions normally but the wave front fails to propagate from the SA node to the atria. Sketch five beats of an ECG with all beats normal except the third, which undergoes sinus exit block.

Problem 44. In *sick sinus syndrome* the SA node has a slow and erratic rate. The AV node and conduction system function properly. You plan to implant a pacemaker in the patient. Should it stimulate the atria or the ventricles? Why?

Problem 45. A patient with *intermittent heart block* has an AV node that functions normally most of the time with occasional episodes of block, lasting perhaps several hours. Design a pacemaker to treat the patient. Ideally, your design will not stimulate the heart when it is functioning normally. Describe

(a) whether you will stimulate the atria or ventricles

(b) which chambers you will monitor with a recording electrode

(c) what logic your pacemaker will use to determine when to stimulate. Your design may be similar to a *demand pacemaker* described in Jeffrey (2001, p. 132).

Problem 46. The Lapicque strength-duration (SD) curve is

$$\frac{i}{i_R} = 1 + \frac{t_C}{t},$$

the SD curve in terms of the error function is

$$\frac{i}{i_R} = \frac{1}{\text{erf}\left(\sqrt{0.228 t/t_C}\right)},$$

and the SD curve derived in Chap. 6 Problem 37 is

$$\frac{i}{i_R} = \frac{1}{1 - e^{-0.693 t/t_C}}.$$

(a) Plot all three curves for $0 < t/t_C < 5$. Use the equation in Problem 34 to evaluate the error function.
(b) Find approximations for each curve for $t/t_C \ll 1$. You may need the Taylor's series expansions $e^x \approx 1 + x$ and $\text{erf}(x) \approx 2x/\sqrt{\pi}$.
(c) Discuss the physical assumptions that were used to derive each curve.

Problem 47. Consider a pacemaker delivering a $2-$ mA, $1-$ V, $1-$ ms pulse every second. Pacemakers are often powered by a lithium-iodide battery that can deliver a total charge of 2 ampere hours.
(a) What is the energy per pulse?
(b) What is the average power?
(c) How long will the battery last?
(d) Your answer to (c) is an overestimate of battery lifetime, in part because the battery voltage begins to decline before all its charge has been delivered, and in part because the pacemaker circuitry requires a small, constant current. For this pacemaker, add a constant current drain of 5 μA and assume that the useful lifetime of the battery is over when 75 % of the total charge has been delivered. How long will the battery last in this case?

Problem 48. During stimulation of cardiac tissue through a small anode, the tissue under the electrode and in the direction perpendicular to the myocardial fibers is hyperpolarized, and adjacent tissue on each side of the anode parallel to the fiber direction is depolarized. Imagine that just before this stimulus pulse is turned on the tissue is refractory. The hyperpolarization during the stimulus causes the tissue to become excitable. Following the end of the stimulus pulse, the depolarization along the fiber direction interacts electrotonically with the excitable tissue, initiating an action potential *(break excitation)*. (This type of break excitation is very different than the break excitation analyzed on page 181.)
(a) Sketch pictures of the transmembrane potential distribution during the stimulus. Be sure to indicate the fiber direction, the location of the anode, the regions that are depolarized and hyperpolarized by the stimulus, and the direction of propagation of the resulting action potential.
(b) Repeat the analysis for break excitation caused by a cathode instead of an anode. For a hint, see Wikswo and Roth (2009).

Problem 49. The signal measured during optical mapping, V, is a weighted average of the transmembrane potential, $V_m(z)$, as a function of depth,

$$V = \int_0^\infty V_m(z) w(z) dz,$$

where $w(z)$ is a normalized weighting function. Suppose the incident light that produces the fluorescence decays with depth exponentially, with an optical length constant δ. Then $w(z) = \exp(-z/\delta)/\delta$. Often a shock will cause $V_m(z)$ to fall off exponentially with depth, $V_m(z) = V_0 \exp(-z/\lambda)$, where V_0 is the transmembrane potential at the tissue surface and λ is the electrical length constant (see Sect. 6.6.12).
(a) Perform the required integration to find an analytical expression for the optical signal, V, as a function of V_0, δ and λ.
(b) What is V in the case $\delta \ll \lambda$? Explain this result physically.
(c) What is V in the case $\delta \gg \lambda$? Explain this result physically.
(d) For which limit do you obtain an accurate measurement of the transmembrane potential at the surface, $V = V_0$?
For additional analysis, see Janks and Roth (2002).

Problem 50. Consider a two-dimensional sheet of cardiac tissue represented as a bidomain having unequal anisotropy ratios: $\sigma_{ix} = \sigma_{ex} = 0.2$, $\sigma_{iy} = 0.02$, and $\sigma_{ey} = 0.08$ S m^{-1}. Assume an insulated obstacle that current must go around is at the center of the sheet. At any point in the tissue, current will divide between the intracellular and extracellular spaces according to their conductivities, with a larger fraction of the current in the space with greater conductivity.
(a) If current is passed through the tissue in the x-direction, determine qualitatively where the tissue is depolarized and where it is hyperpolarized in the region surrounding the insulator. Recall, depolarization occurs where current passes from the intracellular into the extracellular space, and hyperpolarization where current passes from the extracellular into the intracellular space.
(b) Repeat this analysis if current is passed in the y-direction.
(c) What would be the transmembrane potential if the tissue had equal anisotropy ratios?
For additional analysis, see Langrill and Roth (2001).

Section 7.11

Problem 51. When measuring the EEG with electrodes distributed according to the 10–20 system, you obtain measurements of the potential difference between the ith electrode

($i = 1, \ldots, 20$) and the reference electrode ($i = 21$). Show that by computing the average reference $v_i^* = (v_i - v_{21}) - (1/20)\sum_{j=1}^{20}\left(v_j - v_{21}\right)$, the resulting values of v_i^* are independent of the reference potential v_{21}.

Problem 52. Consider a very simple model of the EEG: a dipole **p** pointing in the z direction at the center of a spherical conductor of radius R and conductivity σ_o. The potential v_o can be written as the sum of two terms: the potential of a dipole in an unbounded medium plus a potential that obeys Laplace's equation

$$v_o = \frac{p\cos\theta}{4\pi\sigma_o r^2} + Ar\cos\theta$$

where r and θ are in spherical coordinates, and A is an unknown constant.

(a) Use Appendix L to show that the second term in the expression for v_o obeys Laplace's equation.

(b) If the region outside the spherical conductor is air (an insulator), determine the value of A by using the boundary condition that the radial current at the surface of the sphere is zero.

(c) Calculate v_o as measured at the sphere surface ($r = R$), and determine by what factor v_o differs from what it would be in the case of an unbounded volume conductor.

Problem 53. Suppose you measure the EEG potential v_j at N locations $\mathbf{r}_j = (x_j, y_j, z_j)$, $j = 1, \cdots, N$. Assume v_j is produced by a dipole $\mathbf{p} = (p_x, p_y, p_z)$ located at the origin. Define

$$R = \sum_{j=1}^{N}\left[\frac{p_x x_j + p_y y_j + p_z z_j}{4\pi\sigma\left(x_j^2 + y_j^2 + z_j^2\right)^{3/2}} - v_j\right]^2,$$

which measures the least-squares difference between the data and the potential predicted by a single-dipole model. (Chap. 11 explores the least-squares method in greater detail.) The goal is to find the dipole components p_x, p_y, p_z that fit the data best (minimize R).

(a) Minimize R with respect to p_x (set $dR/dp_x = 0$) and find an equation relating p_x, p_y, and p_z.

(b) Repeat for p_y and p_z.

(c) Write the three equations in the form $\mathbf{Ap} = \mathbf{b}$, where **A** is a 3×3 matrix and **b** is a 3×1 vector. Find expressions for the components of **A** and **b**.

(d) If we had not assumed that we knew the location of the dipole, the problem would be much more difficult. Assume the dipole is at location $\mathbf{r}_p = (x_p, y_p, z_p)$. Modify R and then try to minimize it with respect to $\mathbf{r}_p$. Carry the calculation far enough to convince yourself that you must now solve nonlinear equations to determine $\mathbf{r}_p$. Press et al. (1992) discuss methods for making nonlinear least squares fits.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions with formulas, graphs and mathematical tables. U.S. Government Printing Office, Washington

Barold SS (1985) Modern cardiac pacing. Futura, Mount Kisco

Clark J, Plonsey R (1968) The extracellular potential field of a single active nerve fiber in a volume conductor. Biophys J 8:842–864

Delmar M, Sorgen PL (2009) Molecular organization and regulation of the cardiac gap junction channel connexin43. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 85–92

Dosdall DJ, Fast VG, Ideker RE (2009) Mechanisms of defibrillation. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 499–508

Geddes LA, Bourland JD (1985) The strength-duration curve. IEEE Trans Biomed Eng 32:458–459

George AL Jr (2009) Inheritable sodium channel diseases. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 29–42

Gray RA (2009) Rotors and spiral waves in the heart. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 349–359

Gulrajani RM (1998) Bioelectricity and biomagnetism. Wiley, New York

Harris JW, Stocker H (1998) Handbook of mathematics and computational science. Springer, New York

Henriquez CS (1993) Simulating the electrical behavior of cardiac tissue using the bidomain model. Crit Rev Biomed Eng 21(1):1–77

Henriquez CS, Ying W (2009) The bidomain model of cardiac tissue: from microscale to macroscale. In Efimov IR, Kroll MW, Tchou PJ (eds) Cardiac bioelectric therapy: mechanisms and practical implications. Springer, New York, pp 401–421

Hobbie RK The electrocardiogram as an example of electrostatics. Am J Phys 41:824–831

Janks DL, Roth BJ (2002) Averaging over depth during optical mapping of unipolar stimulation. IEEE Trans Biomed Eng 49:1051–1054

Janks DL, Roth BJ (2009) The bidomain theory of pacing. In Efimov IR, Kroll MW, Tchou PJ (eds) Cardiac bioelectric therapy: mechanisms and practical implications. Springer, New York, pp 63–83

Jeffrey K (2001) Machines in our hearts. Johns Hopkins University Press, Baltimore

Langrill DM, Roth BJ (2001) The effect of plunge electrodes during electrical stimulation of cardiac tissue. IEEE Trans Biomed Eng 48:1207–1211

Lapicque L (1909) Definition experimentale de l'excitabilite. Comptes Rendus Acad Sci 67(2):280–283

Lindemans FW, Denier van der Gong JJ (1978) Current thresholds and liminal size in excitation of heart muscle. Cardiovasc Res 12:477–485

Lopes da Silva FH (2008) The impact of EEG/EMG signal processing and modeling in diagnosis and management of epilepsy. IEEE Rev Biomed Eng 1:143–156

Luo CH, Rudy Y (1994) A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes. Circ Res 74(6):1071–1096

Malmivuo J, Plonsey R (1995) Bioelectromagnetism. Oxford University Press, Oxford

Miller CE, Henriquez CS (1990) Finite element analysis of bioelectric phenomena. Crit Rev Biomed Eng 18(3):207–233

Mitrani RD, Klein LS, Rardon DP, Zipes DP, Miles WM (1995) Current trends in the implantable cardioverter–defibrillator. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd ed. Saunders, Philadelphia, pp 1393–1403

Moses HW, Mullin JC (2007) A practical guide to cardiac pacing, 6th ed. Lippincott Williams and Wilkins, Philadelphia

Nunez PL, Srinivasan R (2005) Electric fields of the brain, 2nd ed. Oxford University Press, Oxford

Oudit GY, Backx PH (2004) Voltage-regulated potassium channels. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 29–42

Peckham PH, Knutson JS (2005) Functional electrical stimulation for neuromuscular applications. Phys Med Biol 47:R67–R84

Perlmutter JS, Mink JW (2006) Deep brain stimulation. Annu Rev Neurosci 29:229–257

Plonsey R (1969) Bioelectric Phenomena. McGraw-Hill, New York

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C: the art of scientific computing, 2nd ed., reprinted with corrections, 1995. Cambridge University Press, New York

Rardon DP, Miles WM, Zipes DP (2000) Atrioventricular block and dissociation. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 3rd ed. Saunders, Philadelphia, pp 451–459

Rattay F (1987) Ways to approximate current-distance relations for electrically stimulated fibers. J Theor Biol 125:339–349

Ripplinger CM, Efimov IR (2009) The virtual electrode hypothesis of defibrillation. In Efimov IR, Kroll MW, Tchou PJ (eds) Cardiac bioelectric therapy: mechanisms and practical implications. Springer, New York, pp 331–356

Rosenbaum DS, Jalife J (2001) Optical mapping of cardiac excitation and arrhythmias. Futura, Armonk

Roth BJ (1992) How the anisotropy of the intracellular and extracellular conductivities influences stimulation of cardiac muscle. J Math Biol 30:633–646

Roth BJ (1997) Electrical conductivity values used with the bidomain model of cardiac tissue. IEEE Trans Biomed Eng 44:326–328

Rudy Y, Burnes JE (1999) Noninvasive electrocardiographic imaging. Ann Noninvasive Electrocardiol 4(3):340–359

Stanley PC, Pilkington TC, Morrow MN (1986) The effects of thoracic inhomogeneities on the relationship between epicardial and torso potentials. IEEE Trans Biomed Eng 33(3):273–284

Tai C, Jiang D (1994) Selective stimulation of smaller fibers in a compound nerve trunk with single cathode by rectangular current pulses. IEEE Trans Biomed Eng 41:286–291

Trayanova N, Plank G (2009) The bidomain model of defibrillation. In Efimov IR, Kroll MW, Tchou PJ (eds) Cardiac bioelectric therapy: mechanisms and practical implications. Springer, New York, pp 85–110

Trayanova N, Henriquez CS, Plonsey R (1990) Limitations of approximate solutions for computing extracellular potential of single fibers and bundle equivalents. IEEE Trans Biomed Eng 37(1):22–35

Ungar IJ, Mortimer JT, Sweeney JD (1986) Generation of unidirectionally propagating action potentials using a monopolar electrode cuff. Ann Biomed Eng 14:437–450

Voorhees CR, Voorhees WD III, Geddes LA, Bourland JD, Hinds M (1992) The chronaxie for myocardium and motor nerve in the dog with chest-surface electrodes. IEEE Trans Biomed Eng 39(6):624–628

Watanabe A, Grundfest H (1961) Impulse propagation at the septal and commissural junctions of crayfish lateral giant axons. J Gen Physiol 45:267–308

Wikswo JP, Roth BJ (2009) Virtual electrode theory of pacing. In Efimov IR, Kroll MW, Tchou PJ (eds) Cardiac bioelectric therapy: mechanisms and practical implications. Springer, New York, pp 283–330

Zeng F-G, Rebscher S, Harrison W, Sun X, Feng H (2008) Cochlear implants: system design, integration and evaluation. IEEE Rev Biomed Eng 1:115–142

Biomagnetism

8

The field of *biomagnetism* has exploded in recent decades. Magnetic signals have been detected from the heart, brain, skeletal muscles, and isolated nerve and muscle preparations. Measurements of the magnetic susceptibility of the lung show the effect of dust inhalation. Susceptibility measurements of the heart can determine blood volume, while the susceptibility of the liver can measure iron stores in the body. Bacteria and some animals contain aggregates of magnetic particles, often attached to neural tissue. Bacteria use these magnetic particles to determine which way is down. Magnetism is used for orientation by birds and other animals.

Sections 8.1 and 8.2 review the basics of magnetism. Section 8.3 calculates the magnetic field of an axon in an infinite conducting medium. This result, which shows that the field is due primarily to the current dipole in the interior of the axon, is approximately true for the magnetocardiogram and evoked responses from the brain, described in Sects. 8.4 and 8.5.

Section 8.6 reviews electromagnetic induction. Section 8.7 describes the use of varying magnetic fields to stimulate nerves or muscles. Section 8.8 introduces diamagnetic, paramagnetic, and ferromagnetic materials and describes biomagnetic effects that depend on magnetic materials. Section 8.9 reviews instrumentation for measuring these weak magnetic signals.

8.1 The Magnetic Force on a Moving Charge

Lodestone, compass needles, and other forms of magnetism have been known for centuries, but it was not until 1820 that Hans Christen Oersted showed that an electric current could deflect a compass needle. We now know that magnetism results from electric forces that moving charges exert on other moving charges and that the appearance of the magnetic force is a consequence of special relativity. An excellent development of magnetism from this perspective is found in Purcell and Morin (2013). The development here is more traditional (Griffiths 2013) and is incomplete.

Suppose that a beam of electrons is accelerated in a cathode-ray tube (as in an oscilloscope, computer display, or television receiver) and causes a spot of light to be emitted where it strikes a fluorescent screen. The electron source is cathode C in Fig. 8.1. The accelerating electrode is E. The fact that the beam is accelerated toward a positively charged electrode confirms that the electrons are negatively charged. The beam normally strikes the screen at point X. Placing a battery between plates A and B creates an electric field that deflects the beam as it passes between the plates. If plate A is positively charged, the beam is deflected upward to point Y. If the battery is removed and the north pole of a bar magnet is brought to the position shown, the beam is deflected to point Z.

We say that a *magnetic field* exists in the space surrounding the bar magnet and that the direction of the magnetic field at any point is the direction a small compass needle located there would point. Experiments show that the force is at right angles to both the direction of the magnetic field and the velocity of the charged particle, and that the magnitude of the force $\mathbf{F}$ is proportional to the charge, the magnitude of the velocity $\mathbf{v}$, and the strength of magnetic field $\mathbf{B}$. (In fact, modern definitions of the magnetic field are based on this proportionality.) The magnitude of the force is greatest when $\mathbf{v}$ and $\mathbf{B}$ are perpendicular. We have seen a relationship like this between three vectors before: the vector product or cross product, which was associated with torque and defined in Sect. 1.5. We write

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B}). \tag{8.1}$$

The SI unit of $\mathbf{B}$ is the tesla, T. An earlier name was the weber per square meter. Another unit is the gauss, G: $1\ \mathrm{T} = 10^4\ \mathrm{G}$.

8.1.1 The Lorentz Force

If a coordinate system is set up so that $\mathbf{v}$ is along the x axis and $\mathbf{B}$ is along the y axis, then $\mathbf{v} \times \mathbf{B}$ and the force

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_8, © Springer International Publishing Switzerland 2015

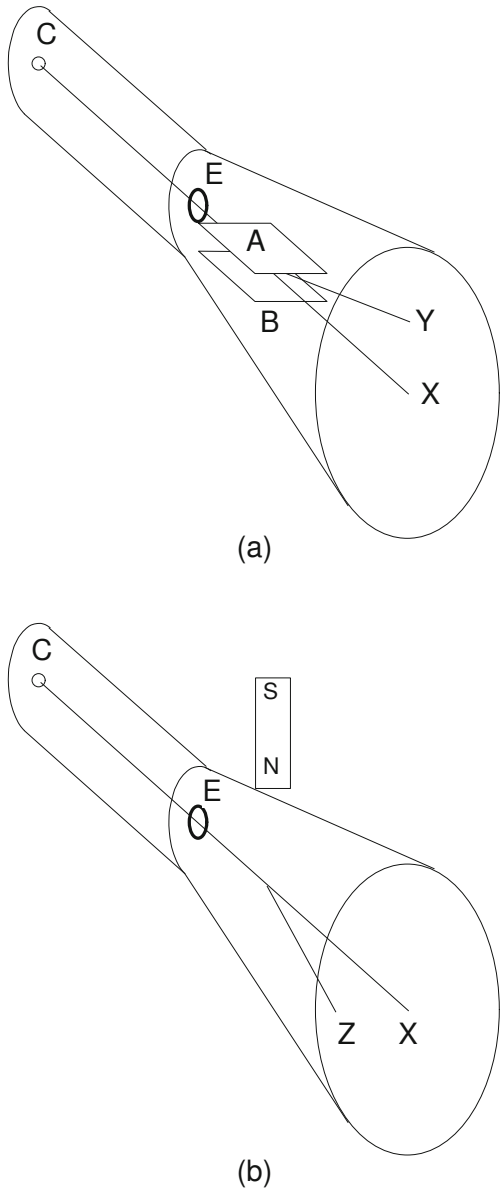


Fig. 8.1 An electron beam generated at cathode C and accelerated through electrode E strikes the fluorescent screen on the *right*. **a** A positive charge on plate A and negative charge on plate B deflects the beam from X to Y. **b** A bar magnet brought close as shown deflects the beam to point Z

on a positive charge are along the $+z$ axis. For negatively charged electrons **F** is in the opposite direction. Combining Eq ref8.01 with the electric force gives the full expression for the electromagnetic force, often called the *Lorentz force*:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{8.2}$$

Since current in a wire is the result of moving charges, there is a force on a segment of wire carrying a current. Suppose that there are C particles per unit volume, each with charge q, drifting with speed v along a segment of wire of length ds and cross sectional area S. In time dt the total

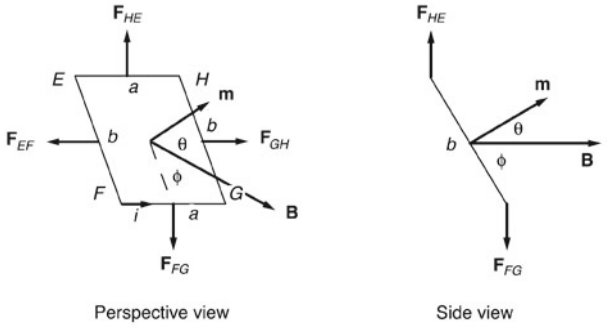


Fig. 8.2 A current-carrying loop is in a uniform magnetic field. The *dashed line* from the *center* of the loop to the *center* of edge FG, vector **B** and vector **m** all lie in the same plane. The sum of angles θ and ϕ is $\pi/2$. The forces on opposite sides add to zero. There is a torque on the loop unless its plane is perpendicular to the field ($\phi = \pi/2$). The magnetic moment **m** is perpendicular to the plane of the loop

charge passing a given plane is $CvqS\,dt$ (see Eq. 4.11) so that the current is $i = CvqS$. If there is a magnetic field perpendicular to the wire, the magnitude of the force on each particle is qvB and the total force is $CS\,ds\,qvB = iB\,ds$. If vector $d\mathbf{s}$ is defined along the wire in the direction of the positive current, then the contribution to the magnetic force from this segment of the wire is

$$d\mathbf{F} = i(d\mathbf{s} \times \mathbf{B}). \tag{8.3}$$

If a small rectangular loop of wire is placed in a uniform magnetic field and a current is made to flow in the wire, there is a magnetic force on each arm of the loop. (The current can be led to and from the rectangle by two parallel closely spaced wires, in which the forces cancel because the currents are in opposite directions. Forces not considered here maintain the position of the loop.) Figure 8.2 shows the orientation of the loop in the horizontal magnetic field. The magnetic moment **m** is perpendicular to the loop and makes an angle θ with the direction of **B**. Sides HE and FG are of length a and perpendicular to the field. The other two sides have length b. The force on side EF has magnitude $iBb\sin\phi$ and is directed as shown. Side GH has a force of equal magnitude in the opposite direction. On side FG the force is down and on side HE it is up, both with magnitude iBa. The vector sum of all the forces is zero. There is a torque, however. If the torque is taken about the center of the loop, the FG force and HE force each exert a torque of magnitude $(iBa)(b/2)\cos\phi$. The total torque is therefore $iBab\cos\phi$. The loop is said to have a magnetic moment **m** of magnitude iS, where $S = ab$ is the area of the loop. Vector **m** is defined to point perpendicular to the loop in the direction of the thumb of the right hand when the fingers curl in the direction of the current around the loop. The units of **m** are A m^2 or J T^{-1}. In terms of angle θ between **m** and **B**, the torque $\boldsymbol{\tau}$ exerted by the magnetic field on the magnetic moment has

magnitude $iabB\sin\theta$, so

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}. \tag{8.4}$$

The torque is zero when **m** and **B** are parallel or antiparallel. When they are parallel the equilibrium is stable: if there is a small rotation of **m** the torque acts to return it to equilibrium. When they are antiparallel, the equilibrium is unstable.

A small current loop can be used to test for the presence of a magnetic field. At equilibrium **m** points in the same direction that a small compass needle would point and gives the direction of **B**. Measuring the torque for a known displacement of **m** from this direction gives the magnitude of **B**.

8.1.2 The Cyclotron

One important application of magnetic forces in medicine is the *cyclotron*. Many hospitals have a cyclotron for the production of radiopharmaceuticals, especially for the generating positron-emitting nuclei for use in *Positron Emission Tomography (PET)* imaging (see Chap. 17).

Consider a particle of charge q and mass m, moving with speed v in a direction perpendicular to a magnetic field **B**. The magnetic force will bend the path of the particle into a circle. Newton's second law states that the mass times the centripetal acceleration, v^2/r, is equal to the magnetic force

$$mv^2/r = qvB. \tag{8.5}$$

The speed is equal to circumference of the circle, $2\pi r$, divided by the period of the orbit, T. Substituting this expression for v into Eq. (8.5) and simplifying, we find

$$T = 2\pi m/(qB). \tag{8.6}$$

In a cyclotron particles orbit at the *cyclotron frequency*, $f = 1/T$. Because the magnetic force is perpendicular to the motion, it does not increase the particles' speed or energy. To do that, the particles are subjected periodically to an electric field that changes direction with the cyclotron frequency so that it is always accelerating, not decelerating the particles. This would be difficult if not for the fortuitous disappearance of both v and r from Eq. (8.6), so that the cyclotron frequency only depends on the charge-to-mass ratio of the particles and the magnetic field, but not on their energy.

Typically, protons are accelerated in a magnetic field of about 1 T, resulting in a cyclotron frequency of approximately 15 MHz. Each orbit raises the potential of the proton by about 100 kV. It must circulate enough times to raise its total energy to at least 10 MeV so that it can overcome the electrostatic repulsion of the target nucleus and cause nuclear reactions. For example, the high-energy protons may be incident on a target of ^{18}O (a rare but stable isotope of oxygen), initiating a nuclear reaction that results in the production of ^{18}F, an important positron emitter used in PET studies.

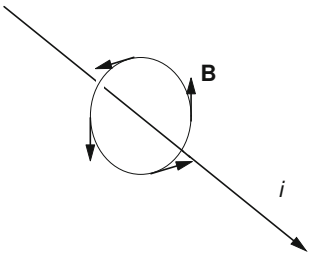


Fig. 8.3 The magnetic field around a current-carrying wire is at right angles to the wire and the perpendicular from the observation point to the wire. The magnitude is inversely proportional to the distance from the wire

8.2 The Magnetic Field of a Moving Charge or a Current

8.2.1 The Divergence of the Magnetic Field is Zero

With a compass needle or small sensing coil we can in principle map the magnetic field surrounding a bar magnet or a wire carrying a current. If we examine the field near a long straight wire carrying current i, we find that **B** is always at right angles to the wire and at distance r has magnitude

$$B = \frac{\mu_0 i}{2\pi r}. \tag{8.7}$$

The constant μ_0 is analogous to ϵ_0 in electrostatics and is $4\pi \times 10^{-7}$ T m A^{-1} (or Ω s m^{-1}). Figure 8.3 shows the direction of **B** at various locations around a wire. The direction of the force is consistent with Eq. 8.2 if the direction of **B** is defined to be the direction in which the fingers of the right hand curl when the thumb points along the wire in the direction of the (positive) current.

Close to the wire **B** is always at right angles to the wire, in contrast to the electric field, which close to a charge always points toward or away from it. In the electric case, the flux of **E** through a closed surface is proportional to the charge within the volume enclosed by the surface (Gauss's law, Sect. 8.3). In contrast, the flux of **B** through a closed surface is always zero. In the notation of Sect. 4.1,

$$\iint_{\text{closed surface}} B_n \, dS = \iint_{\text{closed surface}} \mathbf{B} \cdot d\mathbf{S} = 0. \tag{8.8}$$

If single magnetic charges (magnetic monopoles) existed, the flux would be proportional to the magnetic charge within the volume. Magnetic monopoles have never been observed, in spite of considerable effort to find them.

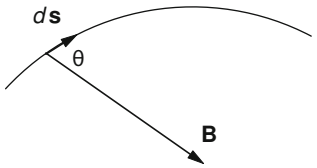


Fig. 8.4 The line integral of $\mathbf{B} \cdot d\mathbf{s}$ is calculated by multiplying $d\mathbf{s}$ by the component of $\mathbf{B}$ parallel to $d\mathbf{s}$, that is $B \cos\theta$

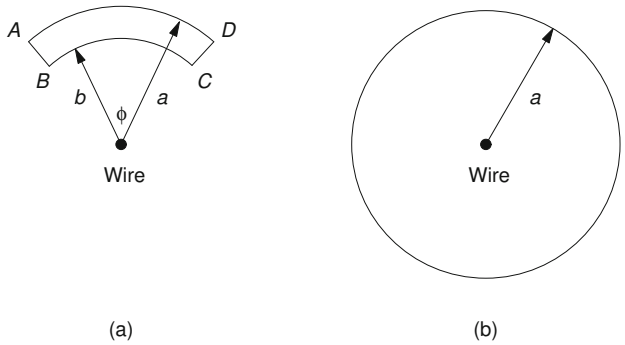


Fig. 8.5 Two paths of integration. In **a** the path does not encircle the wire carrying the current, and $\oint \mathbf{B} \cdot d\mathbf{s} = 0$. In **b** the path encircles the wire and $\oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 i$

As in the electric case, we can construct lines of $\mathbf{B}$. The tangent to the line always points in the direction of $\mathbf{B}$. For the long wire, the lines of $\mathbf{B}$ are circles. One can show from Eq. 8.8 that lines of $\mathbf{B}$ always close on themselves.

Equation 8.8 has the form of the continuity equation, Eq. 4.4, with $\mathbf{B}$ substituted for $\mathbf{j}$ and with $C = 0$. The differential version of Eq. 8.8 can therefore be obtained from Eq. 4.8. It is

$$\operatorname{div} \mathbf{B} = \nabla \cdot \mathbf{B} = 0. \tag{8.9}$$

8.2.2 Ampere's Circuital Law

It is also interesting to consider the line integral of $\mathbf{B}$ around a closed path. That is, for any element of the path $d\mathbf{s}$ shown in Fig. 8.4 take the projection of $\mathbf{B}$ in the direction of $d\mathbf{s}$, $B \cos\theta$. Sum up all the contributions $B \cos\theta \, ds$ along the entire closed path. For path $ABCD$ in Fig. 8.5a, the result is zero. The reason is that $B \cos\theta \, ds$ is zero on segments AB and CD. On segment DA it is $(\mu_0 i/2\pi a)(a\phi) = \mu_0 i\phi/2\pi$, while on segment BC it is $-(\mu_0 i/2\pi b)(b\phi) = -\mu_0 i\phi/2\pi$. In Fig. 8.5b the path is circular with the wire at the center, and the line integral is $B(2\pi a) = \mu_0 i$. This result is general:

$$\oint B \cos\theta \, ds = \oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 i. \tag{8.10}$$

The circle on the integral sign means that the integral is taken around a closed path. The line integral of the magnetic field around a closed path is equal to μ_0 times the current through a circuit enclosed by that path. If two wires carrying equal and opposite currents are enclosed by the path of the line integral, the integral is zero. It does *not* mean that $\mathbf{B}$ is zero everywhere on the path.

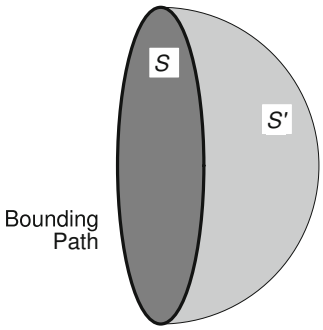


Fig. 8.6 Since the total current or flux of $\mathbf{j}$ through any closed surface is zero, the current through surface S is equal to the current through surface S'

A more general statement is that for steady currents the line integral of $\mathbf{B}$ around a closed path is equal to the integral of the current density $\mathbf{j}$ through any surface enclosed by the path:

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 \iint \mathbf{j} \cdot d\mathbf{S}. \tag{8.11}$$

This is known as *Ampere's circuital law*. Like Gauss's law, it is always true but not always useful. It is true for currents that do not vary with time, but it can be used to calculate the magnetic field only if symmetry can be used to argue that $\mathbf{B}$ is always parallel to the path and has the same magnitude at all points on the path.

The surface used to calculate the right-hand side can be any surface bounded by the path used on the left. Since we are dealing with steady currents for which there is no charge accumulation, the continuity equation, Eq. 4.4, shows that the flux of $\mathbf{j}$ (the total current) through any closed surface is zero. Two surfaces S and S', both bounded by the path, form a closed surface as shown in Fig. 8.6. The total current through surface S is the same as the total current through S'.

8.2.3 The Biot–Savart Law

In situations where the symmetry of the problem does not allow the field to be calculated from Ampere's law, it is possible to find the field due to a steady current in a closed circuit using the *Biot–Savart law*. The contribution $d\mathbf{B}$ to the magnetic field from current i flowing along a line element $d\mathbf{s}$ is

$$d\mathbf{B} = \frac{\mu_0 i}{4\pi} \frac{d\mathbf{s} \times \mathbf{r}}{r^3}. \tag{8.12}$$

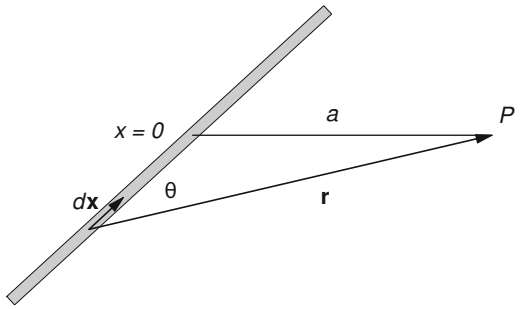


Fig. 8.7 The Biot–Savart law is used to calculate the magnetic field at point P due to an infinite wire

Vector $\mathbf{r}$ is from the current element to the point where the field is to be calculated. The field is found by integrating over the *entire circuit*.

Figure 8.7 shows how this integration is done for an infinitely long straight wire along the x axis. The contribution at point P is obtained by dropping a perpendicular from P to the wire to define $x = 0$. The distance from P to the wire is a. The contribution from an element dx at point x is

$$dB = \frac{\mu_0 i}{4\pi}\frac{dx\,\sin\theta}{r^2} = \frac{\mu_0 i}{4\pi}\frac{a\,dx}{r^3}.$$

Since $r^2 = a^2 + x^2$ the total field is

$$B = \frac{\mu_0 i}{4\pi}\int_{-\infty}^{\infty}\frac{a\,dx}{(a^2+x^2)^{3/2}} = \frac{\mu_0 i a}{4\pi}\left[\frac{x}{a^2(x^2+a^2)^{1/2}}\right]_{-\infty}^{\infty} = \frac{\mu_0 i}{2\pi a}.$$

This agrees with Eq. 8.7 and the result obtained using Ampere's circuital law.

A steady current from a point source which spreads uniformly in all directions generates no magnetic field. To see why consider Fig. 8.8. The source of current is at O. The magnetic field at P can be calculated using the Biot–Savart law. For any element $d\mathbf{s}$ a symmetric element $d\mathbf{s}'$ can be selected, such that $d\mathbf{s}\times\mathbf{r} = -d\mathbf{s}'\times\mathbf{r}'$. Associated with each element is a small area dA, and the current along $d\mathbf{s}$ is $i = j\,dA$. We can set $dA = dA'$ so i is the same in each case. Therefore, $B = 0$. (This can also be shown using Ampere's law; see Problem 11.)

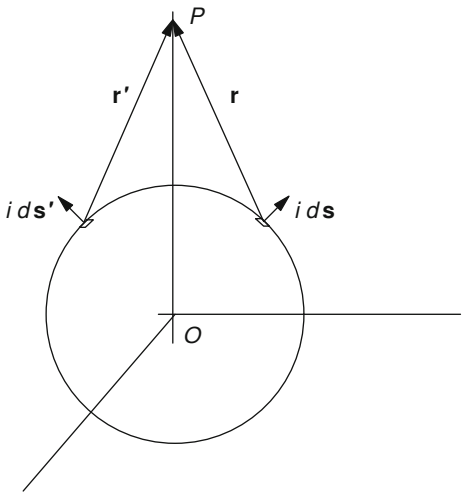


Fig. 8.8 The magnetic field from a spherically symmetric radial distribution of current is zero. The source at O sends current uniformly in all directions. P is the observation point. For any element $d\mathbf{s}$ there is a corresponding $d\mathbf{s}'$ such that $d\mathbf{s}\times\mathbf{r} = -d\mathbf{s}'\times\mathbf{r}'$. The current through a small area dA around $d\mathbf{s}$ is i. The same current flows through a corresponding area around $d\mathbf{s}'$. Can you obtain the same result by a symmetry argument?

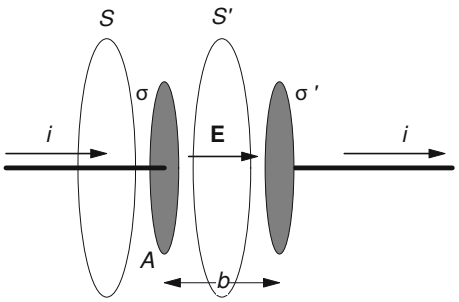


Fig. 8.9 A wire and capacitor plates. The integral of the current density through surface S, which is pierced by the wire, is i. Through surface S', which is between the capacitor plates, the integral is zero. If the displacement current density is included, both surface integrals are the same. (If surfaces S and S' are not large enough, there is also a net displacement current through S, as can be seen from Fig. 8.10)

8.2.4 The Displacement Current

Derivation of Ampere's law requires that there be no charge buildup, so that the total current through a closed surface is zero. However, we will consider an action potential in which the membrane capacitance charges and discharges. To see how this affects Ampere's law, consider current i charging or discharging the two shaded capacitor plates in Fig. 8.9. The area of each capacitor plate is A. The region between the plates, of thickness b, is filled with dielectric of dielectric constant κ. The integral $\iint \mathbf{j}\cdot d\mathbf{S}$ is i for surface S and zero for surface S'. Because of the current, the charge density σ on the left-hand plate is increasing at a rate given by $i = Ad\sigma/dt$, while on the right-hand plate the charge is decreasing because $i = -Ad\sigma/dt$. Since the electric field between the plates is $E = \sigma/\kappa\epsilon_0$ we can say that $i = Ad(\kappa\epsilon_0 E)/dt$. The quantity $D = \kappa\epsilon_0 E$ is called the *electric displacement*, and

$$j_d = \frac{\partial D}{\partial t}$$

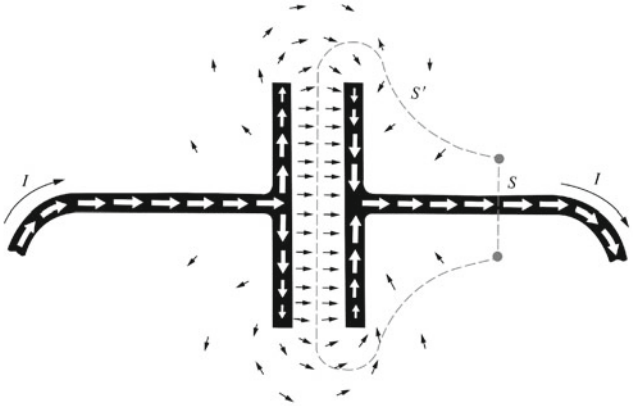


Fig. 8.10 The conduction current (white arrows) and displacement current (black arrows) in a discharging capacitor. The conduction current decreases with distance out the capacitor plates. The displacement current includes the fringing field. (From Purcell and Morin 2013. Used by permission)

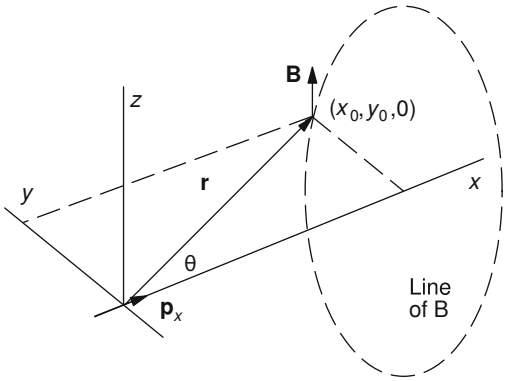


Fig. 8.11 The geometry for calculation of the magnetic field due to a current element $i\,dx$ or current dipole p_x stretched along the x axis

is called the *displacement current density*. More careful consideration shows that Ampere's law is valid when the charge on the plates is changing, if we replace $\mathbf{j}$ by $\mathbf{j} + \mathbf{j}_d$:

$$\oint \mathbf{B} \cdot d\mathbf{s} = \mu_0 \iint (\mathbf{j} + \mathbf{j}_d) \cdot d\mathbf{S}. \tag{8.13}$$

With this change, if S and S' are circles of radius a, Ampere's law gives $B = \mu_0 i/2\pi a$ for either one. (The radius of the circle must be very large; see the discussion in the next paragraph.)

What current should be used in the Biot–Savart law? A very surprising answer is that as long as the fields are relatively slowly varying (so that the emission of radio waves is not important), the displacement current contributes nothing. We are free to include it or ignore it. Purcell and Morin (2013, p. 435) and Shadowitz (1975, p. 416) discuss why this is so. It is not always easy to calculate the entire displacement current. For example, Fig. 8.10 shows how the conduction current and displacement current vary when current charges a capacitor. Notice that some of the displacement current flows to and from the back sides of the capacitor plates. This is why we said in the previous paragraph that the radius of the curve defining surfaces S and S' must be very large in order that one surface has no net flux of displacement current and the other has all of it. Whatever their size, however, Eq. 8.13 is valid.

It was mentioned above that a steady current from a point source that spreads uniformly in all directions generates no magnetic field according to the Biot–Savart law. Yet any circular loop has current flowing through it, so Ampere's law suggests that there is a field. The discrepancy is resolved by noting that the current comes from a charge q at the origin that is being drained off at a rate $i = -dq/dt$. This gives rise to a displacement current $\mathbf{j}_d$ that cancels $\mathbf{j}$ (see Problem 11).

8.3 The Magnetic Field Around an Axon

We can use the Biot–Savart law to calculate the magnetic field due to an action potential propagating down an infinitely long axon stretched along the x axis and embedded in an infinite homogeneous conducting medium. Section 7.1 showed that there are three components to the current: i_i along the interior of the axon, di_o out through the membrane (including both displacement current and conduction current), and current in the surrounding medium.

The principle of superposition allows us to calculate the field due to the exterior current by finding the magnetic field $d\mathbf{B}$ from current di_o into the surrounding medium from axon element dx, and then integrating along the axon. We saw in Chap. 7 that the current in the *external* medium from a small element dx flows uniformly in all directions, as if from a point source. We learned in the preceding section that the magnetic field generated by a spherically symmetric radial current is zero. Therefore, in the approximation that the axon is very thin, we can ignore the external current from each element dx. We can do this only because the medium is infinite, homogeneous, and isotropic. When the exterior conductor has boundaries or structure, the symmetry is broken and the external currents contribute to the magnetic field. Our calculation breaks down very close to the axon. Distortions from the field due to the external current because the axon is not infinitely thin are about 1 % near the axon. The current through the cell membrane gives a very small contribution to the magnetic field—roughly 1 part in 10^6.

The major contribution is therefore from i_i. We use the law of Biot–Savart, Eq. 8.12. The observation point is in the xy plane at $(x_0, y_0, 0)$ and the axon lies along the x axis so that $d\mathbf{s} = \hat{\mathbf{x}}\,dx$, as shown in Fig. 8.11. The product $d\mathbf{s} \times \mathbf{r}$

can be evaluated using Eq. 1.8 or 1.9:

$$d\mathbf{s} \times \mathbf{r} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ dx & 0 & 0 \\ x_0 - x & y_0 & 0 \end{vmatrix} = dx\, y_0\, \hat{\mathbf{z}}.$$

The term in the denominator is $r^3 = \left[(x_0 - x)^2 + y_0^2\right]^{3/2}$. The magnetic field in the xy plane is in the z direction and has magnitude

$$B_z = \frac{\mu_0 y_0}{4\pi} \int \frac{i_i(x)\, dx}{\left[(x_0 - x)^2 + y_0^2\right]^{3/2}}.$$

It was shown in Eqs. 7.17 that $i_i = -\pi a^2 \sigma_i (dv_i/dx)$. The final expression for B_z is

$$B_z = -\frac{\mu_0 a^2 \sigma_i y_0}{4} \int \frac{\left[dv_i(x)/dx\right] dx}{\left[(x_0 - x)^2 + y_0^2\right]^{3/2}}. \tag{8.14}$$

The computer program in Fig. 8.12 evaluates the field for the same crayfish axon whose external potential was studied in Sect. 7.4. The field at a distance $2a$ from the axon is plotted in Fig. 8.13. The results agree well with more sophisticated calculations (Swinney and Wikswo 1980; Woosley et al. 1985). The latter reference is particularly clear and should be accessible to those who have studied the convolution integral in Chap. 12. A three-dimensional plot of their results is shown in Fig. 8.14.

It is worth repeating that a calculation this simple succeeds only because the axon and the exterior medium are infinite. If there are boundaries, or if there are regions in the external medium where the conductivity changes, then current in the external medium does contribute to the magnetic field. For example, an isolated nerve preparation in air would have the external current flowing in a thin layer of ionic solution along the outside of the axon, where it would generate a field that almost completely cancels that from i_i.

An approximation valid at large distances can be obtained from Eq. 8.14 by expanding the denominator in much the same way we did to obtain Eqs. 7.26 and 7.27. The observation point is (R, θ) in the xy plane. In this case we need the expansion of

$$\frac{1}{r^3} = \frac{1}{R^3}\left(1 - 2\frac{x}{R}\cos\theta + \frac{x^2}{R^2}\right)^{-3/2} \approx \frac{1}{R^3}\left(1 + \frac{3x\cos\theta}{R} + \cdots\right).$$

The final result is

$$B_z = \frac{\mu_0\, \pi a^2\, \sigma_i \sin\theta}{4\pi R^2}\left(v_i(x_1) - v_i(x_2)\right) + \frac{\mu_0\, \pi a^2\, \sigma_i 3 \sin\theta\cos\theta}{4\pi R^3}\left[-x v_i(x)\big|_{x_1}^{x_2} + \int_{x_1}^{x_2} v_i(x)\, dx\right]. \tag{8.15}$$

The first term is proportional to the current dipole, $\mathbf{p}$, defined for the depolarization in the previous chapter. For a complete pulse the first term vanishes and the second term is used.

8.4 The Magnetocardiogram

It is now feasible to measure magnetic fields arising from the electrical activity of the heart (the *magnetocardiogram* or *MCG*) and the brain (the *magnetoencephalogram* or *MEG*). The models developed in Sect. 8.3 and in Chap. 7 can be used to compare the electric and magnetic signals from a current dipole $\mathbf{p}$. The instrumentation for these measurements is described in Sect. 8.9.

For a single cell at the origin in a homogeneous conducting medium, the exterior potential at observation point $\mathbf{r}$ is given by Eq. 7.13:

$$v = \frac{\mathbf{p} \cdot \mathbf{r}}{4\pi\sigma_o r^3}.$$

The current dipole $\mathbf{p}$ points along the cell in the direction of the advancing depolarization wave and has magnitude (Eq. 7.12) $p = \pi a^2 \sigma_i \Delta v_i$. An expression analogous to Eq. 7.13 describes the magnetic field of a depolarizing cell. We consider the field due to current along the x axis and then generalize the result. The derivation begins with Eq. 8.15 and uses the geometry of Fig. 8.11. The region of depolarization occupies only a millimeter or so along the cell. Since the measurements are made much farther away, the denominator can be removed from the integral, which is then just $\int (dv/dx)\, dx$. If the depolarization is at the origin, then the expression for B_z for $z = 0$ is

$$B_z = -\frac{\mu_0 a^2 \sigma_i y_0 \left[v(x_2) - v(x_1)\right]}{4(x_0^2 + y_0^2)^{3/2}} = \frac{\mu_0}{4\pi}\frac{p y_0}{(x_0^2 + y_0^2)^{3/2}}. \tag{8.16}$$

Figure 8.11 shows that $y_0 = r\sin\theta$, so that $p y_0 = p\, r \sin\theta = |\mathbf{p} \times \mathbf{r}|$. The direction of $\mathbf{B}$ is also consistent with the cross product. Generalizing, we have for a single cell,

$$\mathbf{B} = \mu_0 \frac{\mathbf{p} \times \mathbf{r}}{4\pi r^3}. \tag{8.17}$$

Note the remarkable similarity between Eqs. 7.13 and 8.17. One involves the dot product, and the other the cross product. For both, the field falls as $1/r^2$. If we are considering the cardiogram, either field from the entire heart is the superposition of the field from many cells. As with the electrocardiogram, the first approximation for the magnetocardiogram is to ignore changes in $1/r^2$ and speak of the total current-dipole vector.

Measurements of either the potential or the magnetic field can be used to determine the location of $\mathbf{p}$. We will adopt the coordinate system usually used for the magnetocardiogram.

```
/*This program integrates to
obtain the magnetic field for
a problem which was first
solved by K. R. Swinney and J.
P. Wikswo, the extracellular
magnetic field of the single
active nerve fiber in a volume
conductor. Biophys. J. 32:719-732
(1980). The nerve pulse is
a series of Gaussians used by
Clark and Plonsey to fit the
data of Watanabe and Grund-
fest, J. Gen. Physiol. 45:267
(1961). Uses Romberg integra-
tion routine qromb from W. H.
Press et al. Numerical Recipes
in C.*/

#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include "nr.h"

/*Global Variables*/
const double pi = 3.14159265359;
double x0, y0,  //coordinates
 of observation point
A[3], B[3], C[3];//parameters
for gaussian
float f(float x)
{
/*Calculates the integrand
dv/dx divided by r^3*/
  double xx, r, dv;
  int i;
  dv = 0;
  for (i=0; i<3; i++)
  {  xx = B[i] * (x- C[i]);
    dv = dv-2 * A[i] * B[i] *
          xx * exp(-xx * xx);}
  r = sqrt((x0 - x) * (x0 - x)
     + y0 * y0);
  return dv / (r * r * r);
}
void main()
{
  FILE *ofp;//output file pointer
  const double Mu0 = 4.0e-7 *
pi,  //permeability free space
  axonradius = 6.0e-5,
  xstart = 0.0,
  xfinish = 0.020;
double integral, MagField,
    y[4];

//Calculate at four different
distances
//from axon
int i;

  if (!(ofp =
fopen("OutputFile","w")))
  {printf("cannnot open output
file\n");  exit(1);}
  A[0] = 0.051;
  A[1] = 0.072;
  A[2] = 0.018;
  B[0] = 800;
  B[1] = 533;
  B[2] = 333;
  C[0] = 0.0054;
  C[1] = 0.0066;
  C[2] = 0.0086;
  y[0] = 2 * axonradius;
  y[1] = 0.001;
  y[2] = 0.003;
  y[3] = 0.01;
printf("%12.5g %12.5g %12.5g
%12.5g\n",
y[0],y[1],y[2],y[3]);
fprintf(ofp, "%12.5g\t
%12.5g\t%12.5g\t%12.5g\n",
y[0],y[1],y[2],y[3]);
  for(x0 = xstart; x0<xfinish;
x0+=0.00025)
  {
    printf("%12.5f",x0);
    fprintf(ofp,"%12.5f",x0);
    for(i=0; i<4; i++)
    {
      y0 = y[i];
      integral = qromb(f,
xstart, xfinish);
      MagField = -Mu0 *
axonradius * axonradius * y0 *
integral / 4.0;
      printf("%12.5e",MagField);
  fprintf(ofp,"\t%12g",MagField);
    }
    printf("\n");
    fprintf(ofp,"\n");
  }
  fclose(ofp);
}
```

Fig. 8.12 The program used to calculate the magnetic field outside an axon in an infinite homogeneous conductor using Eq. 8.14. It uses the Romberg integration routine `qromb` from Press et al. (1992)

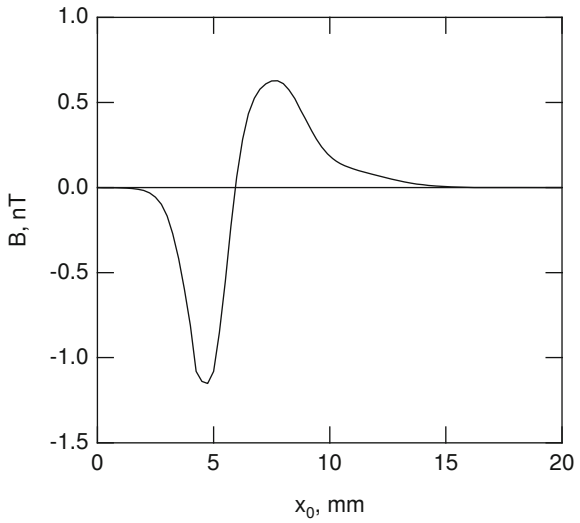


Fig. 8.13 The magnetic field B_z 0.12 mm from a crayfish axon in an infinite homogeneous conducting medium is shown. The field was calculated using the program of Fig. 8.12. The exterior potential for this configuration was calculated in Sect. 7.4

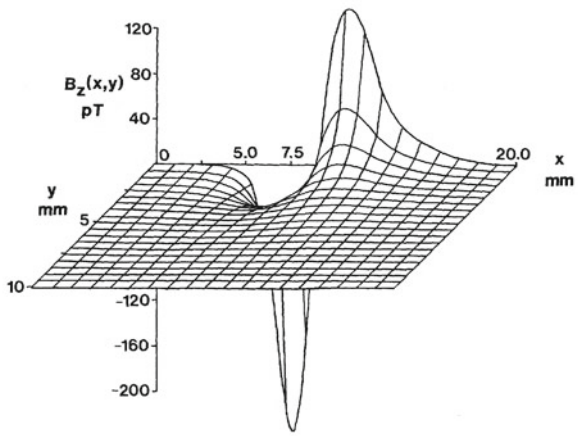


Fig. 8.14 A three-dimensional plot of the magnetic field around the crayfish axon. The minimum distance from the axon is 0.5 mm

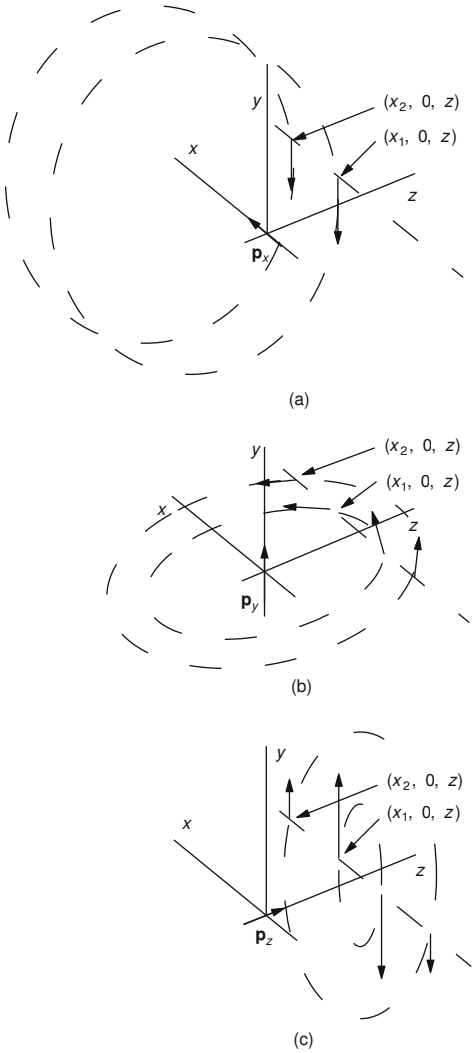


Fig. 8.15 The magnetic field produced by the three components of a current dipole at the origin. The coordinate system is that customarily used for magnetocardiography. The x axis points toward the subject's left, the y axis is vertical, and the z axis points forward through the subject's chest. The coordinate system is viewed over the subject's right shoulder

The x axis points to the patient's left, the y axis points up, and the z axis points toward the front of the patient, roughly perpendicular to the chest wall. Assume that **p** is at the origin and the anterior chest surface is the xy plane at some fixed value of z. We ignore distortions to the field which arise because no current can flow in the region beyond the body, and we assume that the conductivity of the body is homogeneous and isotropic. From Eq. 8.17, we obtain the three components of **B** along the line $(x, 0, z)$:

$$B_x = \frac{\mu_0 p_y z}{4\pi r^3},$$
$$B_y = \frac{\mu_0 (p_z x - z p_x)}{4\pi r^3}, \tag{8.18}$$
$$B_z = -\frac{\mu_0 p_y x}{4\pi r^3}.$$

Compare these results to the lines of **B** in Fig. 8.15, which were drawn for the three components of **p** using the right-hand rule. Along the line being considered ($y = 0, z =$ const), p_x contributes only to B_y, and B_y is always negative. Component p_y contributes to both B_x and B_z; the latter

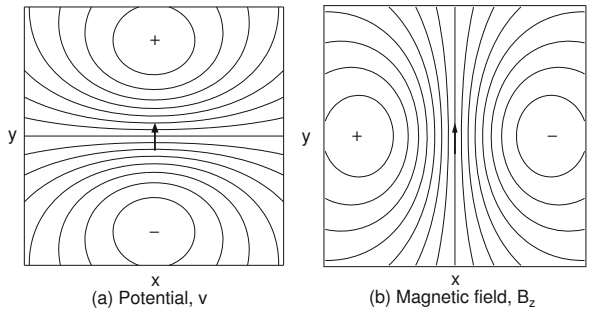


Fig. 8.16 Contour plots in the xy plane for **a** the potential and **b** the z component of the magnetic field from a current dipole **p** pointing along the y axis, calculated for an infinite, isotropic conducting medium

changes sign while the former does not, as we change the value of x. Component p_z gives only a y component of **B** that changes sign as x changes sign. The component normal to the body surface, B_z is given by

$$B_z(x, 0, z) = -\frac{\mu_0}{4\pi} \frac{p_y x}{(x^2 + z^2)^{3/2}}.$$

Figure 8.16 plots contours for the potential and the magnetic field component B_z perpendicular to the body surface when **p** points along the y axis. Again, distortions because of changes in conductivity are ignored. The similarity of the two sets of contours is clear. The contours of constant potential are proportional to $p_y y/r^3$, while the contours for B_z are proportional to $-p_y x/r^3$. Either contour map can be used to determine the location and depth of **p**. To be specific, consider the contours for B_z. The field is proportional to the function $x/(x^2 + z^2)^{3/2}$, which changes sign right over the source and has a maximum and a minimum at $x = \pm z/\sqrt{2}$. The depth of the source z is related to the spacing Δx along the x axis between the maximum and minimum by

$$z = \frac{\Delta x}{\sqrt{2}}. \tag{8.19}$$

The source is located directly beneath the point on the axis where $B_z = 0$, and its strength is related to the maximum value of B_z by

$$B_z(\max) = \frac{\mu_0 p_y}{6\pi\sqrt{3}z^2}. \tag{8.20}$$

Figure 8.17 shows real maps of the potential and the magnetic field on the surface of the chest. While the basic features are described by the simple current dipole model, the exact shape of the contours in Fig. 8.17 differs from the shape in Fig. 8.16. This is due to variations in conductivity of the body. The surface potential is distorted by conductivity differences throughout the thorax; the magnetic field is particularly susceptible to return currents flowing just below the surface of the body. Hosaka et al. (1976) did an early calculation of the effect of currents at the surface of the torso

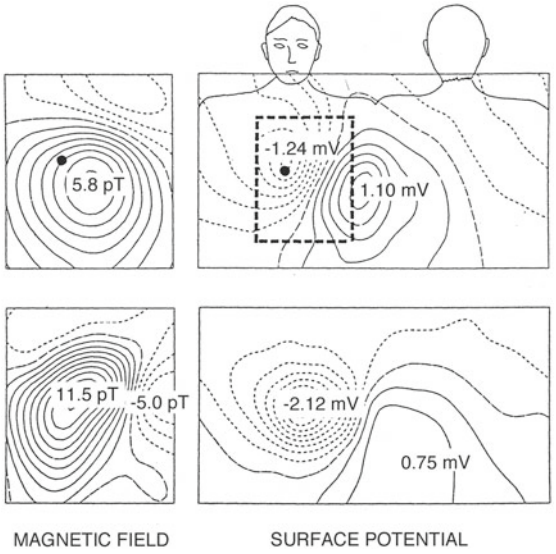


Fig. 8.17 The 56-lead magnetic field map of B_z (left) and the 117-lead potential map (right) during the R-wave maximum for a normal subject (top row) and a patient with an anterior myocardial infarction (bottom row). The maximum and minimum values of each map are indicated on each map. B_z oriented into the page is defined as positive (solid lines) The dashed rectangle in the potential map corresponds to the area for which the magnetic field was measured. The dot in the upper row shows where the midline intersects the level of the fourth intercostal space. Note how the constant contours for the magnetic field are oriented at right angles to the isopotential lines, as in the previous figure. Modified from Stroink (1992) Used by permission

on the magnetocardiogram. They found that the return current modifies the component of **B** perpendicular to the body surface by about 30 %. Tangential components of **B** are influenced more; this is why the normal component B_z is usually measured. Tan et al. (1992) show that using a model of the conductivities that matches the geometry of the patient's thorax allows accurate localization of the current dipole source from the surface measurements. The magnetocardiogram is now being used to record fetal heart arrhythmias (Strasburger et al. 2008).

The magnetic field close to the heart is affected by the anisotropy of the tissue conductivity. Figure 8.18 shows measurements made 1.5 mm from a 1-mm-thick slice of canine myocardium by Staton et al. (1993). Panel A shows the time course of simultaneous recordings from three pickup coils 3 mm in diameter and separated by 4 mm. There are striking differences over 4 mm. Panel B shows a magnetic field contour map during stimulus from another experiment. Instead of having one peak and one valley as in Figs. 8.16 and 8.17, it shows a cloverleaf or *quatrefoil* pattern. Panels C and D show the field contours and the current flow in a third experiment, 6 ms after stimulation. This field and current pattern is predicted by bidomain calculations (Wikswo 1995b).

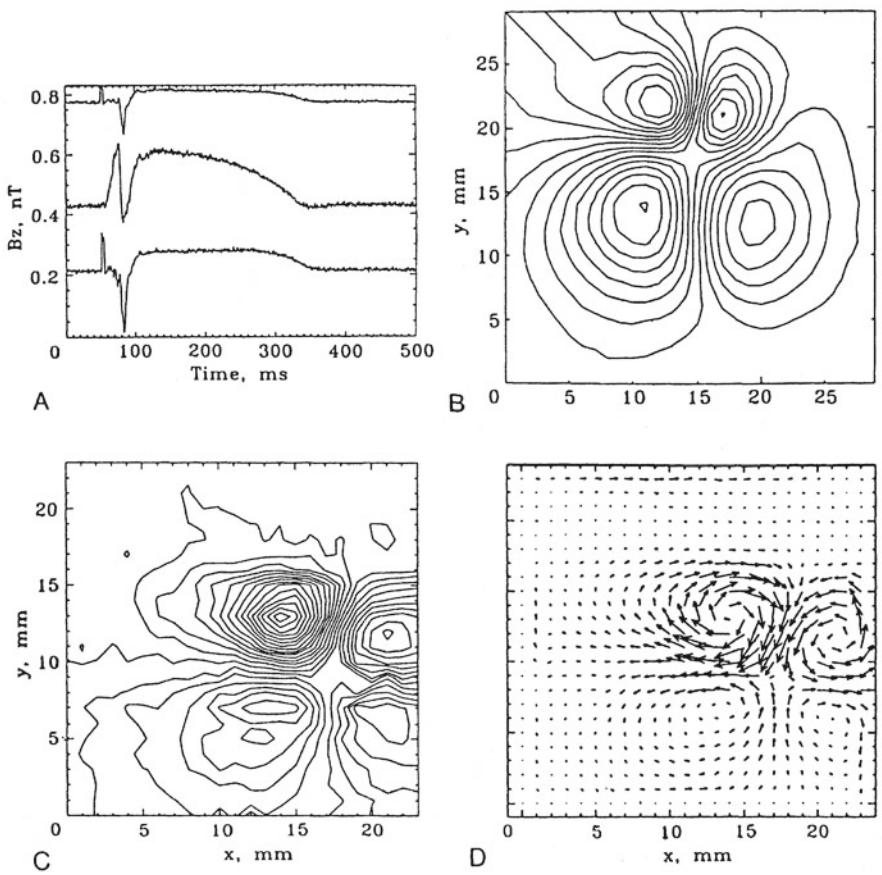


Fig. 8.18 The results of magnetic field measurements very close to a slice of canine myocardium. The panels are described in the text. Part of the figure is reproduced from Wikswo 1995b ©1995 Elsevier, Inc, with permission of Elsevier. The rest is from Staton et al. (1993), ©1993 IEEE

8.5 The Magnetoencephalogram

The magnetic signals from a nerve action potential are weaker than those from the heart for two reasons. First, the current-dipole vector associated with the repolarization follows close behind the depolarization and reduces the field. (The largest unmyelinated axons in the body have a conduction speed of about 1 m s^{-1} and the pulse is about 1 mm long. Myelinated fibers have a pulse length up to 80–100 times longer.) Second, the cross-sectional area of the advancing wavefront is much smaller. However, the magnetic fields accompanying action potentials have been measured in nerve (Barach et al. 1985; Roth and Wikswo 1985) and in muscle (Gielen et al. 1991). They have also been measured in green algae (Trontelj et al. 1994).

We saw in Sect. 6.1 that nerve cells have an input end (dendrites), a cell body, and an axon. The signal that propagates from a synapse through the dendrites to the cell body and axon is much smaller (about 10 mV) and longer (10 ms) than an action potential that travels along the axon. The cells at the surface of the cerebral cortex have dendrites that are like the trunk of a tree perpendicular to the surface of the cortex, with branches from several directions coming to the trunk. The signal from the trunk is the primary contributor to the magnetoencephalogram (MEG) and electroencephalogram (EEG). The problems show that the magnetic field associated with the rise of the postsynaptic potential is more easily observed outside the brain than is the action potential.

One can see from the symmetry argument in the caption of Fig. 8.19 that in a spherically symmetric conducting medium the radial component of **p** and its return currents do not generate any magnetic field outside the sphere. Therefore the MEG is most sensitive to detecting activity in the fissures of the cortex, where the trunk of the postsynaptic dendrite is perpendicular to the surface of the fissure. A tangential component of **p** does produce a magnetic field outside a spherically symmetric conductor. The extracellular current does not contribute to the radial component of the magnetic field (Hämäläinen et al. 1993), so Eq. 8.17 gives B_r

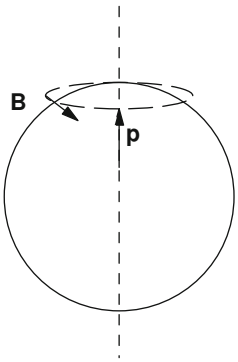


Fig. 8.19 A current dipole **p** is oriented radially inside a homogeneous conducting sphere. The return current is independent of the azimuthal angle, ϕ. By symmetry the magnetic field, if any, must be in the ϕ direction; the current in Ampere's circuital law, which is the sum of the current in p and the return current, is zero

correctly. Extracellular current does influence the tangential components of the magnetic field. Since the skull is not a perfect sphere, there is some effect of the radial component of **p** on the MEG. The EEG is sensitive to both radial and tangential components of **p**. The information available from the EEG and MEG has been reviewed by Wikswo et al. (1993).

In the last decade, the use of the MEG has grown dramatically for conditions such as epilepsy, stroke, chronic pain, and dyslexia (Hari and Salmelin 2012).

Measurements of the magnetoencephalogram are often based on evoked responses. A repetitive stimulus—audible, visual, or tactile—is presented to the subject or the subject is asked to perform a repetitive task such as flexing a finger. Signal-averaging techniques are used to identify the associated changes in magnetic field (see Chap. 11). Figure 8.20 shows averaged magnetic field contours measured over the scalp of a subject who heard a string of words presented in random order every 2.3 s. Sometimes the subject was asked to read something else and ignore the words. At other times the subject was asked to pay attention and count how many of the words were on a list. The first peak, 100 ms after presentation of the word, was the same in both cases. The sustained field peak, SF, was considerably stronger when the subject was paying attention to the list. Magnetic contours and the equivalent current dipole source are also shown.

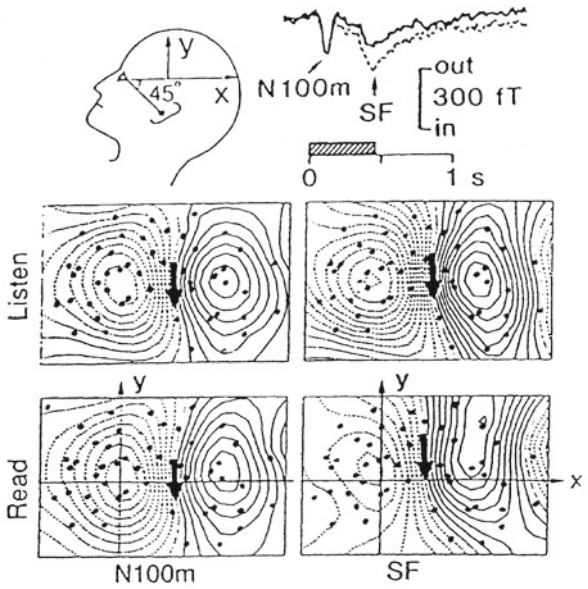


Fig. 8.20 Magnetic field maps recorded over the scalp of a subject who heard a series of words and either ignored them by reading something else or listened carefully and counted how many of the words were in a predetermined list. The features are discussed in the text. Reprinted with permission from Hämäläinen et al. 1993. Copyright © 1993 by the American Physical Society

8.6 Electromagnetic Induction

In 1831 Michael Faraday discovered that a *changing* magnetic field causes an electric current to flow in a circuit. It does not matter whether the magnetic field is from a permanent magnet moving with respect to the circuit or from the changing current in another circuit. The results of many experiments can be summarized in the *Faraday induction law*:

$$\oint \mathbf{E} \cdot d\mathbf{s} = -\frac{d}{dt} \iint \mathbf{B} \cdot d\mathbf{S} = -\frac{d\Phi}{dt}. \tag{8.21}$$

It states that the line integral of **E** around a closed path is equal to minus the rate of change of the magnetic flux through any surface bounded by the path. The relationship between the direction of **S** and $d\mathbf{s}$ is given by a right-hand rule: if the fingers of the right hand curl around the circuit in the direction of $d\mathbf{s}$, the thumb of the right hand points in the direction of a positive normal to **S**. The units of magnetic flux $\Phi = \iint \mathbf{B} \cdot d\mathbf{S}$ are T m^2 or weber (Wb). Rapidly changing magnetic fields can induce currents large enough to trigger nerve impulses. This is discussed in Sect. 8.7.

The differential form of the Faraday induction law is (see Problem 22)

$$\operatorname{curl} \mathbf{E} = \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{8.22}$$

The result of the vector operation *curl* is another vector. In Cartesian coordinates the components of $\nabla \times \mathbf{E}$ are

$$(\nabla \times \mathbf{E})_x = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z},$$
$$(\nabla \times \mathbf{E})_y = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x},$$
$$(\nabla \times \mathbf{E})_z = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}.$$

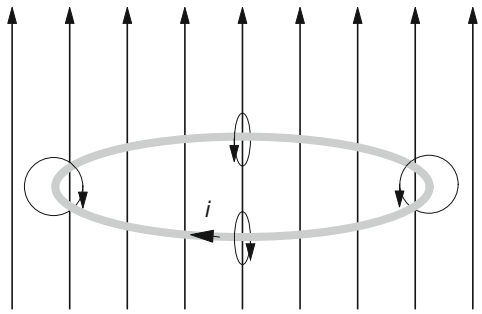


Fig. 8.21 A magnetic field increasing in the direction shown induces a current in the loop. This current generates a magnetic field in the opposite direction, opposing the change in the magnetic field

These can be abbreviated by using determinant notation as

$$(\nabla \times \mathbf{E}) = \begin{vmatrix} \widehat{\mathbf{x}} & \widehat{\mathbf{y}} & \widehat{\mathbf{z}} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix}. \tag{8.23}$$

Similarly (Problem 23) the differential form of Ampere's law is

$$\operatorname{curl} \mathbf{B} = \nabla \times \mathbf{B} = \mu_0 \left(\mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \right). \tag{8.24}$$

The integral form of the Faraday induction law can be used to determine $\mathbf{E}$ only if the symmetry is such that $\mathbf{E}$ is always parallel to $d\mathbf{s}$ and has the same magnitude all along the path. One situation where it can be used is a circular loop of wire in the xy plane centered at the origin. The radius of the loop is a and its normal is along the $+z$ axis. Suppose that everywhere in the xy plane within the boundary of the circle the field points along z and depends only on time: $\mathbf{B}(x, y, z, t) = B(t)\widehat{\mathbf{z}}$. Symmetry shows that $\mathbf{E}$ has the same magnitude everywhere in the wire and is always tangent to the loop. Equation 8.21 gives

$$E = -\frac{a}{2}\frac{dB}{dt}. \tag{8.25}$$

If the loop is made of material that obeys Ohm's law, there is a current of density $j = \sigma E = -(\sigma a/2)(dB/dt)$. If the radius of the wire is b ($b \ll a$), then $i = -(\sigma \pi a b^2/2)(dB/dt)$. Figure 8.21 shows the direction of the induced current if dB/dt is positive. The induced current sets up its own magnetic field which points in the $-z$ direction within the loop, opposing the primary field increase within the loop. The induced current always opposes the *change* of magnetic field that produces it. This is called *Lenz's law*. If it were not true, the induced current, once started, would increase indefinitely.

This result does not require that the ring be hollow; it can be part of a much larger conductor. The larger the conductor, the greater the radius of the path along which the induced current can flow. The currents that changing magnetic fields induce in conductors are called *eddy currents* and cause heating losses in the conductor. Iron, which is a conductor, is often used as a core in transformer windings to increase the intensity of the magnetic field. To reduce the eddy-current losses, the cores are made of thin layers of iron insulated from one another by varnish. This limits the radius of the path in which the eddy currents can flow. Some coils and transformers are wound on cores of powdered iron dispersed in an insulating binder. Rooms with thick conducting walls (aluminum, about 2-cm thick) have been used to shield against 60-Hz magnetic fields from power wiring. The eddy currents induced in the aluminum attenuate the field by about a factor of 200 (Stroink et al. (1981)).

The quantity $\int_a^b \mathbf{E} \cdot d\mathbf{s}$ is the work done per unit charge in moving from a to b and is called the *electromotive force along the path from a to b*. Terminology is not always consistent; see the discussion by Page (1977). The details of how a changing magnetic field causes a current to flow were shown above for a circular conductor. The force on a moving charge due to the induced electric field is balanced by the drag force as the charge drifts through the conductor. Energy supplied by the changing magnetic field is dissipated as heat. If a voltmeter is attached to two points on the circle, the voltmeter reading may seem paradoxical, until one realizes that there may be changing flux in the voltmeter leads as well. An additional complication is that when there is any region of space in which $\nabla \times \mathbf{E} \neq 0$, then it is possible for $\int_a^b \mathbf{E} \cdot d\mathbf{s}$ to depend on the path (rather than just the end points), even if the magnetic field is zero at all points on the path. This is described clearly and in detail by Romer (1982).

8.7 Magnetic Stimulation

Since a changing magnetic field generates an induced electric field, it is possible to stimulate nerve or muscle cells without using electrodes. The advantage is that for a given induced current deep within the brain, the currents in the scalp that are induced by the magnetic field are far less than the currents that would be required for electrical stimulation. Therefore *transcranial magnetic stimulation* (TMS) is relatively painless. It is also safe (Rossi et al. 2009).

Magnetic stimulation can be used to diagnose central nervous system diseases that slow the conduction velocity in motor nerves without changing the conduction velocity in sensory nerves (Hallett and Cohen 1989). It could be used to monitor motor nerves during spinal cord surgery, and to map motor brain function. Because TMS is noninvasive and nearly painless, it can be used to study learning and plasticity (changes in brain organization over time; Wassermann

et al. 2008). Recently, researchers have suggested that repetitive TMS might be useful for treating disorders such as depression (O'Reardon et al. 2007) and Alzheimer's disease (Freitas et al. 2011).

One of the earliest investigations was reported by Barker, Jalinous and Freeston (1985). They used a solenoid in which the magnetic field changed by 2 T in 110 μs to apply a stimulus to different points on a subject's arm and skull. The stimulus made a subject's finger twitch after the delay required for the nerve impulse to travel to the muscle. For a region of radius $a = 10\,\text{mm}$ in material of conductivity $1\ \text{S m}^{-1}$, the induced current density for the field change in Barker's solenoid was $90\ \text{A m}^{-2}$. (This is for conducting material inside the solenoid; the field falls off outside the solenoid, so the induced current is less.) This current density is large compared to current densities in nerves (Chap. 6).

Magnetic stimulators are relatively high-power devices, requiring thousands of amps passed through coils for a few hundred microseconds. Most magnetic stimulators are capacitor discharge devices, in which a large capacitor is charged to a high voltage (several kV) and then discharged through the coil. Different coil geometries have been examined; the most common one is a figure-of-eight shape. Magnetic stimulation is included in the review by Roth (1994) and is compared to other brain imaging methods by Ilmoniemi et al. (1999).

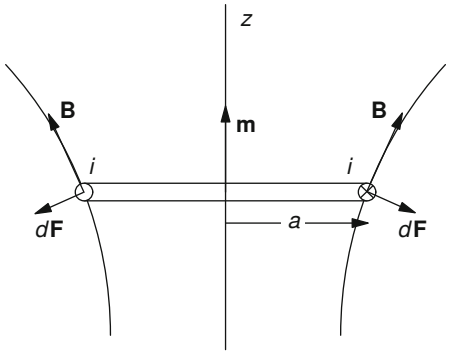


Fig. 8.22 A current loop in an inhomogeneous magnetic field experiences a force toward the region of stronger magnetic field. The circular loop lies in a plane perpendicular to the z axis. Current flows into the page on the *right* and out of the page on the *left*

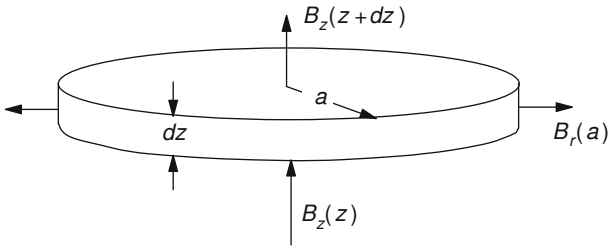


Fig. 8.23 Gauss's law for **B** is applied to a pillbox of radius a and thickness dz

8.8 Magnetic Materials and Biological Systems

Just as the electric field can be altered by the polarization of a dielectric, the magnetic field can be altered by matter. Biological measurements can be based on alterations of the field by an organ in the body. Some cells exhibit permanent magnetism, which is important for measuring direction in some bacteria, birds, and other organisms.

8.8.1 Magnetic Materials

The effects of magnetic fields on material are more complicated than those of electric fields. Since there are no known magnetic charges (monopoles), we must consider the effect of magnetic fields on current loops or magnetic dipoles. Figure 8.22 shows a current loop in a magnetic field that decreases as z increases. As a result the lines of **B** spread apart. The loop has radius a, carries current i, and has magnetic moment[1] **m**. For the orientation shown, there is a force on the loop in the $-z$ direction that is toward the region where the field is stronger. If the magnetic moment of the loop were not parallel to **B**, there would also be a torque on the loop. For ease in calculation, imagine that the loop has been placed in the field in such a way that along the axis of the loop, **B** points in the z direction. Then the spreading of the lines of **B** means that **B** has a component radially outward all around the loop. Because of the symmetry B_r has a constant magnitude everywhere around the loop, and the force on the loop is $-2\pi a\, i\, B_r(a)$.

Field $B_r(a)$ is found by considering the fact that the total magnetic flux through all surfaces of the pillbox in Fig. 8.23 is zero (Eq. 8.8). The net outward flux is

$$\left[B_z(z+dz) - B_z(z)\right]\pi a^2 + B_r(a)2\pi a\, dz = 0.$$

This can be rearranged to give

$$\left[\frac{\partial B_z}{\partial z} + \frac{2}{a}B_r(a)\right]\pi a^2\, dz = 0,$$

from which

$$B_r(a) = -\frac{a}{2}\frac{\partial B_z}{\partial z}.$$

[1] Be careful. We are talking about two different kinds of dipoles in this chapter. The current dipole **p** is a source and sink of current and has units A m. The magnetic dipole **m**, equivalent to a small magnet with north and south poles, has units A m^2. The magnetic field from a magnetic dipole falls off as $1/r^3$.

The force on the loop is therefore

$$F_z = \pi a^2 i \frac{\partial B_z}{\partial z} = m_z \frac{\partial B_z}{\partial z}. \tag{8.26}$$

If **m** is parallel to **B** the force is toward the region of stronger field; if **m** is antiparallel to **B** the force is toward the region of weaker field.

An atom can have a magnetic moment because of two effects.[2] The motion of the electrons in orbit about the nucleus constitutes a current, as a result of which there may be an *orbital magnetic moment*. The intrinsic *spin* of each electron gives rise to a spin magnetic moment, independent of any orbital motion. In most atoms, the orbital magnetic moments average to zero, and most of the electrons are arranged in pairs whose spins cancel. The atom therefore usually has no net magnetic moment.

Most substances placed in an inhomogeneous field experience a weak force *away* from the region of strong field, and the force is roughly proportional to the square of the field strength, an effect called *diamagnetism*. It can be understood with a simple classical model. As the atom is moved into the magnetic field, the Faraday induction effect distorts the orbits of the electrons to induce a magnetic dipole moment proportional to **B** and in the opposite direction, consistent with Lenz's law. The force is therefore proportional to $B_z(\partial B_z/\partial z)$.

A few substances are *attracted* to the region of stronger field, again with a force that is often proportional to the square of the field. Each atom of these *paramagnetic* substances has a permanent magnetic moment associated with the spin of an unpaired electron. Thermal motion normally keeps the magnetic moments of different atoms oriented randomly. As the substance is brought into the magnetic field the spin magnetic moments of different atoms begin to align with the magnetic field. A magnetic dipole moment is induced in the substance, but this time it is in the direction of **B**, and the substance is attracted to the magnet.

Some substances placed in an inhomogeneous magnetic field experience much stronger attraction than do paramagnetic substances. In these substances some of the atomic moments are aligned even in the absence of an external field. They are permanent magnets. Further alignment of the atomic moments may take place in an external field, but complete alignment often takes place in relatively weak external fields. These substances are called *ferromagnets*. The individual atoms have magnetic moments, and there are forces between atoms that cause the spins to align. Section 14-4 of Eisberg and Resnick (1985) provides a relatively simple explanation of the quantum-mechanical effects underlying this spin alignment and the formation of microscopic regions of aligned spins called *domains*. *Ferrimagnets* are similar to ferromagnets, but the crystals contain two different kinds of ions with different magnetic moments.

The *magnetization* **M** is the average magnetic moment per unit volume. It is defined by considering volume ΔV that has total magnetic moment $\Delta\mathbf{m} = \sum \mathbf{m}_i$, where the summation is taken over all atoms in the volume, and taking the ratio

$$\mathbf{M} = \frac{\Delta\mathbf{m}}{\Delta V}. \tag{8.27}$$

We have seen that a current loop possesses a magnetic moment of magnitude $m = iS$. One can imagine a current giving rise to any magnetic moment, even one associated with electron spin. Such currents are called *bound currents* and must be included in Ampere's law. The currents that flow due to conduction—that we can control by changing the conductivity of the material or throwing a switch—are called *free currents*. One can show that if we define the new vector

$$\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}, \tag{8.28}$$

it depends only on the free currents:

$$\oint \mathbf{H} \cdot d\mathbf{s} = \iint \mathbf{j}_{\text{free}} \cdot d\mathbf{S}. \tag{8.29}$$

Vector **H** is called the *magnetic field intensity*. It has units A m^{-1}. It does not have the physical significance of **B** (it does not appear in the Lorentz force or the Faraday induction law). However, it often simplifies computations, because we control free current in the laboratory.

In a vacuum, $\mathbf{B} = \mu_0\mathbf{H}$. It has been traditional to define the *magnetic permeability* of a medium in which **B**, **M**, and **H** are all proportional to one another by the equation

$$\mathbf{B} = \mu\mathbf{H}, \tag{8.30}$$

in which case

$$\frac{\mu}{\mu_0} = 1 + \frac{M}{H} = 1 + \chi_m. \tag{8.31}$$

In diamagnetic materials the *magnetic susceptibility* χ_m is negative and $\mu < \mu_0$. A typical diamagnetic susceptibility is $\approx -1 \times 10^{-5}$. In paramagnetic materials χ_m is positive and $\mu > \mu_0$. A typical paramagnetic susceptibility is $\approx 1 \times 10^{-4}$.

The relationship between B and H in ferromagnetic substances is nonlinear and is characterized by a BH curve. A typical curve is shown in Fig. 8.24. The fact that the curves for increasing and decreasing H do not coincide is called *hysteresis*. The arrows show the direction in which H changes on each branch of the curve. *Saturation* takes place beyond points W and Y. The value of M saturates and $B = \mu_0 (M_{\text{saturated}} + H)$. When $H = 0$ there is a *remanent*

[2] Much weaker magnetic moments of the atomic nucleus are considered in Chap. 18.

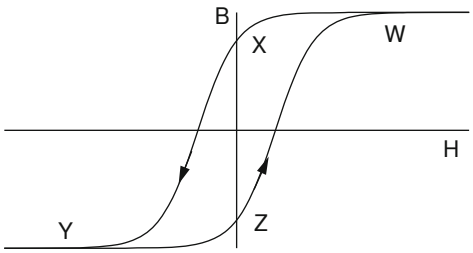


Fig. 8.24 A typical curve of B vs H for a ferromagnetic material. The curve shows hysteresis, and the arrows show the direction of travel around the curve $WXYZ$. Points W and Y show where M saturates. Points X and Z show the remanent magnetic field when $H = 0$

magnetic field (points X and Z). If the temperature of the sample is raised above a critical temperature called the *Curie temperature*, the magnetism is destroyed.

8.8.2 Measuring Magnetic Properties in People

Several kinds of measurements can be based on magnetic effects in materials. A common component of dust inhaled by miners and industrial workers is magnetite, Fe_3O_4, which is ferrimagnetic. By placing the thorax in a fixed magnetic field for a few seconds, the particles can be aligned. The field is turned off and the remanent field measured. The use of *magnetopneumography* in occupational health is described by Stroink (1985). Cohen et al. (1984) have modeled the process by which the particles are magnetized, as well as the relaxation process by which the magnetization disappears after the external field is removed. Relaxation curves are used to estimate intracellular viscosity and the motility of macrophages (scavenger white cells) in the alveoli (Stahlhofen and Moller 1993).

The magnetic susceptibility of blood and myocardium is different from the susceptibility of surrounding lung tissue. An externally applied magnetic field induces a field that changes as the volume of the heart changes. It can be measured externally. The theory and experiments have been described by Wikswo (1980).

Susceptibility measurements can also be used to measure the total iron stores in the body. Normally the body contains 3–4 g of iron. About a quarter of it is stored in the liver. The amount of iron can be elevated from a large number of blood transfusions or in certain rare diseases such as hemochromatosis and hemosiderosis. The liver is an organ whose susceptibility can easily be measured. The susceptibility varies linearly with the amount of iron deposited. Magnetic susceptometry has been used to estimate body iron stores (Nielsen et al. 1995).

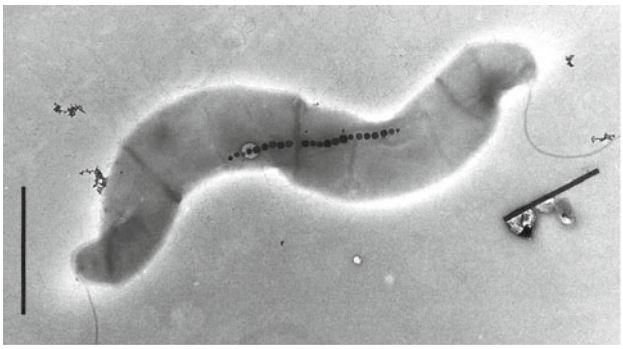


Fig. 8.25 The small black dots are magnetosomes, small particles of magnetite in the magnetotactic bacterium *Aquaspirillum magnetotacticum*. The vertical bar is 1 μm long. The photograph was taken by Y. Gorby and was supplied by N. Blakemore and R. Blakemore, University of New Hampshire.

8.8.3 Magnetic Orientation

Magnetism is used for orientation by several organisms. A history of studies in this area is provided in a very readable book by Mielczarek and McGrayne (2000). Finegold (2012) reviews sensing of static fields. Several species of bacteria contain linear strings of up to 20 particles of magnetite, each about 50 nm on a side encased in a membrane (Frankel et al. 1979; Moskowitz 1995). Over a dozen different bacteria have been identified that synthesize these intracellular, membrane-bound particles or *magnetosomes* (Fig. 8.25). In the laboratory the bacteria align themselves with the local magnetic field. In the problems you will learn that there is sufficient magnetic material in each bacterium to align it with the earth's field just like a compass needle. Because of the tilt of the earth's field, bacteria in the wild can thereby distinguish up from down.

Other bacteria that live in oxygen-poor, sulfide-rich environments contain magnetosomes composed of greigite (Fe_3S_4), rather than magnetite (Fe_3O_4). In aquatic habitats, high concentrations of both kinds of magnetotactic bacteria are usually found near the *oxic–anoxic transition zone* (OATZ). In freshwater environments the OATZ is usually at the sediment–water interface. In marine environments it is displaced up into the water column. Since some bacteria prefer more oxygen and others prefer less, and they both have the same kind of propulsion and orientation mechanism, one wonders why one kind of bacterium is not swimming out of the environment favorable to it. Frankel and Bazylinski (1994) proposed that the magnetic field and the magnetosomes keep the organism aligned with the field, and that they change the direction in which their flagellum rotates to move in the direction that leads them to a more favorable concentration of some desired chemical.

Magnetosomes are found in other species and are likely also to be used for orientation. One species of algae contains about 3000 magnetic particles, each of which is about $40 \times 40 \times 140$ nm (de Araujo et al. 1986). Bees, pigeons, and fish contain magnetic particles. It is more difficult to demonstrate their function, because of the variety of other sensory information available to these animals. For example, homing pigeons with magnets attached to their heads could orient well on sunny days but not on cloudy ones (Walcott et al. 1979). There is evidence that bees orient in a magnetic field (Frankel 1984). The net magnetic moment in the bees is oriented transversely in the body (Gould et al. 1978). In pigeons the magnetic material is located in the dura (the outer covering of the brain) or skull. In all of these cases, the material has been identified as magnetite. In the yellowfin tuna, data are compatible with about 8.5×10^7 magnetic particles, each of which is a single domain of magnetite in the shape of an approximately 50-nm cube (Walker et al. 1984). Recently, Eder et al. (2012) isolated cells from the trout nose and used a rotating magnetic filed to identify cells that are potential magnetite-based magnetoreceptor cells.

Birds may actually have three compasses. Since the magnetic and geographic poles are fairly far apart, migratory birds must correct their magnetic compasses as they fly. The Savannah sparrow is known to have a magnetic compass and a star compass and to take visual cues from the sky at sunset. Able and Able (1995) have shown that adult Savannah sparrows that are subjected to a field pointing in a different direction than the earth's field will at first trust their magnetic compasses, but over a few days they recalibrate their magnetic compasses with their star compasses. An accompanying editorial (Gould 1995) places their work in context. More recently, Cochran et al. (2004) have shown that if migrating thrushes are placed in an eastward-pointing magnetic field at twilight and then released, they fly west instead of south. This strongly suggests that the birds recalibrate their magnetic compass at twilight each day.

The fact that the magnetite particles seem to be about 50 nm on a side is physically significant. Frankel (1984) summarizes arguments that if the particles are smaller than about 35 nm on a side, thermal effects can destroy the alignment of the individual particles. If they are larger than about 76 nm, multiple domains can form within a particle, decreasing the magnetic moment.

8.8.4 Magnetic Nanoparticles

Small single-domain nanoparticles (10–70 nm in diameter) are used to treat cancer (Jordan et al. 1999; Pankhurst et al. 2009). The particles are injected into the body intravenously. Then an oscillating magnetic field is applied. It causes the particles to rotate, heating the surrounding tissue. Cancer cells are particularly sensitive to damage by hyperthermia.

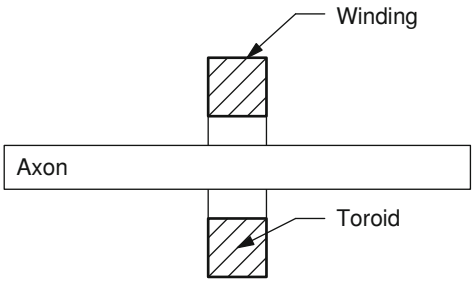


Fig. 8.26 A nerve cell preparation is threaded through the magnetic toroid to measure the magnetic field. The changing magnetic flux in the toroid induces an electromotive force in the winding. Any external current that flows through the hole in the toroid diminishes the magnetic field

Often the surface of the nanoparticle can be coated with antibodies that cause the nanoparticle to be selectively taken up by the tumor, providing more localized heating of the cancer.

8.9 Detection of Weak Magnetic Fields

The detection of weak fields from the body is a technological triumph. The field strength from lung particles is about 10^{-9} T; from the heart it is about 10^{-10} T; from the brain it is 10^{-12} T for spontaneous (α-wave) activity and 10^{-13} T for evoked responses. These signals must be compared to 10^{-4} T for the earth's magnetic field. Noise due to spontaneous changes in the earth's field can be as high as 10^{-7} T. Noise due to power lines, machinery, and the like can be 10^{-5}–10^{-4} T.

If the signal is strong enough, it can be detected with conventional coils and signal-averaging techniques that are described in Chap. 11. Barach et al. (1985) used a small detector through which a single axon was threaded. The detector consisted of a toroidal magnetic core wound with many turns of fine wire (Fig. 8.26). Current passing through the hole in the toroid generated a magnetic field that was concentrated in the ferromagnetic material of the toroid. When the field changed, a measurable voltage was induced in the surrounding coil. This *neuromagnetic current probe* has been used to study many nerve and muscle fibers (Wijesinghe 2010).

The signals from the body are weaker, and their measurement requires higher sensitivity and often special techniques to reduce noise. Hämäläinen et al. (1993) present a detailed discussion of the instrumentation problems. Sensitive detectors are constructed from *superconducting* materials. Some compounds, when cooled below a certain critical temperature, undergo a sudden transition and their electrical resistance falls to zero. A current in a loop of superconducting wire persists for as long as the wire is maintained

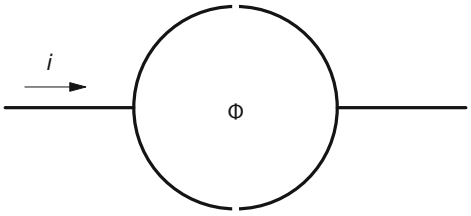


Fig. 8.27 A dc SQUID is shown. The solid lines represent superconducting wires, broken by Josephson junctions at the top and bottom. The total current through both wires depends on Φ, the magnetic flux through the circle

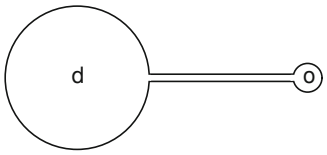


Fig. 8.28 A superconducting loop shaped as shown becomes a flux transporter. Because the total flux in the loop is constant, a change of flux in the detecting loop d is accompanied by an equal and opposite flux change in the output loop o. The diameter of the output loop is matched to the size of the SQUID. Sensitivity is increased because the detecting loop has a larger area

in the superconducting state. The reason there is a superconducting state is a well-understood quantum-mechanical effect that we cannot go into here. It is due to the cooperative motion of many electrons in the superconductor (Eisberg and Resnick 1985, Sect. 14.1; Clarke 1994). The integral $\oint \mathbf{E} \cdot d\mathbf{s}$ around a superconducting ring is zero, which means that $d\Phi/dt$ is zero, and the magnetic flux through a superconducting loop cannot change. If one tries to change the magnetic field with some external source, the current in the superconducting circuit changes so that the flux remains the same.

The detector is called a *superconducting quantum interference device* (*SQUID*). The operation of a SQUID and biological applications are described in the *Scientific American* article by Clarke (1994). Wikswo (1995a) surveys the use of SQUIDs for applications in biomagnetism and nondestructive testing. A technical discussion is also available (Hämäläinen et al. 1993). The dc SQUID requires a superconducting circuit with two branches, each of which contains a very thin nonsuperconducting "weak link" known as a *Josephson junction* (Fig. 8.27). As the magnetic field is changed, these weak links allow the flux in the loop to change. The phase of the quantum mechanical wave function of the collectively moving electrons differs in the two branches by an amount depending on the magnetic flux linked by the circuit. The total current depends on the interference of these two wave functions and is of the form $I = 2I_0 \cos(\pi\Phi/\Phi_0)$, where Φ is the flux through the circuit. The quantity $\Phi_0 = h/2e$, where h is Planck's constant (see Chap. 14) and e is the electron charge, is the *magnetic flux quantum* and has a value equal to 2.068×10^{-15} T m^2. Because interference changes corresponding to a small fraction of this can be measured, the SQUID is very sensitive. The SQUID must be operated at temperatures where it is superconducting. It used to be necessary to keep a SQUID in a liquid-helium bath, which is expensive to operate because of the high evaporation rate of liquid helium. With the advent of high-temperature superconductors, SQUIDS have the potential to operate at liquid-nitrogen temperatures, where the cooling problems are much less severe.

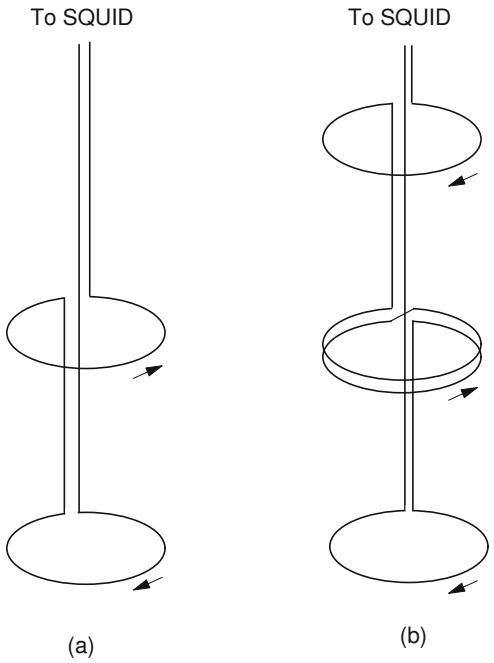


Fig. 8.29 Gradiometers are sensitive to nearby sources of the magnetic field but are much less sensitive to distant sources. **a** A first-order gradiometer. **b** A second-order gradiometer

A typical magnetometer for biomagnetic research contains a *flux transporter*, a superconducting detector coil d a centimeter or so in radius, coupled to a very small multiturn output coil o that matches the size of the SQUID and is placed right next to it. This is shown schematically in Fig. 8.28. The wires between the two loops are close together and have negligible area between them. The total flux, which is constant because the entire circuit is superconducting, is $\Phi = \Phi_d + \Phi_o$. The large area of the detecting coil increases its sensitivity. Any change in the magnetic field at the detector causes an opposite change in the flux and magnetic field at the output coil.

Because ambient natural and artificial background magnetic fields are so high, measurements are often made in special shielded rooms. These can be built of ferromagnetic

materials, or of conductors to take advantage of eddy current attenuation, or they may have active circuits to cancel the background fields. It has proven possible in some cases to eliminate the need for these expensive rooms by using specially designed flux transporters that are less sensitive to distant sources but measure the nearby source with almost the same sensitivity as a single loop. If a distant background source can be represented by a magnetic dipole, the field falls as $1/r^3$. The signal in a magnetometer (Fig. 8.28) would be proportional to this.

Problem 41 shows that the signal from a distant dipole detected by a first-order gradiometer (Fig. 8.29a) is proportional to $1/r^4$ and that the signal in a second order gradiometer (Fig. 8.29b) is proportional to $1/r^5$. Both gradiometers are insensitive to background that does not vary with position. Yet the loop closest to the nearby signal source detects a much stronger signal than the loops that are further away. With modern multi-channel detector systems, one need not use gradiometer coils. Hundreds of coils are used at different locations, and the signals from them are combined to give the same suppression of background from distant sources.

Symbols Used in Chapter 8

Symbol	Use	Units	First used page
a, b	Distance	m	214
e	Elementary charge	C	230
f	Frequency	Hz	215
h	Planck's constant	J s	230
i	Current	A	214
j, $\mathbf{j}$	Current density	A m^{-2}	216
j_d	Displacement current density	A m^{-2}	217
m, $\mathbf{m}$	Magnetic moment	A m^2	214
m	Mass	kg	215
$\mathbf{p}$	Current dipole moment	A m	219
q	Charge	C	213
r, $\mathbf{r}$	Distance	m	215
s, $\mathbf{s}$	Linear displacement	m	213
t	Time	s	217
v, $\mathbf{v}$	Velocity	m s^{-1}	213
v	Electrical potential	V	219
x, y, z	Coordinates	m	217
$\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$	Unit vectors		219
A	Area	m^2	217
B, $\mathbf{B}$	Magnetic field	T	213
C	Particle concentration	m^{-3}	214
D, $\mathbf{D}$	Electric displacement	C m^{-2}	217
E, $\mathbf{E}$	Electric field	V m^{-1}	214
F, $\mathbf{F}$	Force	N	213
H, $\mathbf{H}$	Magnetic field intensity	A m^{-1}	227
I	Current	A	230
M, $\mathbf{M}$	Magnetization	A m^{-1}	227
M	Mutual inductance	V s A^{-1}	236
R	Position	m	219
S, $\mathbf{S}$	Surface area	m^2	214
T	Period	s	215
V	Volume	m^3	227
ϵ_0	Electrical permittivity of free space	N^{-1} m^{-2} C^2	215
θ	Angle		214
κ	Dielectric constant		217
μ	Magnetic permeability	Ω s m^{-1}	227
μ_0	Magnetic permeability of free space	Ω s m^{-1}	215
σ	Charge per unit area	C m^{-2}	217
σ_i, σ_o	Electrical conductivity	S m^{-1}	219
τ, $\boldsymbol{\tau}$	Torque	N m	214
ϕ	Angle		214
χ_m	Magnetic susceptibility		227
Φ	Magnetic flux	T m^2 or Wb	224
Φ_0	Quantum of magnetic flux	T m^2 or Wb	230

Problems

Section 8.1

Problem 1. An electric dipole consists of charges $\pm q$ separated a distance b. Show that the torque $\boldsymbol{\tau}$ on an electric dipole $\mathbf{p}$ in a steady electric field $\mathbf{E}$ is given by $\boldsymbol{\tau} = \mathbf{p} \times \mathbf{E}$, where $\mathbf{p}$ has magnitude qb, pointing in the direction from $-q$ to $+q$.

Problem 2. Show that the units of $\mathbf{m}$, A m^2 or J T^{-1}, are equivalent.

Problem 3. Show that the units of μ_0, T m A^{-1}, are equivalent to Ω s m^{-1}.

Problem 4. It is possible that the Lorentz force law allows marine sharks, skates, and rays to orient in a magnetic field (Frankel 1984). If a shark can detect an electric field strength of 0.5 μV m^{-1}, how fast would it have to swim through the earth's magnetic field to experience an equivalent force on a charged particle? The earth's field is about 5×10^{-5} T.

Problem 5. The introduction to this section says that magnetism is a consequence of special relativity. Consider the following thought experiment. (a) A line of positive charge lies along the x axis and moves in the positive x direction. Is there a magnetic field present? (b) Change to a frame of reference moving with the charge. In this frame, where the charge is stationary, is there a magnetic field present? So is there really a magnetic field present, or not?

Problem 6. The speed of blood in an artery or vein can be measured using an electromagnetic blood flow meter. A blood vessel of radius R is oriented perpendicular to a magnetic field $\mathbf{B}$. Ions in the blood, which is moving with speed U, experience a Lorentz force. Positive ions move to one side of the vessel and negative ions move to the other side, establishing an electric field E whose force just balances the magnetic force.

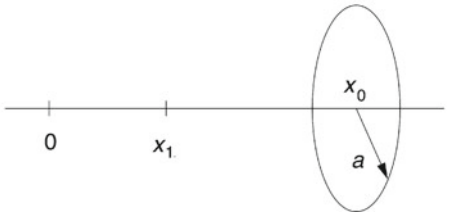
(a) Draw a diagram showing the vessel and the directions of **U**, **E**, and **B**.
(b) Find an expression for U in terms of E and B.
(c) The electric field can be approximated as a voltage v across the vessel divided by the width of the vessel. Find an expression for v in terms of U, B and R.
(d) If $B = 0.1\,\mathrm{T}$, $U = 0.01\,\mathrm{m\,s^{-1}}$ and $R = 1\,\mathrm{mm}$, what is v?

Section 8.2

Problem 7. A very long solenoid of radius a has current i in the windings. The windings are closely spaced and there are N turns per meter. What is the magnetic field in the solenoid? (Hint: if the solenoid is very long, the field inside is uniform and the field outside is zero. Use Ampere's law.)
Problem 8. Figure 8.8 uses the Biot Savart law to show that the magnetic field from a spherically symmetric radial distribution of current is zero. Use a simple symmetry argument to obtain the same result.
Problem 9. Show that $d\mathbf{D}/dt$ has the dimensions of current density.
Problem 10. A circular loop of radius a and area S carries current i. The loop is at the origin and lies in the xy plane. Calculate the magnetic field at any point on the z axis using the Biot–Savart law. Show that it is proportional to the magnetic moment of the loop, $|\mathbf{m}| = iS$, and falls off as z^{-3} if $z \gg a$.
Problem 11. Show that a point source of current in an infinite, homogeneous conducting medium discharges at such a rate that the displacement current density everywhere cancels the current density, so that Ampere's law also predicts that the magnetic field is zero.

Section 8.3

Problem 12. Derive Eq. 8.15 from Eq. 8.14.
Problem 13. The current along an axon is $i_i(x) = i_0,\ 0 < x < x_1$ and is zero everywhere else. The axon is in an infinite homogeneous conducting medium.
(a) What is $v_i(x)$?
(b) Find **B** at a point (x_0, y_0).
Problem 14. One can obtain a very different physical picture of the source of a magnetic field using the Biot–Savart law than one gets using Ampere's law, even though the field is the same. A ring of radius a is perpendicular to the x axis and centered at x_0. Current flows along the x axis from $x = 0$ to $x = x_1$. There is a spherically symmetric current in at $x = 0$ and a spherically symmetric current out from x_1. Calculate the magnetic field at a point on the ring using Ampere's law and using the Biot–Savart law. Discuss the difference in interpretation. Your expression for the field should be the same as Problem 13. A more extensive discussion of three different ways the source of the magnetic field can be viewed is given by Barach (1987).



Problem 15. Suppose that $i_i(t)$ is determined by measurement of the magnetic field around an axon. Numerical differentiation of the data gives derivatives of i_i also. Use the arguments of Sect. 6.11 and Problem 6.60 to show that for an action potential traveling without change of shape, one can determine the membrane current density from

$$j_m = \frac{1}{2\pi a u}\frac{\partial i_i}{\partial t} - c_m u r_i i_i .$$

For an application of this technique, see Barach et al. (1985).
Problem 16. Use Ampere's law to calculate the magnetic field produced by a nerve axon.
(a) First, solve Problem 30 of Chap. 7 to obtain the electrical potential inside (V_i) and outside (V_o) an axon. The solution will be in terms of the modified Bessel functions $I_0(kr)$ and $K_0(kr)$, where k is a spatial frequency and r is the radial distance from the center of the axon. Assume the axon has a radius a.
(b) Find the axial component of the current density both inside and outside the axon, $J_{iz} = -\sigma_i \partial V_i/\partial z$ and $J_{oz} = -\sigma_o \partial V_o/\partial z$, where σ_i and σ_o are the intracellular and extracellular conductivities (Eqs. 6.16b and 6.26).
(c) Integrate J_{iz} over the axon cross-section to get the total intracellular current. Then integrate J_{oz} over an annulus from a to radius r, to get the *return current.* You will need the following integrals:

$$\int x I_0(x)dx = x I_1(x)$$

and

$$\int x K_0(x)dx = -x K_1(x).$$

(d) Use Ampere's law (Eq. 8.11) to calculate the magnetic field. Take the line integral of Ampere's law as a closed loop of radius r concentric with the axon ($r > a$). The current enclosed by this loop is simply the sum of the intracellular and return currents calculated in (c).

Section 8.4

Problem 17. Use the same technique as in Chap. 7 to estimate the magnitude of the magnetocardiogram signal.

Problem 18.

(a) Derive Eqs. 8.19 and 8.20.

(b) What effect will y and z components of $\mathbf{p}$ have for measurements taken along an axis with $y = 0$?

Problem 19. Consider a two-dimensional sheet of cardiac tissue represented using the bidomain model (Sect. 7.9.3). The intracellular and extracellular conductivity tensors are given by

$$\widetilde{\sigma_i} = \begin{pmatrix} \sigma_{ixx} & \sigma_{ixy} \\ \sigma_{ixy} & \sigma_{iyy} \end{pmatrix} = \begin{pmatrix} \sigma_{iL}\cos^2\theta + \sigma_{iT}\sin^2\theta & (\sigma_{iL} - \sigma_{iT})\sin\theta\cos\theta \\ (\sigma_{iL} - \sigma_{iT})\sin\theta\cos\theta & \sigma_{iL}\sin^2\theta + \sigma_{iT}\cos^2\theta \end{pmatrix}$$

$$\widetilde{\sigma_o} = \begin{pmatrix} \sigma_{oxx} & \sigma_{oxy} \\ \sigma_{oxy} & \sigma_{oyy} \end{pmatrix} = \begin{pmatrix} \sigma_{oL}\cos^2\theta + \sigma_{oT}\sin^2\theta & (\sigma_{oL} - \sigma_{oT})\sin\theta\cos\theta \\ (\sigma_{oL} - \sigma_{oT})\sin\theta\cos\theta & \sigma_{oL}\sin^2\theta + \sigma_{oT}\cos^2\theta \end{pmatrix}$$

where "L" means parallel to the fibers, "T" means perpendicular to the fibers, and θ is the angle between the fiber direction and the x axis. The intracellular and extracellular current densities are given by $\mathbf{j}_i = \widetilde{\sigma_i} \cdot \mathbf{E} = -\widetilde{\sigma_i} \cdot \nabla v_i$ and $\mathbf{j}_o = -\widetilde{\sigma_o} \cdot \nabla v_o$. Assume that the intracellular and extracellular potentials are given by $v_i = \sigma_{oxx} v_m(x)/(\sigma_{ixx} + \sigma_{oxx})$ and $v_o = -\sigma_{ixx} v_m(x)/(\sigma_{ixx} + \sigma_{oxx})$, where $v_m(x)$ is the transmembrane potential and is a function only of x, corresponding to a plane wave front propagating in the x direction.

(a) Draw a picture showing the 2-D sheet of tissue, the x and y axes, the fiber direction, and the direction of propagation.

(b) Show that $j_x = j_{ix} + j_{ox}$ is identically zero.

(c) Derive an expression for $j_y = j_{iy} + j_{oy}$ in terms of σ_{iL}, σ_{iT}, σ_{oL}, σ_{oT}, θ, and v_m.

(d) Under what conditions is j_y identically zero?

(e) Describe qualitatively the magnetic field produced by a wave front in a sheet of cardiac tissue. For additional features of this model, see Roth and Woods (1999).

Section 8.5

Problem 20. Consider two cylindrical cells of radius 1 μm. One is an axon with an action potential lasting 1 ms and traveling at 1 m s^{-1} with a depolarization amplitude of 100 mV. The other is a dendrite with a postsynaptic potential depolarization of 10 mV. The conductivity within both cells is 1 S m^{-1}.

(a) Compare the magnetic field 5 cm away from the dendrite with depolarization only and the axon with a complete pulse.

(b) If the minimum magnetic field that can be detected is 100×10^{-15} T, how many dendrites must be simultaneously excited to detect the signal?

(c) Pyramidal cells in the cortex are aligned properly to generate this kind of signal. Assume the dendrite is 2 mm long. There are about 50,000 neurons per mm^3 in the cortex, of which 70 % are pyramidal cells. Find the volume of the smallest excited region that could be detected if all the pyramidal cells in the volume simultaneously had a postsynaptic depolarization of 10 mV.

Problem 21. The magnetic field $\mathbf{B}(\mathbf{r})$ produced by a current dipole $\mathbf{p}$ located at $\mathbf{r}_0$ in a spherical conductor is given by Sarvas (1987)

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi F^2}\left[F\,(\mathbf{p}\times\mathbf{r}_0) - (\mathbf{p}\times\mathbf{r}_0\cdot\mathbf{r})\nabla F\right],$$

where $\mathbf{a} = \mathbf{r} - \mathbf{r}_0$, $a = |\mathbf{a}|$, $r = |\mathbf{r}|$, $F = a(ra + r^2 - \mathbf{r}_0\cdot\mathbf{r})$,

$$\nabla F = \left(\frac{a^2}{r} + \frac{\mathbf{a}\cdot\mathbf{r}}{a} + 2a + 2r\right)\mathbf{r} - \left(a + 2r + \frac{\mathbf{a}\cdot\mathbf{r}}{a}\right)\mathbf{r}_0,$$

and both r and r_0 are measured from the center of the sphere.

(a) Show that if $\mathbf{p}$ is radial, $\mathbf{B} = 0$.

(b) Show that the equation for the radial component of $\mathbf{B}$ reduces to Eq. 8.17. Note that the radius of the sphere does not enter into these equations.

Section 8.6

Problem 22. Consider a rectangular current loop with one corner at $(0, 0, 0)$ and the diagonally opposite corner at $(dx, dy, 0)$, in a changing magnetic field that has components $(0, 0, dB/dt)$. Show that for this configuration the differential form of the Faraday induction law, Eq. 8.22, follows from Eq. 8.21.

Problem 23. Obtain the differential form of Ampere's circuital law, Eq. 8.24, from Eq. 8.13.

Problem 24. The differential form of Ampere's law, Eq. 8.24, provides a relationship between the current density $\mathbf{j}$ and the magnetic field $\mathbf{B}$ that allows you to measure biological current with magnetic resonance imaging (see, for example, Scott et al. (1991)). Suppose you use MRI and find the distribution of magnetic field to be

$$B_x = C(yz^2 - yx^2)$$
$$B_y = C(xz^2 - xy^2)$$
$$B_z = C4xyz$$

where C is a constant with the units of T m^{-3}. Determine the current density. Assume the current varies slowly enough that the displacement current can be neglected.

Problem 25. Write down in differential form (a) the Faraday induction law, (b) Ampere's law including the displacement current term, (c) Gauss's law, and (d) Eq. 8.9. (Ignore the effects of dielectrics or magnetic materials. That is, assume $\mathbf{D} = \varepsilon_0\mathbf{E}$ and $\mathbf{B} = \mu_0\mathbf{H}$.) These four equations together constitute *Maxwell's equations*. Together with the Lorentz force law (Eq. 8.2), Maxwell's equations summarize all of electricity and magnetism.

Problem 26. Consider a square loop of wire in the xy plane that is moving in the positive x direction. There is a static magnetic field with a z component that increases linearly with x. Special relativity implies that the physics should be the same in any inertial frame of reference: that is, the physics should be the same in a reference frame moving with a constant velocity as it is in a frame at rest.

(a) Consider the frame described above, in which the loop moves and the magnetic field is static. Show qualitatively that the Lorentz force on the electrons in the wire induces a current.

(b) Now consider the situation from a frame of reference moving with the loop. Show qualitatively that Faraday induction will induce a current in the wire.

Which "really" caused the current: the Lorentz force or Faraday induction?

Problem 27. Suppose one is measuring the EEG when a time-dependent magnetic field is present (such as during magnetic stimulation). The EEG is measured using a disk electrode of radius $a = 5$ mm and thickness $d = 1$ mm, made of silver with conductivity $\sigma = 63 \times 10^6$ S m^{-1}. The magnetic field is uniform in space, is in a direction perpendicular to the plane of the electrode, and changes from zero to 1 T in 200 μs.

(a) Calculate the electric field and current density in the electrode due to Faraday induction.

(b) The rate of conversion of electrical energy to thermal energy per unit volume (Joule heating) is the product of the current density times the electric field. Calculate the rate of thermal energy production during the time the magnetic field is changing.

(c) Determine the total thermal energy change caused by the change of magnetic field.

(d) The specific heat of silver is 240 J kg^{-1} °C^{-1}, and the density of silver is 10,500 kg m^{-3}. Determine the temperature increase of the electrode due to Joule heating.

The heating of metal electrodes can be a safety hazard during rapid (20 Hz) magnetic stimulation (Roth et al. 1992).

Problem 28. Suppose that during rapid-rate magnetic stimulation, each stimulus pulse causes the temperature of a metal EEG electrode to increase by ΔT (see Problem 27). The hot electrode then cools exponentially with a time constant τ (typically about 45 s). If N stimulation pulses are delivered starting at $t = 0$ with successive pulses separated by a time Δt, then the temperature at the end of the pulse train is $T(N, \Delta t) = \Delta T \sum_{i=0}^{N-1} e^{-i\Delta t/\tau}$. Find a closed form expression for $T(N, \Delta t)$ using the summation formula for the geometric series: $1+x+x^2+\cdots+x^{n-1} = (1-x^n)/(1-x)$. Determine the limiting values of $T(N, \Delta t)$for $N\Delta t \ll \tau$ and $N\Delta t \gg \tau$. (See Roth et al. 1992.)

Problem 29. The concept of *skin depth* plays a role in some biomagnetic applications.

(a) Write Ampere's law (Eq. 8.24) for the case when the displacement current is negligible.

(b) Use Ohm's law (Eq. 6.26) to write the result from (a) in terms of the electric field.

(c) Take the curl of both sides of the equation you found in (b) (Assume the conductivity σ is homogeneous and isotropic).

(d) Use Faraday's law (Eq. 8.22), $\nabla \cdot \mathbf{B} = 0$ (Eq. 8.9), and the vector identity $\nabla \times (\nabla \times \mathbf{B}) = \nabla(\nabla \cdot \mathbf{B}) - \nabla^2\mathbf{B}$ to simplify the result from (c).

(e) Your answer to (d) should be the familiar diffusion equation (Eq. 4.24). Express the diffusion constant D in terms of electric and magnetic parameters.

(f) In Chap. 4, we found that diffusion over a distance L takes a time $T = L^2/2D$. During transcranial magnetic stimulation, $L = 0.1$ m, $\sigma = 0.1$ S $m^{-1}$and $\mu_0 = 4\pi \times 10^{-7}$ T m A^{-1}. How long does the magnetic field take to diffuse into the head? Is this time much longer than or much shorter than the rise time of the magnetic field for the stimulator designed by Barker et al. (1985)?

(g) Solve $T = L^2/2D$ for L, using the expression for D found in (e). Calculate L for $T = 0.1$ ms. Is L much larger than or much smaller than the size of your head? L is closely related to the skin depth defined in electromagnetic theory.

(h) During magnetic resonance imaging (see Chap. 18), an 85-MHz radio-frequency magnetic field is applied to the body. Calculate L using half a period for T. How does L compare to the size of the head? The frequency of the RF field is proportional to the strength of the static magnetic field in an MRI device, and 85 MHz corresponds to 2 T. If the static field is 7 T (common in modern high-field MRI), calculate L. Is it safe to ignore skin depth during high-field MRI?

Section 8.7

Problem 30. Suppose that a magnetic stimulator consists of a single-turn coil of radius $a = 2$ cm. It is desired to have a magnetic field of 2 T on the axis of the coil at a distance $b = 2$ cm away.

(a) Calculate the current required, using symmetry and the Biot–Savart law. (Hint: Use the results of Problem 10.)
(b) Assume that the magnetic field rises from 0 to 2 T in 100 μs. Assume also that the flux through the coil is equal to the field at the center of the coil multiplied by the area of the coil. Calculate the emf induced in the coil.

Problem 31. Assume a sheet of tissue having conductivity σ is placed perpendicular to a uniform, strong, static magnetic field $\mathbf{B}_0$. A weaker but temporally oscillating magnetic field $\mathbf{B}_1(t)$ is parallel to $\mathbf{B}_0$ and is uniform in the region $r < a$, where r is the distance from a line along the direction of $\mathbf{B}_0$.
(a) Derive an expression for the electric field $\mathbf{E}$ induced by the oscillating magnetic field. It will depend on the distance r from the center of the sheet and the rate of change of the magnetic field.
(b) Determine an expression for the current density $\mathbf{j}$ by multiplying the electric field by the conductivity.
(c) The force per unit volume, $\mathbf{F}$, is given by the Lorentz force, $\mathbf{j} \times \mathbf{B}_0$ (ignore the weak $\mathbf{B}_1$). Find an expression for $\mathbf{F}$.
(d) The source of the ultrasonic pressure waves can be expressed as the divergence of the Lorentz Force. Derive an expression for $\nabla \cdot \mathbf{F}$.
(e) Draw a picture showing the directions of $\mathbf{j}$, $\mathbf{B}_0$, and $\mathbf{F}$.

This technique of measuring the ultrasonic signal and determining the conductivity is called *Magnetoacoustic Tomography with Magnetic Induction (MAT-MI)* (Xu and He 2005).

Problem 32.
(a) Rederive the cable equation for the transmembrane potential v, (Eq. 6.55) using one crucial modification: generalize Eq. 6.26 to account for part of the intracellular electric field that arises from Faraday induction and therefore cannot be written as the gradient of a potential,

$$i_i(x) = -\frac{1}{r_i}\left(\frac{dv_i}{dx} + E_{ix}\right).$$

Assume you measure v relative to the resting potential so Eq. 6.53 becomes $j_m = g_m v$, and let the extracellular potential be small so $v_i = v$. Identify the new source term in the cable equation (the *activating function* for magnetic stimulation), analogous to v_r in Eq. 6.55.
(b) Let

$$E_i = E_0 \frac{a^2}{x^2 + a^2}.$$

Calculate the activating function and plot both the electric field and the activating function versus x.
(c) Suppose you stimulate a nerve using this activating function, first with one polarity of the current pulse and then the other. What additional delay in the response of the nerve (as measured by the arrival time of the action potential at the far end) will changing polarity cause because of the extra distance the action potential must travel? Assume $a = 4$ cm and the conduction speed is 60 m s^{-1}.

Section 8.8

Problem 33. Magnetite, Fe_3O_4, has a density of 5.24 g cm^{-3} and a magnetic moment of 3.75×10^{-23} A m^2 per molecule. If a cubic sample 50 nm on a side is completely magnetized, what is the total magnetic moment? What is the magnitude of M?

Problem 34. The magnetic moment of a magnetosome, one of the small particles of magnetite in a bacterium, is about 6.40×10^{-17} A m^2. Assume that the magnetic activity in all the species listed is due to a collection of magnetosomes of this size. The table shows values given in the references cited in the text. The earth's magnetic field is about 5×10^{-5} T. Fill in the remaining entries in the table.

Organism	Number of magnetosomes	Total magnetic moment (A m^2)	mB_{earth}/k_BT
Bacterium	20		
Bee		1.2×10^{-9}	
Pigeon		5.0×10^{-9}	
Tuna	8.5×10^7		

Problem 35. In this problem you will work out the orientation of a bacterium if the entire organism simply aligns like a compass needle in the earth's field of 5×10^{-5} T.
(a) Show that $\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}$ implies an orientation energy $U = -mB\cos\theta$.
(b) The bacterium has a single flagellum that causes it to swim in the direction of its long axis with speed v_0. The component of its velocity in the direction of the earth's field is $v_x = v_0 \cos\theta$. In the absence of the magnetic torque, the probability that a bacterium is at an angle between θ and $\theta + d\theta$ with the earth's field is proportional to $d\Omega = 2\pi \sin\theta\, d\theta$. With the torque, the probability that a bacterium is at angle with the earth's field is modified by a Boltzmann factor $\exp(-U/k_BT)$. Find the average velocity in the direction of the earth's field. Use $m = 1.28 \times 10^{-15}$ A m^2.

Problem 36. Suppose that the bacterium of Problem 35 is swimming in a tank aligned with the earth's field. An external coil suddenly reverses the direction of the field but leaves the magnitude unchanged. Assume that the bacterium is a sphere of radius a. A torque on a small sphere of radius a in a medium of viscosity η causes the sphere to rotate at a rate $d\theta/dt$, such that $\tau = 8\pi a^3 \eta (d\theta/dt)$. For simplicity, assume that all motion takes place in a plane.

(a) Show that $d\theta/dt = \sin\theta/t_0$, where $t_0 = 8\pi a^3\eta/mB$.
(b) Evaluate t_0 for a bacterium of radius 2 μm in the earth's magnetic field. Use $m = 1.28 \times 10^{-15}$ A m^2
(c) The velocity component perpendicular to the field is $v_y = v_0 \sin\theta$. Show that when the bacterium rotates from angle θ_1 to θ_2 it has moved a distance $y = v_0 t_0(\theta_2 - \theta_1)$.
(d) Show that the time required to change from angle ε to $\pi - \varepsilon$ is $t_0 \ln\left[(1+\cos\varepsilon)/(1-\cos\varepsilon)\right]$.

Problem 37. Magnetic cell sorting is a way to isolate cells of a particular type. Small *superparamagnetic particles* (about 50 nm diameter) are bound to an antibody that attaches specifically to the cell type of interest. (Superparamagnetic means that they behave linearly but have a magnetic susceptibility $\chi_m \gg 1$.) These cells are then placed in a magnetic field gradient, and the resulting force is used to manipulate the cell. What is the force if 100 spherical 50-nm diameter particles are attached to a cell that is in a magnetic field of 1 T with a magnetic field gradient of 10 T m^{-1}?

Section 8.9

Problem 38. The spatial gradient in the earth's field is about 10^{-11} T m^{-1}. How much lateral movement can be tolerated in measuring a magnetoencephalogram of about 10^{-13} T?

Problem 39. Show that the units of $h/2e$ are V s, and that this is also a unit of magnetic flux.

Problem 40. Suppose that a SQUID of area 0.1 cm^2 can resolve a magnetic flux change $\Delta\Phi = 10^{-3}\Phi_0$. What is the corresponding change in B?

Problem 41. The first difference of B is $B(x+a) - B(x)$. What is the second difference? Compare the first and second differences to what is detected by a first-order and second-order gradiometer. Assume that B is constant over the area of each gradiometer loop. Use these results to determine the signal resulting from a distant but unwanted dipole source with a magnetic field that falls as $1/r^3$.

Problem 42. A first-order gradiometer is used to measure the magnetic field at a point $(x_0, 0, z_0)$ from a current dipole described by Eq. 8.18. The gradient is measured at position $z = x_0/\sqrt{2}$. The coils are at x_0 and $x_0 + a$ and are perpendicular to the z axis. Find the net flux in the gradiometer in terms of x_0 and a and the radius b of the coils. Assume B is uniform across each coil.

Problem 43. Figure 8.29a shows a gradiometer for measuring $\partial B_z/\partial z$. Sketch a gradient coil for measuring $\partial B_z/\partial x$.

Problem 44. Consider a nerve threaded through the center of a toroid of magnetic permeability μ, wound with N turns of wire, as shown in Fig. 8.26. The inner radius of the toroid is c, the outer radius is d, and the width is e. Assume a current I flows inside the axon and is uniform along its length.

(a) Calculate the magnetic flux $\Phi = \int \mathbf{B}\cdot d\mathbf{S}$ through the toroid winding caused by the current in the axon.
(b) The magnetic flux divided by the current is called the *mutual inductance*, M, of the axon and coil. Show that the mutual inductance is $M = \mu N e \ln(d/c)/2\pi$.
(c) By the Faraday Induction Law, the electromotive force (EMF) induced in the windings is $EMF = -M(dI/dt)$. Calculate the EMF when $\mu = 10\,000\mu_0$, $N = 100$, $c = 1$ mm, $d = 2$ mm, $e = 1$ mm, and I changes from zero to 1 μA in 1 ms. For additional information about using a toroid to detect currents along an axon, see Gielen et al. (1986).

Problem 45. A coil on a magnetic toroid as in Problem 44 is being used to measure the magnetic field of a nerve axon.
(a) If the axon is suspended in air, with only a thin layer of extracellular fluid clinging to its surface, use Ampere's law to determine the magnetic field, B, recorded by the toroid.
(b) If the axon is immersed in a large conductor such as a saline bath, B is proportional to the sum of the intracellular current plus that fraction of the extracellular current that passes through the toroid (see Problem 14). Suppose that during an experiment an air bubble is trapped between the axon and the inner radius of the toroid. How is the magnetic signal affected by the bubble? See Roth et al. (1985).

Problem 46. When comparing calculated and measured magnetic fields, the calculated field should be integrated over the area of the detector coil to give the magnetic flux through the coil. Assume the detector coil is circular with radius a. The flux can be approximated by

$$\iint \mathbf{B}\cdot d\mathbf{S} \approx \frac{\pi a^2}{3}\sum_{i=1}^{3} B_n\left(r = \frac{a}{\sqrt{2}}, \theta = i\frac{2\pi}{3}\right),$$

where B_n is the component of $\mathbf{B}$ normal to the coil, and r and θ are polar coordinates with the origin at the coil center. Show that this equation is exact up to second order. In other words, show that this equation is exact for magnetic fields given by

$$B_n = c + dx + ey + fx^2 + gxy + hy^2,$$

where c, d, e, f, g, and h are constants, $x = r\cos\theta$, and $y = r\sin\theta$. Higher order formulas for averaging the magnetic field can be found in Roth and Sato (1992).

References

Able KP, Able MA (1995) Interactions in the flexible orientation system of a migratory bird. Nature 375:230–232

Barach JP (1987) The effect of ohmic return currents on biomagnetic fields. J Theor Biol 125:187–191
Barach JP, Roth BJ, Wikswo JP (1985) Magnetic measurements of action currents in a single nerve axon: a core conductor model. IEEE Trans Biomed Eng 32:136–140
Barker AT, Jalinous R, Freeston IL (1985) Non-invasive magnetic stimulation of the human cortex. Lancet 1(8437):1106–1107
Clarke J (1994) SQUIDs. Sci Am (Aug 1994):46–53
Cochran WW, Mouritsen H, Wikelski M (2004) Migrating songbirds recalibrate their magnetic compass daily from twilight cues. Science 304:405–408
Cohen D, Nemoto I, Kaufman L, Arai S (1984) Ferrimagnetic particles in the lung part II: the relaxation process. IEEE Trans Biomed Eng 31:274–285
de Araujo FF, Pires MA, Frankel RB, Bicudo CEM (1986) Magnetite and magnetotaxis in algae. Biophys J 50:375–378
Eder SHK, Cadiou H, Muhamad A, McNaughton PA, Kirschvink JL, Winklhofer M (2012) Magnetic characterization of isolated candidate vertebrate magnetoreceptor cells. PNAS 109:12022–12027
Eisberg R, Resnick R (1985) Quantum physics of atoms, molecules, solids, nuclei and particles, 2nd ed. Wiley, New York
Finegold L (2012) Resource letter BSSMF-1: biological sensing of static magnetic fields. Am J Phys 80:851–861
Frankel RB (1984) Magnetic guidance of organisms. Ann Rev Biophys Bioeng 13:85–103
Frankel RB, Bazylinski DA (1994) Magnetotaxis and magnetic particles in bacteria. Hyperfine Interact 90:135–142
Frankel RB, Blakemore RP, Wolfe RS (1979) Magnetite in freshwater magnetotactic bacteria. Science 203:1355–1356
Freitas C, Mondragon-Llorca H, Pascual-Leone A (2011) Noninvasive brain stimulation in Alzheimer's disease: systematic review and perspectives for the future. Exp Gerontol 46(8):611–627
Gielen FLH, Roth BJ, Wikswo JP Jr (1986) Capabilities of a toroid-amplifier system for magnetic measurement of current in biological tissue. IEEE Trans Biomed Eng 33:910–921
Gielen FLH, Friedman RN, Wikswo JP Jr (1991) In vivo magnetic and electric recordings from nerve bundles and single motor units in mammalian skeletal muscle. Correlations with muscle force. J Gen Physiol 98(5):1043–1061
Gould JL (1995). Constant compass calibration. Nature 375:184
Gould JL, Kirschvink JL, Deffeyes KS (1978) Bees have magnetic remanence. Science 201:1026–1028
Griffiths DJ (2013) Introduction to electrodynamics, 4th ed. Addison-Wesley, Boston
Hallett M, Cohen LG (1989) Magnetism: a new method for stimulation of nerve and brain. JAMA 262:538–541
Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65(2):413–497
Hari R, Salmelin R (2012) Magnetoencephalography: from SQUIDs to neuroscience. Neuroimage 20th Anniversary Special Edition. Neuroimage 61:386–396
Hosaka H, Cohen D, Cuffin BN, Horacek BM (1976) The effect of torso boundaries on the magnetocardiogram. J Electrocardiol 9:418–425
Ilmoniemi RJ, Ruohonen J, Karhu J (1999) Transcranial magnetic stimulation—A new tool for functional imaging of the brain. Crit Rev Biomed Eng 27(3–5):241–284
Jordan A, Scholz R, Wust P, Fahline H, Roland F (1999) Magnetic fluid hyperthermia (MFH): cancer treatment with AC magnetic field induced ecitation of biocompatible superparamagnetic nanoparticles. J Magn Magn Mater 201:413–419
Mielczarek EV, McGrayne SB (2000) Iron, nature's universal element: why people need iron and animals make magnets. Rutgers U. Press, New Brunswick
Moskowitz BM (1995) Biomineralization of magnetic minerals. Rev Geophys 33(Part 1 Suppl. S):123–128
Nielsen P, Fischer R, Englehardt R, Tondury P, Gabbe EE, Janka GE (1995) Liver iron stores in patients with secondary haemosiderosis under iron chelation therapy with deferoxamine or deferiprone. Brit J Hematol 91:827–833
O'Reardon JP, Solvason HB, Janicak PG, Sampson S, Eisinberg KE, Nahas Z, McDonald WM, Avery D, Fitzgerald PB, Loo C, Demitrack MA, George MS, Sackheim HA (2007) Efficacy and safety of transcranial magnetic stimulation in the acute treatment of major depression. A multisite randomized controlled trial. Biol Psychiatr 62(11):1208–1216
Page CH (1977) Electromotive force, potential difference, and voltage. Am J Phys 45:978–980
Pankhurst QA, Thanh NTK, Jones SK, Dobson J (2009) Progress in applications of magnetic nanoparticles in biomedicine. J Phys D Appl Phys 42:224041
Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd ed. reprinted with corrections, 1995. Cambridge University Press, New York
Purcell EM, Morin DJ (2013) Electricity and magnetism, 3rd ed. Cambridge University Press, New York
Romer RH (1982) What do "voltmeters" measure?: Faraday's law in a multiply connected region. Am J Phys 50:1089–1093
Rossi S, Hallett M, Rossini PM, Pascual-Leone A, The safety of the TMS Consensus Group (2009) Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. Clin Neurophysiol 120:2008–2039
Roth BJ (1994) Mechanisms for electrical stimulation of excitable tissue. Crit Rev Biomed Eng 22(3/4):253–305
Roth BJ, Sato S (1992) Accurate and efficient formulas for averaging the magnetic field over a circular coil. In Hoke M, Erne SN, Okada TC, Romani GL (eds) Biomagnetism: clinical aspects. Elsevier, Amsterdam
Roth BJ, Wikswo JP Jr (1985) The magnetic field of a single axon: a comparison of theory and experiment. Biophys J 48:93–109
Roth BJ, Woosley JK, Wikswo JP Jr (1985) An experimental and theoretical analysis of the magnetic field of a single axon. In Weinberg H, Stroink G, Katila T (eds) Biomagnetism: applications and theory. Pergamon, New York, pp 78–82
Roth BJ, Pascual-Leone A, Cohen LG, Hallett M (1992) The heating of metal electrodes during rapid-rate magnetic stimulation: a possible safety hazard. Electroencephalogr Clin Neurophysiol 85:116–123
Sarvas J (1987) Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys Med Biol 32:11–22
Scott GC, Joy MLG, Armstrong RL, Henkelman RM (1991) Measurement of nonuniform current density by magnetic resonance. IEEE Trans Med Imag 10:362–374
Shadowitz A (1975) The electromagnetic field. McGraw-Hill, New York
Stahlhofen W, Moller W (1993) Behaviour of magnetic micro-particles in the human lung. Rad Env Biophys 32(3):221–238
Staton DJ, Friedman RN, Wikswo JP Jr (1993) High resolution SQUID imaging of octupolar currents in anisotropic cardiac tissue. IEEE Trans Appl Superconduct 3(1):1934–1936
Strasburger JF, Cheulkar B, Wakai RT (2008) Magnetocardiography for fetal arrhythmias. Heart Rhythm 5:1073–1076
Stroink G (1985) Magnetic measurements to determine dust loads and clearance rates in industrial workers and miners. Med Biol Eng Comput 23:44–49
Stroink G (1992) Cardiomagnetic imaging. In: Zaret BL et al (eds) Frontiers in cardiovascular imaging. Raven Press, New York, pp 161–177

Stroink G, Blackford B, Brown B, Horacek BM (1981) Aluminum shielded room for biomagnetic measurements. Rev Sci Instrum 52(3):463–468

Swinney KR, Wikswo JP Jr (1980) A calculation of the magnetic field of a nerve action potential. Biophys J 32:719–732

Tan GA, Brauer F, Stroink G, Purcell CJ (1992) The effect of measurement conditions on MCG inverse solutions. IEEE Trans Biomed Eng 39(9):921–927

Trontelj Z, Zorec R, Jazbinsek V, Erne SN (1994) Magnetic detection of a single action potential in *Chara corallina* internodal cells. Biophys J 66:1694–1696

Walcott C, Gould JL, Kirschvink JL (1979) Pigeons have magnets. Science 205:1027–1029

Walker MM, Kirschvink JL, Chang S-BR, Dizon AE (1984) A candidate magnetic sense organ in the yellowfin tuna, *Thunnus albacares*. Science 224:751–753

Wassermann E, Epstein C, Ziemann U (2008) Oxford Handbook of transcranial stimulation. Oxford University Press, Oxford

Wijesinghe R (2010) Magnetic measurements of peripheral nerve function using a neuromagnetic current probe. Exp Biol Med 235:159–169

Wikswo JP Jr (1980) Noninvasive magnetic detection of cardiac mechanical activity: theory. Med Phys 7:297–306

Wikswo JP Jr (1995a) SQUID magnetometers for biomagnetism and nondestructive testing: important questions and initial answers. IEEE Trans Appl Superconduct 5(2):74–120

Wikswo JP Jr (1995b) Tissue anisotropy, the cardiac bidomain, and the virtual cathode effect. In Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd ed. Saunders, Philadelphia, pp 348–361

Wikswo JP Jr, Gevins A, Williamson SJ (1993) The future of the EEG and MEG. Electroencephalogr Clin Neurophysiol 87:1–9

Woosley JK, Roth BJ, Wikswo JP Jr (1985) The magnetic field of a single axon; A volume conductor model. Math Biosci 76:1–36

Xu Y, He B (2005) Magnetoacoustic tomography with magnetic induction (MAT-MI). Phys Med Biol 50:5175–5187

Electricity and Magnetism at the Cellular Level

9

This chapter describes a number of topics related to charged membranes and the movement of ions through them. Topics range from the basics of how the presence of impermeant ions alters the concentration ratios of permeant ions, to the movement of ions under the combined influence of an electric field and diffusion, and to simple models for gating in ion channels in cell membranes. It also discusses mechanisms for the detection of weak electric and magnetic fields and the possible effects of weak low-frequency electric and magnetic fields on cells.

Section 9.1 discusses Donnan equilibrium, in which the presence of an impermeant ion on one side of a membrane, along with other ions that can pass through, causes a potential difference to build up across the membrane. This potential difference exists even though the bulk solution on each side of the membrane is electrically neutral. Section 9.2 examines the Gouy–Chapman model for the charge buildup at each surface of the membrane that gives rise to this potential difference. This same model is extended in three dimensions to the cloud of counterions surrounding each ion in solution—the Debye–Hückel model of Sect. 9.3.

Since water molecules have a net dipole moment, they align themselves so as to nearly cancel the electric field of each ion. Very close to the ion, the electric field is so strong that even complete alignment is insufficient to cancel the ion's field. This saturation of the dielectric is described in Sect. 9.4.

Ions move in solution by diffusion if there is a concentration gradient and by drift if there is an applied electric field. The Nernst–Planck equation (Sect. 9.5) describes this motion. When several ion species are moving through a membrane, there can be zero total electric current, even though there is a flow of each species. A constant-field model for this situation leads to the Goldman equations of Sect. 9.6.

The next two sections discuss channels in active cell membranes. Section 9.7 describes a simple model for gating—the opening and closing of channels—as well as limitations to the conductance of each channel imposed by diffusion to the mouth of the channel. Section 9.8 introduces noise—the fluctuations in channel current that limit measurement accuracy but also can be used to determine properties of the channels.

Section 9.9 shows how channels can detect very small mechanical motions, as in the ear, and how certain fish can detect very small electric fields in sea water. Both of these processes are working near the limit of sensitivity set by random thermal motion.

Section 9.10 introduces an area of great interest and controversy: whether weak, low-frequency electric and magnetic fields can have any effect on cells. We discuss some of the physical aspects of the problem and conclude that such effects are highly unlikely.

There are many similarities between the models for biological physics presented in this chapter and the models used in plasma physics (Uehara et al. 2000).

9.1 Donnan Equilibrium

There is usually an electrical potential difference across the wall of a capillary. There is also a potential difference across the cell membrane (or plasma membrane or cytoplasmic membrane), and the concentration of certain ion species is different in the intracellular and extracellular fluid. In Chap. 3, we saw that if the potential difference across the membrane is $v' - v$, an ion of valence z is in equilibrium when $C'/C = e^{-ze(v'-v)/k_BT}$. For this concentration ratio, there is no current, even if the membrane is permeable to the species. This result is a special case of the Boltzmann factor, more familiar in physiology as the Nernst equation (Eq. 3.34):

$$v' - v = -\frac{k_BT}{ze} \ln\left(\frac{C'}{C}\right) = -\frac{RT}{zF} \ln\left(\frac{C'}{C}\right).$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_9, © Springer International Publishing Switzerland 2015

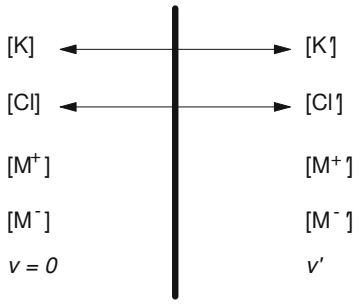


Fig. 9.1 Ion concentrations on either side of a membrane. Species that can pass through the membrane are indicated by double-headed arrows

It is often said—incorrectly—that the Nernst equation shows how the concentration of an ion species causes the potential difference across the membrane. We saw in Chap. 6 that the potential difference across the membrane is caused by layers of charge on each side of the membrane that create an electric field in the membrane. The solutions on each side of the membrane are electrically neutral except at the boundary with the membrane. (If there was an electric field in the solution, ions would move until the field was zero; then Gauss's law could be used to show that any volume contains zero charge.) We will learn in Sect. 9.2 the typical distance from the membrane occupied by the charged layer, and in Sect. 9.3, we will find the distance scale over which there are microscopic departures from neutrality in a bulk ionic solution.

The concentration differences do not *directly* cause the potential difference. However, if the concentration of an ion species on one side of the membrane is varied, the potential often changes in a manner that is approximated by the Nernst equation over a wide range of concentrations. We will now explore one mechanism by which this can happen. This is particularly important for the walls of capillaries, where charged proteins in the blood are too large to pass through the gaps between cells in the capillary walls, but it is also applicable to the cell membrane.

In Donnan equilibrium, the potential difference arises because one ion species cannot pass through the membrane at all. Consider the hypothetical case of Fig. 9.1. Permeant potassium ions exist on either side of the membrane in concentrations [K] and $[K']$. In this case, potassium is the only permeant cation; in a real situation, there might be several permeant ions. The membrane is also permeable to chloride ions, which exist in concentrations [Cl] and $[Cl']$. Chloride is the only permeant anion. In addition, there are large charged molecules $[M^+]$ and $[M^-]$ that cannot pass through the membrane. Their concentrations are $[M^+]$, $[M^{+\prime}]$, $[M^-]$, and $[M^{-\prime}]$. For simplicity, we assume they are monovalent. The potential on the left is 0; on the right, it is v'. Assume that the concentrations of the large molecules are fixed. The potassium concentration on the left side of the membrane will be assumed known, and we must solve for four variables: $[K']$, [Cl], $[Cl']$, and v'. Therefore, four equations are needed.

The first two equations state that the solutions on either side are electrically neutral:

$$[M^+] + [K] = [Cl] + [M^-], \tag{9.1}$$

$$[M^{+\prime}] + [K'] = [Cl'] + [M^{-\prime}]. \tag{9.2}$$

Equation 9.1 can be solved for [Cl]. It will be convenient to define $[M] = [M^+] - [M^-]$ and $[M'] = [M^{+\prime}] - [M^{-\prime}]$:

$$[Cl] = [K] + ([M^+] - [M^-]) = [K] + [M]. \tag{9.3}$$

Note that adding any amount of KCl to the solution on the left automatically satisfies this equation, since any increase in [K] is accompanied by the same increase in [Cl].

The other two equations state that the concentrations of potassium and chloride on the two sides of the membrane are related by a Boltzmann factor. Since the valence $z = +1$ for [K] and -1 for [Cl], we have

$$\frac{[K']}{[K]} = \frac{[Cl]}{[Cl']} = e^{-ev'/k_BT}. \tag{9.4}$$

The chloride concentration on the right is $[Cl'] = [Cl]\left([Cl']/[Cl]\right) = [Cl]\left([K]/[K']\right)$, so that from Eq. 9.2 $[K'] + [M'] = [Cl]\left([K]/[K']\right)$. This can be rewritten as a quadratic equation in $[K']$, since [K] and $[M']$ are known and [Cl] is calculated from Eq. 9.3:

$$[K']^2 + [M'][K'] - [K][Cl] = 0.$$

The solution is

$$[K'] = \frac{-[M'] + \sqrt{[M']^2 + 4[K][Cl]}}{2}. \tag{9.5}$$

(The negative square root is discarded because it would give a negative potassium concentration.) Once we have solved for $[K']$, $[Cl']$ and v' are determined from Eq. 9.4. Solutions for different values of [K] are shown in Table 9.1 and Figs. 9.2 and 9.3 for the conditions

$$[M^+] = 145\,\text{mmol}\,\text{l}^{-1}, \quad [M^{+\prime}] = 15\,\text{mmol}\,\text{l}^{-1},$$

$$[M^-] = 30\,\text{mmol}\,\text{l}^{-1}, \quad [M^{-\prime}] = 156\,\text{mmol}\,\text{l}^{-1},$$

$$[M] = 115\,\text{mmol}\,\text{l}^{-1}, \quad [M'] = -141\,\text{mmol}\,\text{l}^{-1}.$$

The temperature $T = 310$ K, for which $k_BT/e = 26.75$ mV.

Table 9.1 Variation of concentrations (mmol l^{-1}) and voltage (mV) as [K] is varied

[K]	[Cl]	[K′]	[Cl′]	[Cl]/[Cl′] = [K′]/[K]	v'
0.01	115.01	141.01	0.00816	14101	−255.57
0.10	115.10	141.08	0.08	1410.8	−193.99
0.20	115.20	141.16	0.16	705.8	−175.46
0.50	115.50	141.41	0.41	282.8	−151.00
1.00	116.00	141.82	0.82	141.8	−132.53
2.00	117.00	142.64	1.64	71.32	−114.15
5.00	120.00	145.13	4.13	29.03	−90.10
10.00	125.00	149.37	8.37	14.94	−72.33
20.00	135.00	158.08	17.08	7.904	−55.30
50.00	165.00	185.48	44.48	3.710	−35.07
100.00	215.00	233.20	92.20	2.332	−22.65
200.00	315.00	331.21	190.21	1.656	−13.49
500.00	615.00	629.49	488.49	1.259	−6.16

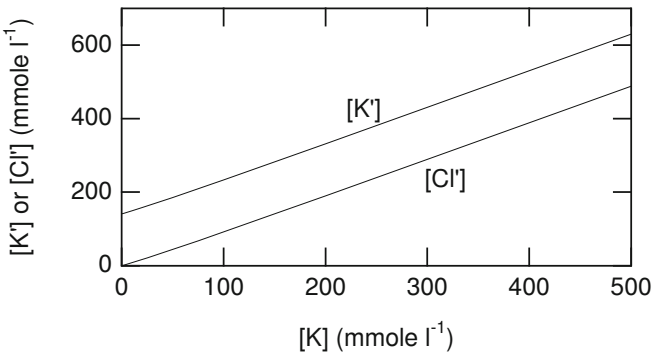


Fig. 9.2 Variation of [K′] and [Cl′] with [K] in the example of Donnan equilibrium

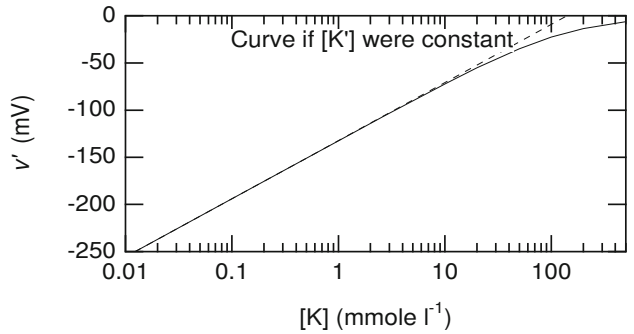


Fig. 9.3 Membrane potential v' vs. [K] for the example of Donnan equilibrium. For [K] $<$ 10 mM the curve is like the Nernst equation because [K′] has a nearly constant value of 141 mM. The dashed line shows the relationship if [K′] were constant

Several features of this solution are worth noting. First, changing [K] does change the potential, but the mechanism is indirect. The Boltzmann factor still applies; minuscule changes in concentration are sufficient to provide layers of charge on the membrane surface that generate a potential difference such that these concentrations are at equilibrium. Table 9.1 shows that [K] can vary by three orders of magnitude—from 0.01 to 10, and $[K']$ changes very little. Therefore, the curve of v' vs. ln [K] in Fig. 9.3 is nearly a straight line. The dashed line in Fig. 9.3 shows v' vs. ln [K] if $[K']$ is held constant. We could equally well have regarded [Cl] as the independent variable.

The impermeable ions enter the equation only as their net charge, $[M] = [M^+] - [M^-]$ and $[M'] = [M^{+\prime}] - [M^{-\prime}]$. As the concentrations [K] and [Cl] get larger, the impermeant ions become less important, the potential approaches zero, and the ratios $[K']$ / [K] and $[Cl']$ / [Cl] approach unity.

Donnan equilibrium may well explain the potential that exists across the capillary wall, which is impermeable to negatively charged proteins but is permeable to other ions. There is evidence that it does not adequately explain the potential across a cell membrane. For example, the membrane is known to be slightly permeable to sodium, although the sodium concentration is nowhere near what it would be if the sodium were in equilibrium.

9.2 Potential Change at an Interface: The Gouy–Chapman Model

In this section, we study one model for how ions are distributed at the interface in Donnan equilibrium. The model was used independently by Gouy and Chapman to study the interface between a metal electrode and an ionic solution. They investigated the potential changes along the x-axis perpendicular to a large plane electrode. The same model is used to study the charge distribution in a semiconductor. Biological applications are described by Mauro (1962). We show the features of the model by examining the transition region for the Donnan equilibrium example described in the preceding section.

An infinitely thin membrane at $x = 0$ is assumed to be permeable to potassium and chloride ions. Their concentrations are $K(x)$ and $Cl(x)$. An impermeant positive cation has concentration $M(x)$ for $x > 0$. For negative x, $M(x) = 0$. There are no impermeant anions. Far to the left, the potential is zero and the concentrations are [K] and [Cl]. Far to the right, they are v', $[K']$, $[Cl']$, and $[M']$.

The first step is to relate the charge distribution to the potential. If v and E change only in the x direction, then Gauss's law can be applied to a slab of cross-sectional area S between x and $x + dx$ as shown in Fig. 9.4. The net flux out through the surface at $x + dx$ is $E_x(x + dx)S$. The net outward flux at x is $-E_x(x)S$. There is no contribution to the flux through the other surfaces. The total ionic charge in the volume is $\rho_{\text{ext}}(x)Sdx$. We include the effect of water polarization by using the dielectric constant for water, which

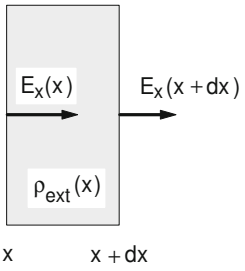


Fig. 9.4 Gauss's law is applied to the shaded volume to derive Poisson's equation in one dimension

is about $\kappa = 80$. Applying Gauss's law in the form of Eq. 6.21b, we obtain[1]

$$E_x(x+dx) - E_x(x) = \frac{4\pi\rho_{\text{ext}}(x)\,dx}{4\pi\epsilon_0\kappa},$$

$$\frac{dE_x}{dx} = \frac{4\pi\rho_{\text{ext}}(x)}{4\pi\epsilon_0\kappa}.$$

Finally, since $E_x = -\partial v/\partial x$, we have the one-dimensional Poisson equation,

$$\frac{d^2v}{dx^2} = -\frac{4\pi\rho_{\text{ext}}(x)}{4\pi\epsilon_0\kappa}. \tag{9.6}$$

This equation was derived in much the same way that the equation of continuity was combined with Fick's first law to derive Fick's second law (Sect. 4.8). The same procedure can be used in three dimensions to derive the general form of Poisson's equation:

$$\nabla^2 v = -\frac{4\pi\rho_{\text{ext}}(\mathbf{r})}{4\pi\epsilon_0\kappa}. \tag{9.7}$$

For the model being considered the ions are all univalent, so the ionic charge density at x is related to the concentrations by

$$\rho_{\text{ext}}(x) = e\,[K(x) + M(x) - Cl(x)]\,. \tag{9.8a}$$

More generally, for a series of ion species each with concentration C_i and valence z_i,

$$\rho_{\text{ext}}(\mathbf{r}) = e\sum_i z_i C_i(\mathbf{r}). \tag{9.8b}$$

[1] Throughout this section, we keep 4π in both numerator and denominator that could be canceled. We do this for two reasons. First, the quantity $1/4\pi\epsilon_0$ has a numerical value of about 9×10^9, which is easy to remember; second, for those who do not use SI units, the factor $1/4\pi\epsilon_0$ does not appear, but the other factor of 4π remains.

The next step is to assume that the concentrations of all ions are given by Boltzmann factors and are therefore related to the potential by

$$\begin{aligned} K(x) &= [\text{K}]\,e^{-ev(x)/k_BT} \quad \text{for all } x,\\ Cl(x) &= [\text{Cl}]\,e^{ev(x)/k_BT} \quad \text{for all } x,\\ M(x) &= \left[\text{M}'\right]e^{-e(v(x)-v')/k_BT}, \quad x>0. \end{aligned} \tag{9.9a}$$

(Remember that $M(x) = 0$ to the left of the origin.) An equivalent general expression is

$$\rho_{\text{ext}}(\mathbf{r}) = e\sum_i z_i\,[C_i]\exp\left[\frac{-z_i e v(\mathbf{r})}{k_BT}\right], \tag{9.9b}$$

where C_i is the concentration in the region where $v = 0$.

Combining Eqs. 9.7 and 9.9b gives the *Poisson–Boltzmann equation* for a dielectric:

$$\nabla^2 v = -\frac{4\pi e}{4\pi\epsilon_0\kappa}\sum_i z_i\,[C_i]\exp\left(\frac{-z_i e v(\mathbf{r})}{k_BT}\right). \tag{9.10}$$

For the specific problem at hand, the Poisson–Boltzmann equation takes the form

$$\frac{d^2v}{dx^2} = \frac{-4\pi e}{4\pi\epsilon_0\kappa}\left([\text{K}]\,e^{-ev(x)/k_BT} - [\text{Cl}]\,e^{ev(x)/k_BT}\right).$$

This applies for $x < 0$ only. While it is possible to solve this using numerical techniques (Mauro 1962), we will confine ourselves to the case in which $\xi = ev/k_BT \ll 1$, and we can make the approximation $e^{\xi} \approx 1+\xi$. (This is accurate to 0.5 % for $\xi = 0.1$, to 10 % for $\xi = 0.5$, and to 25 % for $\xi = 0.8$.) With this approximation

$$\begin{aligned} \rho_{\text{ext}} &= e\sum[C_i]\,z_i\left(1-\frac{z_i ev}{k_BT}\right) = \\ &e\sum[C_i]\,z_i - \frac{e^2}{k_BT}\sum[C_i]\,z_i^2 v. \end{aligned} \tag{9.11}$$

Far from the membrane the solution is electrically neutral, so the first term vanishes. We are left with the *linear Poisson–Boltzmann equation*:

$$\nabla^2 v(\mathbf{r}) = \frac{4\pi e^2\sum[C_i]\,z_i^2}{4\pi\epsilon_0\kappa\,k_BT}v(\mathbf{r}). \tag{9.12}$$

The coefficient of $v(\mathbf{r})$ on the right has the dimensions of $1/(\text{length})^2$. This length will also appear in other contexts. It is known as the *Debye length*, λ_D:

$$\frac{1}{\lambda_D^2} = \frac{4\pi e^2\sum[C_i]\,z_i^2}{4\pi\epsilon_0\kappa\,k_BT}. \tag{9.13}$$

The linearized Poisson–Boltzmann equation is

$$\nabla^2 v = \frac{v}{\lambda_D^2}. \tag{9.14}$$

For the one-dimensional problem and $x < 0$, it is

$$\frac{d^2v}{dx^2} = \frac{v}{\lambda_D^2}, \tag{9.15}$$

where

$$\frac{1}{\lambda_D^2} = \frac{4\pi e^2 ([\mathrm{K}] + [\mathrm{Cl}])}{4\pi\epsilon_0 \kappa k_B T}. \tag{9.16}$$

The methods of Appendix F can be applied to solve this equation.[2] The characteristic equation is $s^2 = 1/\lambda_D^2$, so the solution for $x < 0$ is $v(x) = Ae^{-x/\lambda_D} + Be^{x/\lambda_D}$. The potential is zero far to the left, so $A = 0$. Therefore, the solution is

$$v(x) = Be^{x/\lambda_D}, \quad x < 0. \tag{9.17}$$

It is most convenient to write the concentrations for $x > 0$ in terms of the concentrations far to the right. It is now necessary to include the impermeant ions.

$$\begin{aligned} K(x) &= [\mathrm{K}']\, e^{-e[v(x)-v']/k_B T}, \\ Cl(x) &= [\mathrm{Cl}']\, e^{e[v(x)-v']/k_B T}, \\ M(x) &= [\mathrm{M}']\, e^{-e[v(x)-v']/k_B T}. \end{aligned} \tag{9.18}$$

The linearized Poisson–Boltzmann equation for $x > 0$ is then

$$\begin{aligned} \frac{d^2v}{dx^2} = -\frac{4\pi e}{4\pi\epsilon_0\kappa}\Bigg(& [\mathrm{K}'] - \frac{[\mathrm{K}']\, ev(x)}{k_B T} + \frac{[\mathrm{K}']\, ev'}{k_B T} \\ & - [\mathrm{Cl}'] - \frac{[\mathrm{Cl}']\, ev(x)}{k_B T} + \frac{[\mathrm{Cl}']\, ev'}{k_B T} \\ & + [\mathrm{M}'] - \frac{[\mathrm{M}']\, ev(x)}{k_B T} + \frac{[\mathrm{M}']\, ev'}{k_B T} \Bigg). \end{aligned} \tag{9.19}$$

Neutrality requires that $[\mathrm{K}'] + [\mathrm{M}'] - [\mathrm{Cl}'] = 0$. With the definition

$$\frac{1}{\lambda_D'^2} = \frac{4\pi e^2([\mathrm{K}'] + [\mathrm{Cl}'] + [\mathrm{M}'])}{4\pi\epsilon_0\kappa k_B T}, \tag{9.20}$$

Eq. 9.19 can be written as

$$\frac{d^2v}{dx^2} - \frac{v(x)}{\lambda_D'^2} = -\frac{v'}{\lambda_D'^2}. \tag{9.21}$$

[2] We have seen this equation before in electrotonus when the membrane capacitance is fully charged (Sect. 6.12).

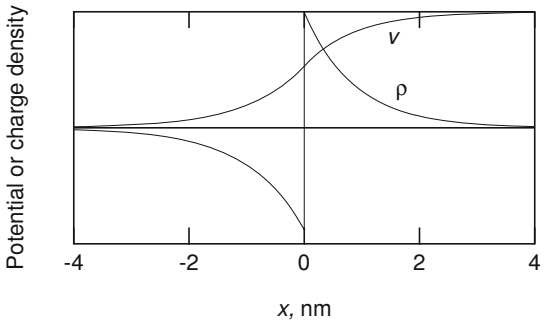


Fig. 9.5 The potential and charge density in the vicinity of the Donnan membrane. There is a layer of negative charge on the left of the membrane and of positive charge on the right. Each decays with the Debye length given by the ion concentrations far from the membrane

This is an inhomogeneous linear differential equation with constant coefficients. As pointed out in Appendix F, the most general solution is the sum of the solution to the homogeneous equation (i.e., with the right hand side equal to 0) and any solution of the inhomogeneous equation, with the constants adjusted to satisfy whatever boundary conditions exist. In this case, $v(x) = v'$ satisfies the inhomogeneous equation, so the most general solution is $v(x) = A'e^{-x/\lambda_D'} + B'e^{x/\lambda_D'} + v'$. Far to the right, $v = v'$ so $B' = 0$. Therefore, the solution we need is

$$v(x) = A'e^{-x/\lambda_D'} + v' \quad x > 0. \tag{9.22}$$

This solution for $x > 0$ must be combined with the solution for $x < 0$, Eq. 9.17. At $x = 0$ the potential must be continuous. Therefore $B = A' + v'$. Also at $x = 0$ the electric field, and therefore dv/dx, is continuous. (If dv/dx were not continuous, the second derivative and ρ_{ext} would be infinite.) This requirement gives the equation $B/\lambda_D = -A'/\lambda_D'$. Solving these two equations, we obtain

$$A' = \frac{-v'\lambda_D'}{\lambda_D' + \lambda_D}, \quad B = \frac{v'\lambda_D}{\lambda_D' + \lambda_D}. \tag{9.23}$$

Figures 9.5 and 9.6 show the potential, concentration, and charge density for the case $[\mathrm{K}] = 100$ and $[\mathrm{M}'] = 50$ mmol l^{-1}. The other parameters are given in Table 9.2. The value of $ev'/k_B T$ is 0.23.

Since the radii of ions are about 0.2 nm, the Debye length is several ionic diameters, and the continuous model we have used is reasonable.

The Poisson–Boltzmann equation is widely used to study charged molecules in solution (Honig and Nicholls 1995) and has implications for how proteins bind to DNA (Rohs et al. 2009). However, in small-scale systems such as ion channels, which have a size similar to or smaller than the Debye length, continuous models may not be entirely reliable (Moy et al. 2000).

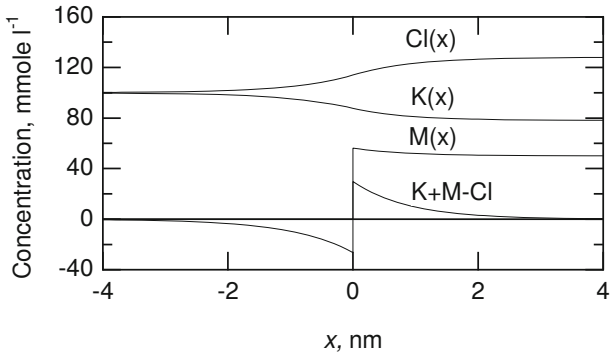


Fig. 9.6 Concentration profiles across the Donnan membrane. The concentration $K(x)+M(x)-Cl(x)$ is proportional to the charge density

Table 9.2 Parameters for the Donnan interface when [K] = 100, [M] = 0, and $[M'] = 50\,\text{mmol}\,l^{-1}$ at $T = 310$ K

[Cl]	$100\,\text{mmol}\,l^{-1}$
[K]	$100\,\text{mmol}\,l^{-1}$
[M]	$0\,\text{mmol}\,l^{-1}$
[K′]	$78.1\,\text{mmol}\,l^{-1}$
[Cl′]	$128.1\,\text{mmol}\,l^{-1}$
[M′]	$50\,\text{mmol}\,l^{-1}$
v'	6.617 mV
λ_D	0.991 nm
λ'_D	0.875 nm

The Gouy–Chapman model has been compared to detailed *molecular dynamics* simulations (in which every molecule is individually accounted for) for the case of salt water surrounding a lipid bilayer. The two computations are consistent as long as the adsorption of ions on the bilayer surface is accounted for (Yi et al. 2008).

9.3 Ions in Solution: The Debye–Hückel Model

In an ionic solution, ions of opposite charge attract one another. A model of this neutralization was developed by Debye and Hückel a few years after Gouy and Chapman developed the model in the previous section. The Debye–Hückel model singles out a particular ion and assumes that the average concentration of the counterions surrounding it is given by the Boltzmann factor. Screening by the counterions causes the potential to fall much more rapidly than $1/r$. One major difficulty with this assumption is that each counterion is also a central ion; therefore, the notion of a continuous cloud of counterions represents some sort of average.

We consider a situation in which the electric field, potential, and charge distribution are spherically symmetric. We could begin with Eq. 9.7 and look up the Laplacian operator in spherical coordinates. However, it is instructive to derive Poisson's equation for the spherically symmetric case. Consider two concentric spheres of radius r and radius $r + dr$. Apply Gauss's law to the volume contained between the two surfaces. If **E** is spherically symmetric, the flux through the inner sphere is $4\pi r^2 E(r)$. It points into the sphere and is therefore negative. The outward flux at $r + dr$ is

$$4\pi(r+dr)^2 E(r+dr) = 4\pi\left[r^2 + 2rdr + (dr)^2\right]\left[E(r) + \frac{dE}{dr}dr\right].$$

If we keep only terms of order dr or less, the outward flux through the outer sphere is

$$4\pi r^2 E(r) + 8\pi r E(r)dr + 4\pi r^2 \frac{dE}{dr}dr.$$

The net flux out of the volume is $8\pi r E(r)dr + 4\pi r^2(dE/dr)dr$. The total charge in the shell is $\rho_{\text{ext}}(r)$ times the volume of the shell, $4\pi r^2 dr$. Therefore, Gauss's law is

$$8\pi r E(r)dr + 4\pi r^2 \frac{dE}{dr}dr = \rho_{\text{ext}}(r)\frac{4\pi r^2}{\kappa\epsilon_0}dr$$

or

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 E(r)\right) = \frac{4\pi\rho_{\text{ext}}(r)}{4\pi\epsilon_0\kappa}. \tag{9.24}$$

Since $E(r) = -dv/dr$, the final equation for the potential is

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dv}{dr}\right) = -\frac{4\pi\rho_{\text{ext}}(r)}{4\pi\epsilon_0\kappa}. \tag{9.25}$$

The Poisson–Boltzmann equation in spherical coordinates, the analog of Eq. 9.10, is

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dv}{dr}\right) = -\frac{4\pi e}{4\pi\epsilon_0\kappa}\sum z_i\,[C_i]\exp\left(\frac{-z_i e v(r)}{k_B T}\right). \tag{9.26}$$

We again make a linear approximation to the Boltzmann factor to obtain the linear Poisson–Boltzmann equation for spherical symmetry:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dv}{dr}\right) = \frac{1}{\lambda_D^2}v(r). \tag{9.27}$$

The Debye length λ_D is defined in Eq. 9.13. With the substitution $v(r) = u(r)/r$, the equation becomes

$$\frac{d^2u}{dr^2} = \frac{1}{\lambda_D^2}u(r), \tag{9.28}$$

which is the same as Eq. 9.15. Therefore, the solution is

$$v(r) = \frac{u(r)}{r} = \frac{Ae^{-r/\lambda_D} + Be^{r/\lambda_D}}{r}.$$

Requiring that $v(r)$ approaches 0 as $r \to \infty$ means that $B = 0$. For small r, the electric field (dv/dr) is that of an unshielded ion of charge ze. Therefore $A = ze/4\pi\epsilon_0\kappa$, and the final solution is

$$v(r) = \left(\frac{ze}{4\pi\epsilon_0\kappa}\right)\left(\frac{e^{-r/\lambda_D}}{r}\right). \tag{9.29}$$

Table 9.3 The Debye–Hückel potential for a monovalent ion in a solution of ions at the concentration given in Fig. 6.2 for the interior of an axon. Also shown are the unscreened potential, the parameter zev/k_BT, and the charge inside a sphere of radius r

r (nm)	$v(r)$ (mV)	$e/(4\pi\epsilon_0\kappa r)$ (mV)	ev/k_BT	$q(r)/e$
0.3	40.6	59.9	1.52	0.94
0.4	26.8	44.9	1.00	0.90
0.5	18.8	36.0	0.70	0.86
0.6	13.8	30.0	0.51	0.82
0.7	10.4	25.7	0.39	0.77
0.8	8.0	22.5	0.30	0.72
0.9	6.2	20.0	0.23	0.67
1.0	4.9	18.0	0.18	0.63
1.2	3.2	15.0	0.12	0.54
1.4	2.1	12.8	0.08	0.46
1.6	1.4	11.2	0.05	0.39
1.8	1.0	10.0	0.04	0.32
2.0	0.7	8.99	0.03	0.27
2.2	0.5	8.17	0.02	0.22
2.4	0.3	7.49	0.01	0.18
2.6	0.2	6.91	0.01	0.15
2.8	0.2	6.42	0.01	0.12
3.0	0.1	5.99	0.00	0.10

This is the potential of a point charge ze in a dielectric, modified by an exponential decay over the Debye length. From Eq. 9.13, one sees that the greater the concentration of counterions, the shorter the Debye length.

Table 9.3 shows the values of $v(r)$, $\xi = ev/k_BT$, and the potential from an unscreened point charge in water of dielectric constant 80, when the ion concentrations are those given in Fig. 6.3. A typical ion radius is about 0.2 nm. We will discover in the next section that the dielectric constant saturates for $r < 0.25$ nm. Therefore, values are given in Table 9.3 only for $r > 0.3$ nm. The table shows that the assumption $e^{\xi} \approx 1 + \xi$ is reasonable only for $r > 0.5$ nm. The Debye length is $\lambda_D = 0.77$ nm.

The charge density of the ion cloud can be obtained from Eqs. 9.25 and 9.29. The result is

$$\rho_{\text{ext}}(r) = \frac{-ze}{4\pi\lambda_D^2 r} e^{-r/\lambda_D}. \tag{9.30}$$

The total charge in the counterion cloud inside a sphere of radius a is

$$\int_0^a 4\pi r^2 \rho_{\text{ext}}(r)\, dr.$$

Adding to this a point charge ze at the origin gives the total charge due to both the ion and the counterion cloud inside radius a:

$$q(a) = ze\left(1 + \frac{a}{\lambda_D}\right) e^{-a/\lambda_D}. \tag{9.31}$$

This function approaches ze, the charge of the point ion, as $a \to 0$, and it approaches 0 as $a \to \infty$. Table 9.3 also shows the values of $q(a)/e$. Ninety percent of the counterion charge resides within 3 nm of the central ion. The charge on the central ion is half neutralized by charge in a sphere of radius 1.3 nm, about six ionic radii. Figure 9.7 shows schematically an ion of radius 0.2 nm. Since a monovalent ion will be neutralized by a single counterion, it is clear that the assumption of a continuous charge distribution equal to the average is a bit strained. The shaded circle of radius 0.25 nm represents the region in which the water molecules are completely polarized and the dielectric constant is less than 80; this is discussed in the next section. (We have ignored the fact that close to the central ion the linear approximation is not valid.)

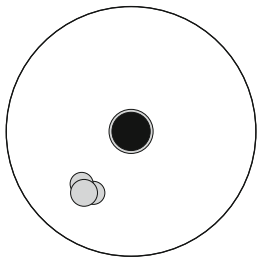


Fig. 9.7 Schematic picture of the regions surrounding an ion. The solid circle in the center represents the ion of radius 0.2 nm. The shaded circle shows the region in which the polarization of the water is saturated. The outer circle of radius 1.3 nm represents the region within which the cloud of counterions has neutralized half of the charge on the ion, which means that on the average a counterion will be in this region half of the time. This radius depends on the ion concentrations that are those for the interior of a squid axon. A scale drawing of a water molecule is also shown

When a highly charged molecule (for example, a strand of DNA) is surrounded by multivalent counterions, the counterions may interact so strongly that they are correlated with each other. Such effects are not included in the Debye–Hückel model. In some cases, these counterions cause charge inversion: so many correlated positive counterions form around a central negatively charged molecule that from a distance the molecule appears to have a net positive charge (Grosberg et al. 2002).

9.4 Saturation of the Dielectric

The electric field in vacuum at distance r from a point charge q is $E = q/(4\pi\epsilon_0 r^2)$. If the charge is in a dielectric, the field is reduced by a factor $1/\kappa$, except at very small distances, where the electric field is so strong that the polarization of the dielectric is saturated.

A molecule of water appears schematically as shown in Fig. 6.18. The radius of each hydrogen atom is about 0.12 nm; the radius of the oxygen is about 0.14 nm. Each hydrogen nucleus is 96.5 pm from the oxygen; the angle between them is 104 °. The hydrogen atoms share their electrons with the oxygen in such a way that each hydrogen atom has a net positive charge and the oxygen has a net negative charge. A pair of charges $\pm q$ separated by distance b has an *electric dipole moment* $\mathbf{p}_e$ of magnitude $p_e = qb$.

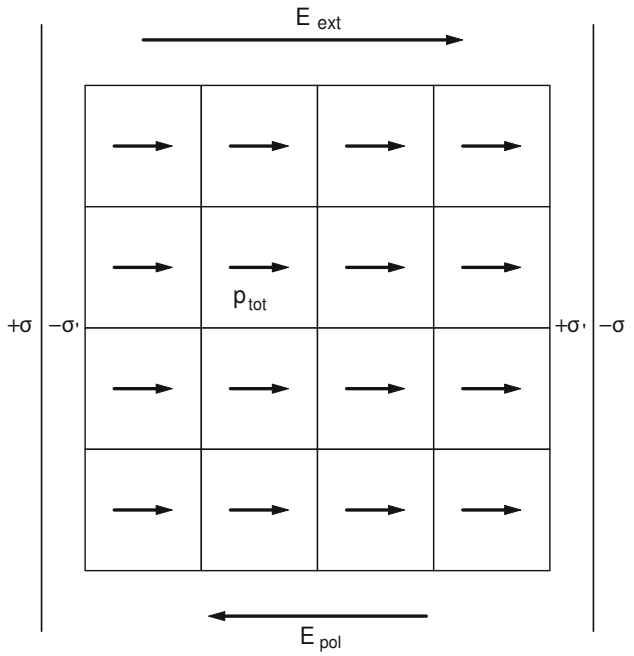


Fig. 9.8 A dielectric is placed in a parallel-plate capacitor that has charge density $\pm\sigma$ on each plate. A dipole moment of magnitude p_{tot} is induced in each volume element of the dielectric. The total effect is the same as a charge density $\pm\sigma'$ induced on the surfaces of the dielectric

The vector points from the negative to the positive charge. The magnitude of the dipole moment of a water molecule is 6.237×10^{-30} C m.

Each molecule of a dielectric in an applied electric field has an induced dipole moment that reduces the field. This dipole moment can be caused by a displacement of the electron cloud with respect to the nucleus, or it can represent (as for a polar molecule like water) an average molecular alignment against the tendency of thermal motion to orient the water molecules randomly.

The average induced dipole moment gives rise to the polarization field $\mathbf{E}_{\text{pol}}$ (Eqs. 6.19–6.20). To see the relationship, consider a small volume in the dielectric with N molecules per unit volume. Each molecule has an electric dipole moment $p_e = qb$. Far from this volume, the potential is primarily due to the dipole moment of each molecule. This can be shown by arguments like those in Sects. 7.3 and 7.4. The potential depends on the total dipole moment of the volume. The total number of dipoles in the volume is $NSdx$, so $p_{\text{tot}} = p_e NSdx$. This is equivalent to a charge $q' = p_e NS$ on the ends of the volume element, or a surface of charge density

$$\sigma'_q = \frac{q'}{S} = p_e N. \tag{9.32}$$

Now consider a parallel-plate capacitor as shown in Fig. 9.8. Imagine a series of small volume elements in the dielectric. The induced charges $\pm\sigma'_q$ on adjacent surfaces of each row of volume elements cancel except at the end of each row. The polarization field is therefore entirely due to the induced charge of surface density $\pm\sigma'_q$ at each end of the dielectric. The magnitude of the field is

$$E_{\text{pol}} = \frac{\sigma'_q}{\epsilon_0} = \frac{N p_e}{\epsilon_0}. \tag{9.33}$$

The quantity Np_e is the dipole moment per unit volume and is called the *polarization* P.

As the external electric field is increased, E_{pol}, which points in the opposite direction, also increases. This corresponds to the water molecules becoming more and more aligned. From the definition of susceptibility and the dielectric constant in Sect. 6.7, the magnitudes are related by

$$|E_{\text{pol}}| = \frac{\chi}{1+\chi}|E_{\text{ext}}| = \left(1 - \frac{1}{\kappa}\right)|E_{\text{ext}}|.$$

For a monovalent ion in water, $E_{\text{pol}} = (79/80)E_{\text{ext}} = (79/80)e/(4\pi\epsilon_0 r^2)$. When the dipoles are completely aligned, E_{pol} saturates at its maximum value, given by Eq. 9.33 with the molecular dipole moment substituted for p_e. The number of water molecules per unit volume is obtained from the fact that 1 mol has a mass of 18 g, occupies 1 cm^3 g^{-1}, and contains N_A molecules:

$$\begin{aligned} &E_{\text{pol}}(\text{max}) \\ &= \left[\frac{(N_A \text{ molecule mol}^{-1})(1 \text{ g cm}^{-3})(10^6 \text{ cm}^3 \text{ m}^{-3})}{(18 \text{ g mol}^{-1})\epsilon_0 \text{ C V}^{-1} \text{ m}^{-1}}\right] \\ &\times \left[6.237 \times 10^{-30} \text{ C m molecule}^{-1}\right] \\ &= 2.36 \times 10^{10} \text{ V m}^{-1}. \end{aligned}$$

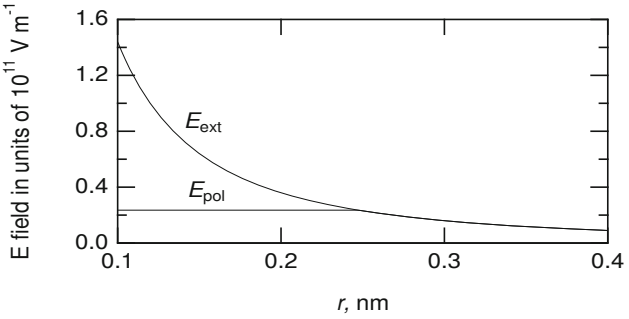


Fig. 9.9 The electric field around a monovalent point charge and the polarization electric field due to the water. The polarization field saturates for $r < 0.23$ nm

Figure 9.9 shows the fields E_{ext} and E_{pol} around a monovalent ion. As E_{pol} saturates, E_{tot} rises toward the value it would have without a dielectric. The dielectric constant falls

from 80 to 1 at about 0.23 nm. A more accurate model predicts similar behavior, but with a more gradual transition of the dielectric constant from 80 to 1.[3]

Close to an ion the potential is larger than $q/(4\pi\epsilon_0\kappa\, r)$. This changes the Born charging energy (Eq. 6.22), and the free energy change as an ion dissolves in a solvent (Bockris and Reddy 1970, Chap. 2). Also, close to an ion, the continuum approximation breaks down.

9.5 Ion Movement in Solution: The Nernst–Planck Equation

Solute particles can move by diffusion. They can also move if they have an average velocity V_{solute}. There are two ways they can acquire an average velocity. The first is if they are at rest on average with respect to a moving solution. This is called *solvent drag*. The second is for the solute particles to be dragged through the solution by an external force that acts on them, such as gravity or an electric force, balanced by the viscous force on the particles. In both cases, number per unit area per unit time crossing a plane is CV_{solute}. The solute particle fluence rate (particle current density) due to both diffusion and the solute velocity in the x direction is[4] (Sect. 4.12)

$$j_s = -D\frac{dC}{dx} + CV_{\text{solute}}. \tag{9.34}$$

Suppose that an external force $\mathbf{F} = ze\mathbf{E}$ acts on the solute particles in the x direction. They will be accelerated until the viscous drag on them is equal to the magnitude of F. But we saw in Chap. 4 that the viscous drag is $f = -\beta(V_{\text{solute}} - V_{\text{solvent}})$ where $V_{\text{solute}} - V_{\text{solvent}}$ is the relative velocity of the solute through the solvent. Coefficient β is related to the diffusion constant by $\beta = k_BT/D$. Therefore, the particles are no longer accelerated when

$$V_{\text{solute}} - V_{\text{solvent}} = zeE/\beta. \tag{9.35}$$

Equation 9.34 can be rewritten as

$$j_s = -D\frac{dC}{dx} + C\left[V_{\text{solvent}} + (V_{\text{solute}} - V_{\text{solvent}})\right].$$

Now V_{solvent} is the volume of solvent that flows per unit area per unit time and is just j_v. With this substitution and using Eq. 9.35, the particle current density is

$$j_s = -D\frac{dC}{dx} + Cj_v + CzeE\frac{D}{k_BT}. \tag{9.36}$$

[3] A more sophisticated model for the alignment of the electric dipoles in the electric field is analogous to that for magnetic moments in Sect. 8.3.

[4] We use x for the distance in the direction parallel to $\mathbf{E}$ because z is used for valence.

Table 9.4 Conductivities of ions at various concentrations at 25°C, calculated using Eq. 9.39. Diffusion constants for each ion are from Hille (2001, p. 317). Concentrations are typical of mammalian nerve and are from Hille (2001, p. 17). The conductivities of each species add, and $\rho = 1/\sigma$. Larger ions with very small diffusion constants make the solutions electrically neutral

	D (m^2s^{-1})	C (mmol l^{-1})	σ (S m^{-1})	ρ (Ω m)
Extracellular squid axon				
Na	1.33×10^{-9}	145	0.723	
K	1.96×10^{-9}	4	0.029	
Cl	2.03×10^{-9}	123	0.936	
			1.688	0.592
Intracellular squid axon				
Na	1.33×10^{-9}	12	0.060	
K	1.96×10^{-9}	155	1.139	
Cl	2.03×10^{-9}	4.2	0.032	
			1.231	0.812

The first term represents solute motion due to diffusion, the second represents solute dragged along with the bulk flow of the solution (solvent drag), and the third represents drift due to the applied electric field.

We will consider only the case in which there is no bulk flow of solution, so $j_v = 0$. The equation then reduces to the *Nernst–Planck equation*:

$$j_s = -D\frac{dC}{dx} + \frac{zeE}{k_BT}DC. \tag{9.37}$$

Diffusion is always toward the region of lower concentration, while for positive charge the V_{solute} term is in the direction of $\mathbf{E}$. For negative charges, it is in the opposite direction.

Consider the current density in bulk solution between planes at $x = 0$ where $v(x) = 0$ and $x = L$ where $v(x) = v$. If there is no concentration gradient and the potential changes uniformly, then $E = -dv/dx = -v/L$ points in the negative x direction, and the particle current density is $j_s = -zeDCv/k_BTL$. The electrical current density j is obtained by multiplying j_s by the charge on each particle, ze:

$$j = -\frac{z^2e^2DCS}{k_BTL}\frac{v}{S} = -\frac{G(C)}{S}v. \tag{9.38}$$

If $v(L) > v(0)$, the current is to the left and is negative. Recalling that $G = \sigma S/L = 1/R = S/\rho L$, we obtain the conductivity in the bulk solution

$$\sigma = \frac{1}{\rho} = \frac{z^2e^2DC}{k_BT}. \tag{9.39}$$

If several ion species carry current and can be assumed to move independently, then the total conductivity is the sum of the conductivities for each ion. Table 9.4 shows contributions to the conductivity for various species at typical concentrations.

This model is satisfactory for material such as the inside of an axon where the concentrations are constant and the material is electrically neutral, so that the ions themselves do not on average contribute to the electric field. We have assumed that the ions move independently, which will happen only if the electric field of other ions can be ignored.

We can model ions flowing from a region of one concentration to another (such as crossing the axon membrane) with the Nernst–Planck equation. Writing it for the electric current density and using the fact that $E(x) = -dv/dx$, we have

$$j = -zeD\frac{dC}{dx} - \frac{z^2e^2D}{k_BT}\frac{dv}{dx}C. \tag{9.40}$$

It is simpler to use the dimensionless variable $u(x) = zev(x)/k_BT$, which is the ratio of an ion's energy to thermal energy:

$$j = -zeD\left(\frac{dC}{dx} + C\frac{du}{dx}\right). \tag{9.41}$$

If we assume that dv/dx is constant throughout the region, $v(0) = 0$ and $v(L) = v$, then the gradient is $dv/dx = v/L$, and Equation 9.40 becomes

$$\frac{dC}{dx} - \frac{1}{\lambda}C = -\frac{j}{zeD}, \tag{9.42}$$

where the characteristic length for this model (*not* the Debye length) is

$$\lambda = -\frac{L}{u} = -\frac{k_BTL}{zev}. \tag{9.43}$$

Equation 9.42 is the same as Eq. 4.58, except for the denominator of the term involving j. Here the denominator is zeD because j is the electric current density instead of the particle current density. The solution analogous to Eq. 4.62 is

$$j = \frac{zeD}{\lambda}\frac{C_0e^{L/\lambda} - C_0'}{e^{L/\lambda} - 1} = \frac{zeD}{\lambda}\frac{C_0e^{-u} - C_0'}{e^{-u} - 1}, \tag{9.44}$$

where C_0 is the ion concentration at $x = 0$ and C_0' is the concentration at $x = L$.

The current vanishes if $C_0e^{L/\lambda} - C_0' = 0$, or $C_0'/C_0 = e^{L/\lambda} = e^{-zev/k_BT} = e^{-u}$. This is the Boltzmann factor.

Equation 9.44 can be written in terms of the original variables:

$$j = -\frac{z^2e^2Dv}{k_BTL}\frac{C_0e^{-zev/k_BT} - C_0'}{e^{-zev/k_BT} - 1} = -\frac{zeDu}{L}\frac{C_0e^{-u} - C_0'}{e^{-u} - 1}. \tag{9.45}$$

It is interesting to compare this to Eq. 9.38. Since G depends on concentration, it is useful to factor out C_0 and write

$$\begin{aligned} j &= -\frac{z^2e^2DC_0}{k_BTL}\frac{e^{-zev/k_BT} - C_0'/C_0}{e^{-zev/k_BT} - 1}v \\ &= -\frac{G(C_0)}{S}\frac{e^{-zev/k_BT} - C_0'/C_0}{e^{-zev/k_BT} - 1}v. \end{aligned} \tag{9.46}$$

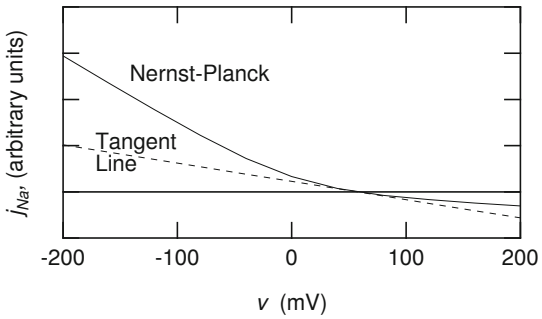


Fig. 9.10 Sodium current versus applied potential for the constant field Nernst–Planck model when the sodium concentration is 145 mM on the *left* and 15 mM on the *right*. The calculation was done using Eq. 9.45 for $T = 293$ K. The *tangent line* was calculated using Eq. 9.47. The nonlinearity or rectification occurs because of the different ion concentrations on each side

If $C_0 = C_0'$, we recover Eq. 9.38. Figure 9.10 shows the current density in A m^{-2} for a situation where $C_0 = 145$ and $C_0' = 15$ mmol l^{-1}. The diffusion constant for sodium from Table 9.4 has been used. As $C_0 > C_0'$, equilibrium occurs when $v = +57.3$ mV at 20°C.

Note the nonlinearity of the current–voltage relationship that arises because $C_0 \neq C_0'$. For very negative potentials, the flow is almost entirely from left to right and the current density approaches $G(C_0)v/S$ while for very positive potentials, the flow is from right to left and the current density approaches $G(C_0')v/S$. This asymmetry is fundamental. It occurs because there are different numbers of charge carriers on the left and right. When this behavior is seen in channels in cell membranes, they are often called *rectifier channels*. This same asymmetry in differences in the concentration of charge carriers is responsible for rectification in semiconductors.

Near the Nernst potential, the current density has the form $j = -(G/S)(v - v_{\text{Nernst}})$ if

$$\frac{G}{S} = \frac{G(C_0)(zev_{\text{Nernst}}/k_BT)}{S\left(e^{zev_{\text{Nernst}}/k_BT} - 1\right)}. \tag{9.47}$$

This equation was used to derive the tangent line shown in Fig. 9.10.

The constant-field model is an oversimplification. The field can be distorted by fixed charges near the channel through which the ions are flowing. Moreover, the model is internally inconsistent. There are electric fields generated by the flowing ions, which become important at high concentrations. The fact that $j = 0$ when the potential is equal to the Nernst potential is fundamental and holds for any ion or model for conduction. It can be derived in the general case from Eq. 9.42 (Problem 15). A self-consistent analytic solution for the case of a single ion species has been known for

50 years. The solution has been extended by many workers and has been generalized by Leuchtag and Swihart (1977) to the case in which all the ions have the same charge.

9.6 Zero Total Current in a Constant-Field Membrane: The Goldman Equations

The Nernst–Planck equation can be used to calculate the current due to movement of ions through a membrane in which there is a constant electric field. We assume a constant field because it leads to an analytic solution and because we have no knowledge of internal structure or the behavior of counterions which could change the field. The resulting equations are called the *Goldman* or the *Goldman–Hodgkin–Katz* (GHK) equations.

The GHK equations can be derived by assuming either a homogeneous membrane, in which case the Nernst–Planck equation is simply applied to each species, or cylindrical pores of constant cross section. Since we know that the pores do not have a constant electric field (Sect. 9.7) and it is quite unlikely that they have constant cross section, the GHK equations are an approximation. Nevertheless, they have been used widely in the study of excitable membranes.

We will show the derivation for a cylindrical pore that has a constant circular cross section. We use cylindrical coordinates (r, ϕ, x), where x is the axis of the cylinder. (Again, z denotes the valence of the ions.) Let the outside of the membrane be at $x = 0$ and the inside at $x = L$, where the potential is v and $u = zev/k_BT$. The arguments of Sect. 5.9 about the r and x dependence can be applied to Eq. 9.41. The analog of Eq. 5.37 is

$$j(r) = -zeD(r, a, R_p)\left(\frac{\partial C(r,x)}{\partial x} + \frac{u}{L}C(r,x)\right). \quad (9.48)$$

Again the concentration can be written as $C(r,x) = C(x)\Gamma(r)$. Equation 9.48 becomes

$$j(r) = -ze\Gamma(r)D(r, a, R_p)\left(\frac{\partial C(x)}{\partial x} + \frac{u}{L}C(x)\right). \quad (9.49)$$

This can be multiplied by $2\pi r\,dr$ and integrated over the pore area. There are two integrals to consider. The first defines the average current density for a particular species:

$$\int_0^{R_p} j(r)2\pi r\,dr = \pi R_p^2\,\bar{j}. \quad (9.50)$$

The second defines an effective diffusion constant:

$$\int_0^{R_p} \Gamma(r)D(r, a, R_p)2\pi r\,dr = \pi R_p^2 D_{\text{eff}}. \quad (9.51)$$

The integrated current density equation is

$$\bar{j} = -zeD_{\text{eff}}\left(\frac{dC(x)}{dx} + \frac{u}{L}C(x)\right). \quad (9.52)$$

Consideration of the r dependence in the pore has given an equation exactly like Eq. 9.41, but with D_{eff} instead of D. Equations 9.42 and 9.43 are still valid. The form of λ is unchanged: $\lambda = -k_BTL/zev$. Conversion from a single pore to unit area of the membrane requires multiplying $\bar{j}$ by $n\pi R_p^2$. As in Eq. 5.49 we define $\omega_s RT = n\pi R_p^2 D_{\text{eff}}/L$ and call the concentration outside C_1 and the concentration inside C_2. The electric current density per unit area of membrane is

$$\begin{aligned} J' &= \frac{z^2e^2\omega_s RTv}{k_BT}\frac{C_1e^{-zev/k_BT} - C_2}{1 - e^{-zev/k_BT}} \\ &= z^2e^2v\omega_s N_A\frac{C_1e^{-zev/k_BT} - C_2}{1 - e^{-zev/k_BT}}. \end{aligned} \quad (9.53)$$

Suppose that three species can pass through the membrane: sodium, potassium, and chloride. Equation 9.53 can be applied separately to each species to obtain the *GHK current equation* for each ion species:

$$J'_{\text{Na}} = e^2v\omega_{\text{Na}}N_A\frac{[\text{Na}_1]\,e^{-ev/k_BT} - [\text{Na}_2]}{1 - e^{-ev/k_BT}}, \quad (9.54a)$$

$$J'_{\text{K}} = e^2v\omega_{\text{K}}N_A\frac{[\text{K}_1]\,e^{-ev/k_BT} - [\text{K}_2]}{1 - e^{-ev/k_BT}}, \quad (9.54b)$$

$$J'_{\text{Cl}} = e^2v\omega_{\text{Cl}}N_A\frac{[\text{Cl}_1]\,e^{+ev/k_BT} - [\text{Cl}_2]}{1 - e^{+ev/k_BT}}. \quad (9.54c)$$

The *reversal potential*, v_{rev}, is the potential for which the total membrane current or fluence rate, that is the sum of the three fluence rates, is zero. The amount of charge within the cell does not change with time, but the concentration of each species within the cell changes with time. This less stringent requirement becomes $J'_{\text{Na}} + J'_{\text{K}} + J'_{\text{Cl}} = 0$. Adding Eqs. 9.54 together and factoring out $N_Ae^2v/(1 - e^{-ev/k_BT})$ gives

$$\begin{aligned} &(\omega_{\text{Na}}[\text{Na}_1] + \omega_{\text{K}}[\text{K}_1] + \omega_{\text{Cl}}[\text{Cl}_2])\,e^{-ev/k_BT} \\ &= \omega_{\text{Na}}[\text{Na}_2] + \omega_{\text{K}}[\text{K}_2] + \omega_{\text{Cl}}[\text{Cl}_1], \end{aligned}$$

or the *GHK voltage equation*

$$v_{\text{rev}} = \frac{k_BT}{e}\ln\left(\frac{\omega_{\text{Na}}[\text{Na}_1] + \omega_{\text{K}}[\text{K}_1] + \omega_{\text{Cl}}[\text{Cl}_2]}{\omega_{\text{Na}}[\text{Na}_2] + \omega_{\text{K}}[\text{K}_2] + \omega_{\text{Cl}}[\text{Cl}_1]}\right). \quad (9.55)$$

As an example of the use of the GHK voltage equation, consider how the reversal potential depends on the concentration of some external ion. We will use the concentrations of Fig. 6.2, except for the ion whose concentration is being changed. The particle concentrations are in mmol l^{-1} (any units can be used since ratios are taken):

$$[\text{Na}_1] = 145, \qquad [\text{Na}_2] = 15,$$

$$[\text{K}_1] = 5, \qquad [\text{K}_2] = 150,$$

$$[\text{Cl}_1] = [\text{Na}_1] + [\text{K}_1] - 25, \qquad [\text{Cl}_2] = [\text{Na}_2] + [\text{K}_2] - 156.$$

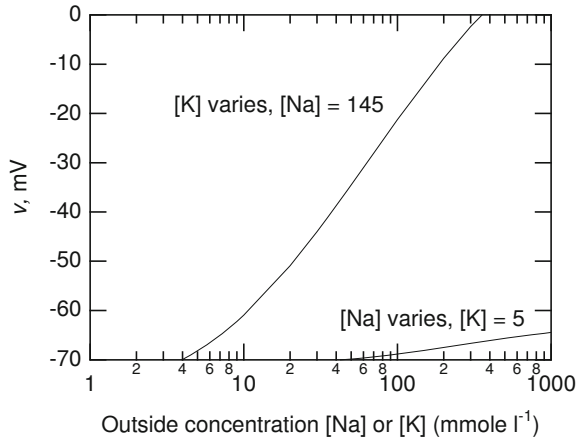


Fig. 9.11 The potential difference across a cell membrane as a function of changes in the exterior concentration of KCl or NaCl, calculated using the Goldman equation

The permeabilities are not known. However, only the ratio to ω_K matters. If we take the ratio $\omega_K : \omega_{Na} : \omega_{Cl}$ to be $1.0 : 0.04 : 0.45$ and use $T = 300$ K, then Eq. 9.55 is (in mV)

$$v = 25.88 \ln\left(\frac{[K_1] + 0.04\,[Na_1] + 0.45([Na_2] + [K_2] - 156)}{[K_2] + 0.04\,[Na_2] + 0.45([Na_1] + [K_1] - 25)}\right).$$

This has been plotted in Fig. 9.11 for variations of $[K_1]$ and $[Na_1]$. In each case, Cl ions are also added to the external solution in an equal amount. There is a region of potassium concentration over which the behavior is nearly exponential, and one could be misled into thinking that the potential–concentration relation was given either by the Nernst equation alone or by Donnan equilibrium. The potential change with sodium concentration is much less because of the low permeability of the membrane to sodium.

The assumption that the total current through the membrane is zero guarantees that there will be no charge buildup inside the cell; however, the individual currents are not zero, so there may be concentration changes with time. We will next investigate the magnitude of this effect. Equation 9.53 can be converted to particle flux instead of charge flux by dividing by ze. The result for ion s is

$$J_s = zev\omega_s \frac{C_1 e^{-zev/k_BT} - C_2}{1 - e^{-zev/k_BT}}.$$

The concentrations are converted from mmol l^{-1} to particles m^{-3} by multiplying by Avogadro's number. (The factors of 10^3 in the conversion happen to cancel out.) Consider the previous example at $T = 300$ K, $[K_1] = 5$, $[Na_1] = 145$, and $v = -68.17$ mV. The exponential factor for the positive ions is $e^{-ev/k_BT} = 13.929$, while for the chloride ion it is the reciprocal, 0.0718. If we write $\omega_{Na} = 0.04\omega_K$ and $\omega_{Cl} = 0.45\omega_K$, then the fluxes for the three ions are

$$J_K = +(6.55 \times 10^3)\omega_K(6.215),$$
$$J_{Na} = -(6.55 \times 10^3)\omega_K(6.202),$$
$$J_{Cl} = -(6.55 \times 10^3)\omega_K(0.013),$$

and the total current is zero.

Although the GHK equations are widely used because of their simplicity, some cautions are in order. Their derivation assumed independence of the moving ions. We know that this is an oversimplification for several reasons. Experiments show that the currents saturate for high concentrations. The distortion of the electric field by other ions was ignored. The permeability (diffusion constant) was assumed to be constant. The pore was assumed to have a constant cross-section and constant electric field. A somewhat less restrictive model for the reversal potential (the potential at which the current density becomes zero and changes sign) can be derived for a pair of ions with the same valence if we assume that any variations in $D(x)$ for the two ions are similar (Problem 20). With that assumption, the reversal potential is

$$v_{rev} = \frac{k_BT}{ze} \ln\left(\frac{\omega_a C_{a1} + \omega_b C_{b1}}{\omega_a C_{a2} + \omega_b C_{b2}}\right). \tag{9.56}$$

When ions have different valences, the GHK equation becomes more complicated. Lewis (1979) has derived an analogous equation for transport of sodium, potassium, and calcium.

9.7 Membrane Channels

In Chap. 6, we described some of the properties of the sodium and potassium channels in a squid axon. There are many other kinds of channels. Variations exist not only from one organism to another, but in different kinds of cells in the same organism. The classic monograph on ion channels is the book by Hille (2001). Genetic mutations of these channel proteins can cause diseases known as *channelopathies* (Ashcroft 2012).

There are several different kinds of potassium channels. Most open after depolarization; a few open after hyperpolarization. Potassium channels in axons (like the ones we encountered in Chap. 6) are called *delayed rectifiers* because of their delay in opening after a voltage step.

The properties of sodium channels are more uniform from one cell type to another.

Calcium channels pass much smaller currents than sodium or potassium channels because calcium concentrations are much smaller; the calcium current density is usually about one tenth the current density for sodium or

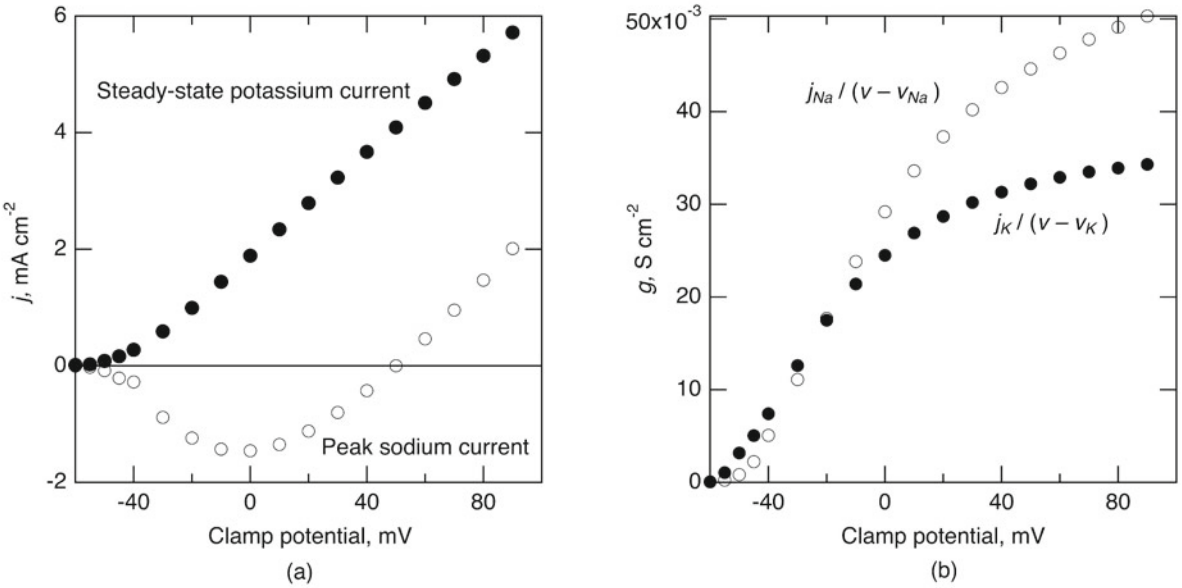


Fig. 9.12 Steady-state potassium current and peak sodium current for a squid axon subject to a voltage clamp vs. the transmembrane potential during the clamp. These are not real data, but were generated using the Hodgkin–Huxley model. **a** Current density. **b** Current density divided by the difference between the potential and the Nernst potential, to give the conductance per unit area. (see Eq. 6.61)

potassium. Calcium channels typically activate with depolarization. Since the concentration of calcium inside cells is usually very small, the interior calcium concentration can increase 20-fold in response to depolarization. This increase in concentration can initiate a chemical reaction, for example, to cause contraction of a muscle cell.

Chloride channels often have a large conductivity. The chloride concentration ratio in some muscle cells is such that the resting potential is close to the chloride Nernst potential. As a result, small changes in the potential cause relatively large chloride currents, which tend to stabilize the resting potential.

The earliest voltage-clamp measurements were difficult to sort out. Hodgkin and Huxley changed the concentration of extracellular sodium, substituting impermeant choline ions, to determine what part of the current was due to sodium and what was due to potassium. Figure 9.12(a) shows typical currents.

In the mid-1960s, various drugs were found that at very small concentrations selectively block conduction of a particular ion species. We now know that these drugs bind to the channels that conduct the ions. An example is *tetrodotoxin* (TTX), which binds to sodium channels and blocks them, making it a deadly poison.

The next big advance was *patch-clamp recording* (Neher and Sakmann 1976). Micropipettes were sealed against a cell membrane that had been cleaned of connective tissue by treatment with enzymes. A very-high-resistance seal

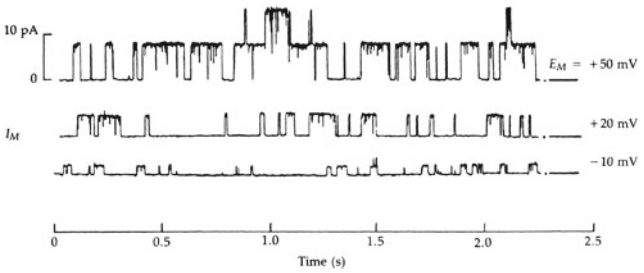


Fig. 9.13 Opening of single K(Ca) channels. (From Pallotta et al. (1981). Reprinted with permission from Nature (London))

resulted [$(2–3)\times10^7$ Ω] that allowed one to see the opening and closing of individual channels. For this work, Erwin Neher and Bert Sakmann received the Nobel Prize in Physiology or Medicine in 1991. Around 1980, Neher's group found a way to make even higher resistance (10^{10}–10^{11} Ω) seals that reduced the noise even further and allowed patches of membrane to be torn from the cell while adhering to the pipette (Hamill et al. 1981). The relationship of noise to resistance will be discussed below.

The patch-clamp studies revealed that the pores open and close randomly, as shown in Fig. 9.13. Thus, the Hodgkin–Huxley model describes the average behavior of many pores, not the kinetics of single pores. Note how the current through

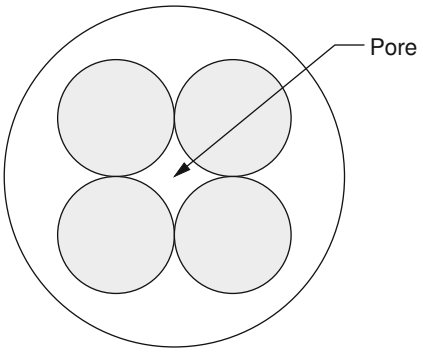


Fig. 9.14 The structure of a *Shaker* potassium channel. There are four subunits that traverse the membrane and create a pore at their center

an open pore changes as a function of the applied potential. A single open pore can pass at least 1 pA of current or 6×10^6 monovalent ions per second. Most can pass much more. While no perfectly selective channel is known, most channels are quite selective; for example, some potassium channels show a 100:1 preference for potassium over sodium.

Gene splicing combined with patch-clamp recording provided a wealth of information. Regions of the DNA responsible for synthesizing the membrane channel have been identified. One example that has been extensively studied is a potassium channel from the fruit fly, *Drosophila melanogaster*. The *Shaker* fruit fly mutant shakes its legs under anesthesia. It was possible to identify exactly the portion of the fly's DNA responsible for the mutation. When *Shaker* DNA was placed in other cells that do not normally have potassium channels, they immediately made functioning channels.

The current view is that the *Shaker* potassium channel consists of four subunits that span the membrane. The pore presumably runs along the four fold-symmetry axis, as shown in end view in Fig. 9.14. Sodium and calcium channels are very similar. Voltage-gated channels are reviewed by Sigworth (1993) and by Keynes (1994).

Roderick MacKinnon and his colleagues determined the three-dimensional structure of a potassium channel using X-ray diffraction (Doyle et al. 1998; Jiang et al. 2003). MacKinnon received the 2003 Nobel Prize in Chemistry for his work on the potassium channel.

The channel protein contains four identical subunits, arranged with four-fold symmetry around a central pore (Fig. 9.14). Each subunit has two alpha helices that cross the membrane and an inner pore region. One of the remarkable features of this channel is that potassium ions are 10,000 times more likely to pass through than sodium ions. Yet, potassium and sodium have similar chemistry (they are in the same column of the periodic table), and their ions are identical except for size (0.133 nm radius for potassium, 0.095 nm for sodium). The channel structure suggests that a narrow, 1.2 nm long region of the pore is responsible for selectivity. As the ion enters this region, there is not enough room for the polar water molecules that normally surround it and shield its charge. Instead, carbonyl oxygen atoms on the channel protein come in close contact with the potassium ion and provide the shielding. The size of the pore is such that potassium ions fit snugly with the surrounding carbonyl oxygen atoms, but sodium does not fit as well.

X-ray diffraction studies have also clarified the mechanism of voltage dependence in potassium channels. The pore is surrounded by the charged structures on the channel's perimeter that sense the transmembrane voltage. These structures act somewhat like levers, opening and closing the pore in response to the voltage. The movement of these structures is responsible for gating currents in these channels.

The structure of the sodium channel has recently been determined (Payandeh et al. 2011).

Let us now explore some of the physics of ion channels. Combining the macroscopic current density with the current in a single channel shows that there are not many channels per unit area of the membrane (see Problems 21 and 22). It is illuminating to consider what effect currents of this magnitude and duration have on the transmembrane potential. The capacitance per unit area of biological membranes is about 0.01 F m^{-2} (1 μF cm^{-2}). A channel conducting 1 pA for 1 ms allows 10^{-15} C to pass. This is enough charge to change the potential 100 mV on an area of 10^{-12} m^2 or 1 μm^2. This charge transfer corresponds to about 6000 monovalent ions per μm^2.

Figure 9.12a shows the steady-state potassium and peak sodium current densities for a squid axon. The ion concentrations are known, and we saw in Chap. 6 that the Nernst potentials at 6.3 °C were +50 mV for sodium and −77 mV for potassium. Figure 9.12(b) shows the conductance per unit area, obtained by dividing the current by $v - v_{\text{Nernst}}$. Figure 9.15 shows a semilogarithmic plot of the conductance per unit area.

The sodium current density changes sign at the sodium Nernst potential. While a measured zero crossing is an accurate way to determine the Nernst potential, extrapolation to find the zero-crossing can be quite misleading. The potassium current density appears to be linear over a large region, and it is tempting to extrapolate to find v_K. The extrapolation shows zero current at about −40 mV, which is far from v_K. The reason can be seen in Fig. 9.12(b), which shows that g_K is varying considerably over the region where j_K appears to be linear; this distorts the slope and changes the extrapolated intersection.

A simple two-state model can explain the general shape of the curves in Fig. 9.15. The conductance per unit area of a membrane is the product of the conductance of an open pore and the average number of pores per unit area that are open.

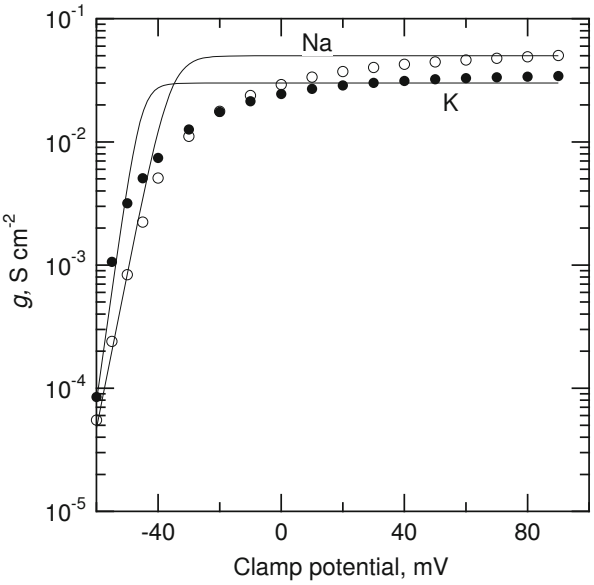


Fig. 9.15 Semilog plot of sodium and potassium conductivities from Fig. 9.12(b) with fits by Eq. 9.57. For sodium $u_o = -10.5$ and $z = -7$; for potassium $u_o = -19$ and $z = -10$

The model assumes that each channel has a gate that is either open or closed. When the gate is open, the channel has a conductance determined by the passive properties of the rest of the channel. The rapid increase of conductance between -60 and -30 mV corresponds to a rapidly increasing probability that the gate is open.

Suppose that each channel has a gate with two states: open and closed. When there is no average electric field in the membrane ($v = 0$), the energy of the open state is $w = u_o k_B T$ greater than the closed state. Suppose also that as the gate opens and closes, a charge q associated with the gate moves a small distance parallel to the axis of the pore. When there is a potential v across the membrane, the charge moves through a potential difference αv, where $\alpha < 1$. The total energy change when the gate opens with potential v across the membrane is then $w + q\alpha v$. The quantity $q\alpha$ is often written as ze and called the *equivalent gating charge*. In terms of $k_B T$, the energy change when the pore opens is $u = u_o + zev/k_B T$.

Let p_o be the probability that a pore is open and p_c be the probability that it is closed. The probabilities are related by a Boltzmann factor: $p_o = p_c e^{-u}$. Since $p_o + p_c = 1$, $p_o = e^{-u}/(1 + e^{-u}) = 1/(1 + e^{u})$,

$$p_o = \frac{1}{1 + e^{u_o + zev/k_B T}}. \tag{9.57}$$

For very large values of u (small values of p_o),

$$p_o \approx e^{-(u_o + zev/k_B T)}. \tag{9.58}$$

The conductance per unit area of the membrane is the conductance of an open pore times the number of pores per unit area (that is g_∞), times p_o. Figure 9.15 shows plots of the "data" and lines generated from Eq. 9.57. The multiplicative constant has been adjusted to fit the flat region of the "data" at high v. Parameters u_o and z have been adjusted to provide good fits at the lowest conductances. For sodium $u_o = -10.5$ and $z = -7$; for potassium $u_o = -19$ and $z = -10$. The fact that u_o is very negative means that when $v = 0$ the energy of an open gate is much less than the energy of a closed gate. Nearly all of the pores are open, as can be seen from the $v = 0$ point in Fig. 9.15. The fact that $z = -7$ or -10 means that when the pore opens or closes the equivalent of 7 (or 10) electron charges must move through the full transmembrane potential difference. Many more charges could be displaced a much smaller distance and experience a much smaller potential change. More sophisticated multilevel models are discussed by Sigworth (1993).

This charge movement constitutes a very small current called the *gating current*. It is different from the current to charge the membrane capacitance. We saw above that during a 1-pA pulse lasting 1 ms, about 6000 monovalent ions flow through the membrane. The gating charge is about 10 monovalent charges, a ratio of about 600. The gating current is so small that it has not yet been measured in a single channel, but it can be measured by manipulating the ions bathing the membrane in a patch-clamp experiment. Figure 9.16 shows the results of a set of experiments with *Shaker* potassium channels. Panel A shows the macroscopic depolarizations to $+20$ and $+80$ mV for a patch with about 400 channels. The peak current at $+80$ mV is 1.25 pA per channel. Panel B shows the gating current recorded from another patch containing about 8000 channels. Potassium was removed from the solution bathing the interior surface of the membrane. The gating current lasts slightly less than 1 ms and peaks at about 4.5×10^{-15} A per channel, about 300 times less than the channel current. The agreement with our first estimate of 600 times less is satisfactory, given the accuracy of the data. Panels C and D show recordings similar to panel A, but with only a few channels in the patch. The results from three successive depolalrizing pulses are shown in each case. The channel openings are similar to those in Fig. 9.13, but are recorded at a much shorter time scale. The increased current through an open channel and the higher probability of being open for a clamp of $+80$ mV are both apparent. The smooth macroscopic current shown in Fig. 9.16a is the sum of many discrete channel currents like those shown in Fig. 9.16c.

A very simple approximate calculation shows that there is not much ion–ion interaction in a channel. A current of 1 pA is 6.25×10^6 monovalent ions per second, so that the average time between the passage of successive ions through the channel is 1.6×10^{-7} s. In a uniform electric field giving

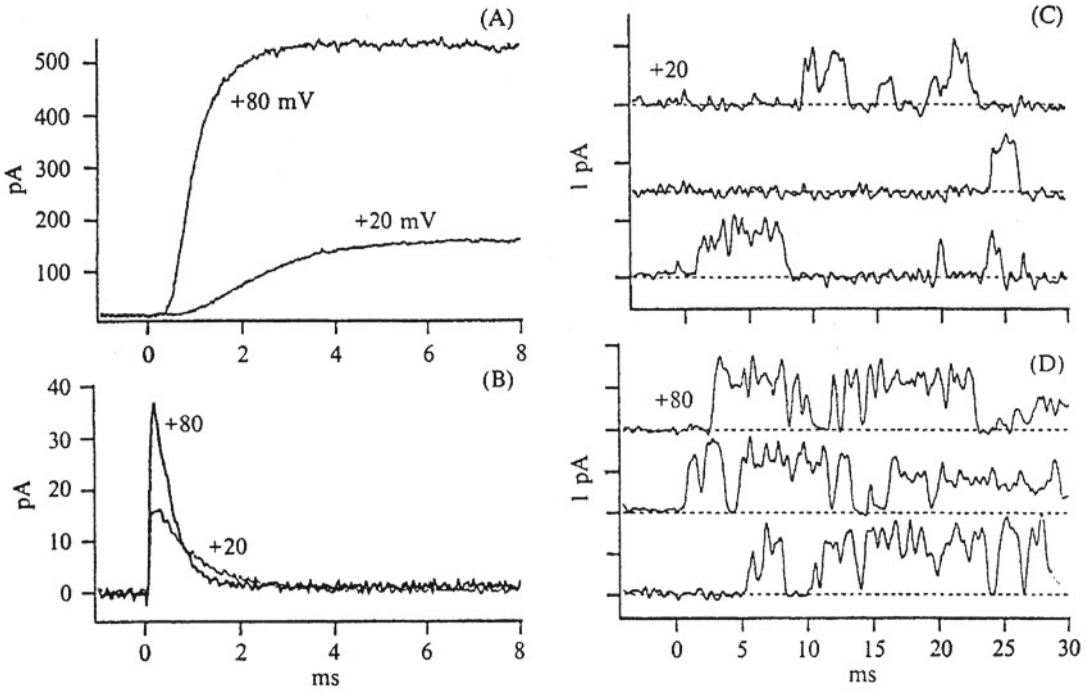


Fig. 9.16 The results of a set of experiments with *Shaker* potassium channels. Panel A shows the macroscopic depolarizations to +20 and +80 mV for a patch with about 400 channels. Panel B shows the gating current recorded from another patch containing about 8000 channels. Potassium was removed from the solution bathing the interior surface of the membrane. Panels C and D show recordings similar to *panel A*, but with many fewer channels in the patch. The results from three successive depolarizing pulses are shown in each case. (From F. J. Sigworth (1993). Reprinted with permission of Cambridge University Press)

80 mV across the membrane, a monovalent ion would have a drift velocity of 0.6 m s^{-1} based on the bulk diffusion constant. (See the discussion surrounding Eq. 4.22.) As ions in the pore are confined, let us use $\frac{1}{10}$ of this, or 0.06 m s^{-1}. (The diffusion constant is proportional to the solute permeability; see Sect. 5.9. Ignoring electric forces, we see from Fig. 5.16 that $\omega/\omega_0 = 0.1$ corresponds to $a/R_p = 0.4$. So this is probably still a high drift velocity.) Then the time it takes the ion to pass through the channel is its length (assume 6 nm) divided by the average speed, or 10^{-7} s. The fraction of the time there is an ion in the channel is $f = 0.625$.

We can make some other estimates of channel parameters. Over some part of its length, the channel must be narrow enough so the wall can interact directly with the ion that is passing through, not shielded by water molecules. The pore must therefore be narrowed to a radius of 0.3 to 0.7 nm in some region. Let us assume a cylindrical pore of radius $a = 0.7$ nm and length $h = 6$ nm. The average number of water molecules in the channel is 308; the average concentration of ions is $f/(\pi a^2 h) = 113$ mmol l^{-1}, which is about right. The resistance of a channel while it is open is $R = v/i = 80$ mV/1 pA $= 8 \times 10^7$ Ω. (We should actually use $v - v_{\text{Nernst}}$, but this is a rough estimate. If we were going to be more accurate, we should also use the Nernst–Planck equation, recognizing that the ions move by diffusion as well as drift.)

9.8 Noise

The current fluctuates while a channel is open, as can be seen in Figs. 9.13 and 9.16. Some of the fluctuation is due to noise in the measurement apparatus. However, there are some fundamental physical lower limits to the fluctuations resulting from noise in the membrane patch itself. We discuss these briefly here, with a more extensive discussion in Chap. 11. DeFelice (1981) wrote an excellent book on noise in membranes.

9.8.1 Shot Noise

The first (and smallest) limitation is called *shot noise*. It is due to the fact that the charge is transported by ions that move randomly and independently through the channels. Imagine a single open conducting channel with an average current $\bar{i}$ of monovalent ions. During time Δt (which can be any interval

shorter than the time the channel is open), the average charge flow is $\bar{i}\Delta t$ and the average number of ions is $\overline{n} = \bar{i}\Delta t/e$. Since there are a very large number of ions that might flow through the channel (occurrences) and the probability that any one ion moves through the channel during Δt is very small, we have the Poisson limit of the binomial distribution (Appendix J). The variance in the number of ions is $\sigma_n^2 = \overline{n} = \bar{i}\Delta t/e$. Since the average charge transported is $\overline{q} = \overline{n}e$, the variance in the charge is $\sigma_q^2 = e^2\sigma_n^2 = e\bar{i}\Delta t$. When many samples of length Δt are measured, the variance in the current is $\sigma_i^2 = \sigma_q^2/(\Delta t)^2$. The standard deviations are

$$\begin{aligned} \sigma_n &= \left(\frac{\bar{i}\Delta t}{e}\right)^{1/2}, \\ \sigma_q &= \left(e\bar{i}\Delta t\right)^{1/2}, \\ \sigma_i &= \left(\frac{e\bar{i}}{\Delta t}\right)^{1/2}, \end{aligned} \tag{9.59}$$

and the fractional standard deviations are

$$\frac{\sigma_n}{\overline{n}} = \frac{\sigma_q}{\overline{q}} = \frac{\sigma_i}{\bar{i}} = \left(\frac{e}{\bar{i}\Delta t}\right)^{1/2}. \tag{9.60}$$

For a current of 1 pA, the fractional standard deviation is 0.013 when the sampling time is 1 ms and 0.04 when the sampling time is 0.1 ms. These are much smaller than what is observed in the figures.

9.8.2 Johnson Noise

The next source of noise is called *Johnson noise*. It arises from thermal fluctuations or Brownian movement of the ions. It can be derived from a microscopic model of conduction (either in an ionic solution or a metal), but we will do it using the equipartition of energy.

First, we need an expression for the energy U contained in a charged capacitor. To obtain it, imagine that an amount of charge $+dq$ is transferred from the negative to the positive conductor. This increases the amount of positive charge on the positive conductor and also increases the amount of negative charge on the negative conductor. The work required to transfer the charge when the potential difference between the conductors is v is $v dq$. The energy stored in the capacitor is the total work required to charge the conductor from 0 to q. Remembering that $q = Cv$, we have

$$U = \int_0^q v\, dq = \frac{1}{C}\int_0^q q\, dq = \frac{q^2}{2C} = \frac{Cv^2}{2}. \tag{9.61}$$

If the capacitor is completely isolated, there can be a constant charge on each conductor with no fluctuations. If the capacitor is in thermal contact with its surroundings and is in equilibrium, then the equipartition theorem applies. The capacitor can be brought into thermal equilibrium with its surroundings by connecting a resistance R between the conductors. This will discharge the capacitor so $\overline{q} = \overline{v} = 0$. There will be fluctuations around these zero values. As the expression for the energy depends on the square of the variables, the mean square value is given by the equipartition of energy theorem, Eq. 3.38. We will assume that when the capacitor is charged, thermal fluctuations give the same variances as when it is discharged:

$$\sigma_v^2 = \left(\overline{v^2} - \overline{v}^2\right) = \overline{v^2} = k_B T/C, \tag{9.62a}$$

$$\sigma_q^2 = \left(\overline{q^2} - \overline{q}^2\right) = \overline{q^2} = Ck_B T. \tag{9.62b}$$

In a simple RC circuit, $i = v/R$, so

$$\sigma_i^2 = \sigma_v^2/R^2 = k_B T/R^2 C. \tag{9.62c}$$

Since changes in current or voltage in an RC circuit occur with time constant $\tau = RC$, we can also write these as

$$\sigma_v^2 = Rk_B T/\tau, \quad \sigma_i^2 = k_B T/R\tau. \tag{9.63}$$

These are special cases of a more general relationship that will be discussed in Chap. 11.

We can use these to determine some of the requirements for patch-clamp recording. In order to see the current from a single channel with some accuracy, let us require that the standard deviation of the current fluctuation be less than $\frac{1}{8}$ of the signal we want to measure. (This signal-to-noise ratio, $SNR = 8$, is arbitrary.) First, consider the limitation due to Johnson noise. We want $\sigma_i < \bar{i}/8$ or $\sigma_i^2 < \left(\bar{i}\right)^2/64$. From this, we obtain

$$R > \left(\frac{k_B T}{C}\right)^{1/2} \frac{8}{\bar{i}}. \tag{9.64}$$

The capacitance of a patch of membrane of 1 μm radius is 3.1×10^{-14} F. At a temperature of 300 K and for an average current of 1 pA, this gives $R > 3 \times 10^9\ \Omega$. Larger values of R will give an even higher SNR. There are several sources of thermal noise in a recording electrode, all discussed in the paper by Hamill et al. (1981). These are order-of-magnitude results; one must determine carefully which capacitances and resistances provide the dominant effects.

We can also see when shot noise is important. The ratio of Johnson noise to shot noise is

$$\frac{\sigma_i^2(\text{Johnson})}{\sigma_i^2(\text{shot})} = \frac{k_B T/R\tau}{e\bar{i}/\Delta t} = \frac{k_B T}{Re\bar{i}}. \tag{9.65}$$

This ratio is less than 1 and shot noise is important when $R > k_B T/e\bar{i} = 2.6 \times 10^{10}\ \Omega$. Shot noise has been detected in channel gating currents and subjected to very sophisticated analysis. See the paper by Crouzy and Sigworth (1993) and the references therein.

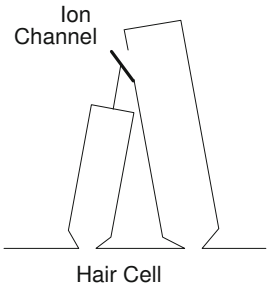


Fig. 9.17 A schematic diagram of two stereocilia linked by a filament that opens a channel as the cillia move back and forth

9.9 Sensory Transducers

Animals have very acute senses. We will see (Problem 13.19) that the ear can hear sounds at 1000 Hz that are just greater than the pressure fluctuations due to molecular collisions on the ear drum. An eye that is adapted to the dark can detect flashes of light corresponding to a few photons (Chap. 14). Many animals can smell chemicals when only a few molecules strike their sensory organs. The electric skate can detect extremely small electric fields. In each case a transducer converts the sensory stimulus into a series of nerve impulses. The transducer must have sufficient sensitivity to respond to the stimulus, and it must also absorb an amount of energy from the stimulus that is greater than what it receives from random thermal bombardment (Brownian movement).[5] We describe here two transducers: the mechanoreceptors (hair cells) of the inner ear and the electric organ of the skate.

Various transduction mechanisms are reviewed in Chap. 8 of Hille (2001). The mechanoreceptors of the bullfrog inner ear have been studied for many years. The hair-cell current rises from 0 to 100 pA with a 0.5-μm displacement. Each hair cell is cylindrical. On its end face are found about 60 very small *stereocilia*, each 1–50 μm long and with a 100–500-nm radius. The tips of these stereocilia are linked by thin filaments. The hair cells and stereocilia that detect sound in the ear are attached to the basilar membrane in the cochlea of the ear and move in a very viscous fluid as the basilar membrane vibrates. Hair cells detecting accelerations of the entire animal are attached to a suspended dense body. It is believed that as the stereocilia move, these filaments pull open flaps at the end of ion channels, allowing ions to enter the cell and initiate the conduction process. This is shown schematically in Fig. 9.17. Denk and Webb (1989) have used a laser interferometer to measure the motion of the hair cells. They found that the spontaneous motion consists primarily of thermal excitation (Brownian motion). Fluctuations in the intracellular voltage were also measured. They often correlated with the motion of the hair cells.

Freshwater catfish respond to electric fields as low as 10^{-4} V m^{-1}. Saltwater sharks and rays can detect fields of 5×10^{-7} V m^{-1}. A brief review has been given by Bastian (1994); Kalmijn (1988) provides a very complete review. The saltwater fish have a more complicated sensory apparatus than the freshwater fish, known as the *ampullae of Lorenzini*. Kalmijn et al. discovered that the ocean flounder generates a current dipole of 3×10^{-7} A m. Sea water with resistivity of 0.23 Ω m gives an electric field of 2×10^{-5} V m^{-1} at a distance 10 cm in front of the flounder. They were able to show in a beautiful series of behavioral experiments that dogfish (a small shark) could detect the electric field 0.4 m from a current dipole of 4×10^{-7} A m, corresponding to an electric field of 5×10^{-7} V m^{-1}. The fish would bite at the electrodes, ignoring a nearby odor source. A field of 10^{-4} V m^{-1} would elicit the startle response. A field $\frac{1}{10}$ as large caused a physiologic response. The animals responded to a constant field or a sinusoidally alternating field up to 4 Hz. At 8 Hz the threshold increased by a factor of 2.

In a series of experiments Lu and Fishman (1994) dissected out the ampulla of Lorenzini and measured its response in the laboratory. They found that the resting rate of firing of the organ is about 35 Hz (impulses per second) and that an applied electric potential increases or decreases the firing rate by about 1 Hz μV^{-1}, depending on its sign. The firing rate saturated for potential differences of 100 μV.

The behavioral experiments showed sensitivity to an electric field of 0.5 μV m^{-1}. The anatomy of the ampulla is such that the organ senses the potential difference between the surface of the fish and deep in its interior. Pickard (1988) shows that for a spherical fish of radius a, this gives a potential of $3aE/2$, where E is the external electric field. The amplitude of the potential difference oscillation of a fish of length $\frac{1}{3}$ m is therefore 0.25 μV. This is enough to cause a 1 % change in firing rate, which could be detected by neuronal circuits (Adair et al. 1998).

The Johnson noise is somewhat smaller than the signal detected. To estimate it, use Eq. 9.62a with the ampullary capacitance of 0.15 μF measured by Lu and Fishman. The standard deviation of the noise is 0.17 μV.

9.10 Possible Effects of Weak External Electric and Magnetic Fields

There is a lingering controversy over whether radio-frequency (cell phone, 450 MHz–2 GHz) and power-line-frequency (50–60 Hz) electric or magnetic fields can cause

[5] For the detection of light, the amount of energy per photon is so much greater than $k_B T$ that shot noise dominates.

cancer. While the effect, if any, is quite small, the literature is extensive, involving both epidemiological and laboratory studies. Results are conflicting, and the mechanisms by which such an effect might occur are not yet understood. Mechanisms have been proposed, some of which are inconsistent with basic physical principles such as the Boltzmann factor, the mean free path of ions, and thermal fluctuations. A review in the physics-teaching literature was provided by Hafemeister (1996).

It is beyond our scope to do more than provide pointers to the field and discuss some basic underlying physics. We consider here the power-line frequencies.

We have seen that electric charges give rise to electric fields, and moving electric charges (currents) generate magnetic fields. The electric field lines start and end on charges, and the magnetic field lines surround the currents. We will see in Chap. 14 that accelerated charges generate electromagnetic *radiation*, in which the electric and magnetic fields are interrelated and the field lines close on themselves far from the source charges. Energy is radiated; it leaves the source and never returns. This radiated energy is in the form of discrete packets or *photons*, whose energy is related to the frequency of oscillation of the fields. The energy of each photon is $E = h\nu$, where h is Planck's constant and ν is the frequency. At room temperature, the energy of random thermal motion is $k_B T = 4 \times 10^{-21}$ J. At 60 Hz, the energy in each photon is much smaller: 4×10^{-32} J. At 100 MHz, it is 7×10^{-26} J. For electromagnetic radiation in the ultraviolet and beyond, which certainly can harm cells, the photon energy is 5×10^{-19} J or greater, quite large compared to $k_B T$. At the very low frequencies we are considering it is the strength of the electric or magnetic field that is important, not the energy of individual photons. A more detailed discussion of the distinction between these low-frequency "near fields" and "radiation fields" is found in Polk (1996).

9.10.1 Strong Fields

Electrical burns, cardiac pacing, and nerve and muscle stimulation are produced by electric or rapidly changing magnetic fields. Even stronger electric fields increase membrane permeability. This is believed to be due to the transient formation of pores (*electroporation*). Pores can be formed, for example, by microsecond-length pulses with a field strength in the membrane of about 10^8 V m^{-1} (Weaver 2000). Microwaves are used to heat tissue. Nerve stimulation requires a few millivolts across the cell membrane, or about 10^5–10^6 V m^{-1}. Both electric and magnetic fields are used to promote bone healing, with field strengths in tissue in the fracture region of 10^{-1} V m^{-1} (Tenforde 1995), though these results are controversial (Adair 2000).

9.10.2 Power Frequency (50–60 Hz) Fields

9.10.2.1 Fields in Homes are Weak

Much weaker fields in homes are produced by power lines, house wiring, and electrical appliances. Barnes (1995) found average electric fields in air next to the body of about 7 V m^{-1}, with peak values of 200 V m^{-1}. (We will find that since the body is a conductor, the fields within the body are much less.) Average residential magnetic fields are about 0.1 μT, with peaks up to four times as large. Within the body they are about the same. Tenforde (1995) reviews both power-line and radio-frequency field intensities.

9.10.2.2 Epidemiological Studies

Epidemiological studies have been very valuable in tracing the cause of infectious outbreaks. They have also indicated that smoking increases the probability of developing lung cancer by 3000 %—a factor of 30. However, there are difficulties with epidemiological studies when the effect is small: there are inescapable statistical fluctuations unless the number of subjects is huge; associations do not prove causality; and there may be unrecognized variables that are confusing the picture. The problem is exacerbated when positive findings receive widespread publicity and negative findings are ignored by the press.

Epidemiological studies usually report *relative risk*: the incidence in an exposed group divided by the incidence in an unexposed group. A relative risk of one means no effect. John Moulder, the author of a web site about power lines and cancer that unfortunately no longer exists, said,

> A strong association is one with a relative risk (RR) of 5 or more. Tobacco smoking, for example, shows a strong association, with the risk of lung cancer in smokers being 10–30 times that of non-smokers. A relative risk of less than about 3 indicates a weak association. A relative risk below about 1.5 is essentially meaningless unless it is supported by other data.
>
> Most of the positive power-frequency studies have relative risks of two or less. The leukemia studies as a group have relative risks of 0.8-1.9, while the brain cancer studies as a group have relative risks of 0.8-1.6. This is a weak association. Interestingly, as the sophistication of the studies has increased, the relative risks have not increased.

One would also expect an increased response with increasing dose. Moulder continued,

> No published power-frequency exposure study has shown a statistically-significant dose-response relationship between measured fields and cancer rates, or between distances from transmission lines and cancer rates. However, there is some indication of a dose-response in some of the older childhood leukemia studies when wire codes or calculations of historic fields are used as exposure metrics. The lack of a clear relationship between exposure and increased cancer incidence is a major reason why most scientists are skeptical about the significance of much of the epidemiology.

9.10.2.3 Laboratory Studies

The many laboratory studies were also reviewed by Moulder. He concluded:

> Power-frequency fields show little evidence of the type of effects on cells, tissues or animals that point towards their being a cause of cancer, or to their contributing to cancer. In fact, the existing laboratory data provides strong evidence that power-frequency fields of the magnitude to which people are exposed are not carcinogenic.[6]

9.10.2.4 Reviews and Panel Reports

Reviews by Moulder and Foster (1995, 1999) find that the association between power-frequency fields and cancer is weak[7] for magnetic fields and even weaker for electric fields. Carstensen (1995) and Bren (1995) reach similar conclusions.

A report by a committee of the National Research Council concludes that

> the current body of evidence does not show that exposure to these fields presents a human-health hazard.... The committee reviewed residential exposure levels to electric and magnetic fields, evaluated the available epidemiological studies, and examined laboratory investigations that used cells, isolated tissues, and animals. (National Research Council (1997), p. 2)
>
> There is no convincing evidence that exposure to 60-Hz electric and magnetic fields causes cancer in animals.... There is no evidence of any adverse effects on reproduction or development in animals, particularly mammals, from exposure to power-frequency 50- or 60-Hz electric or magnetic fields. (National Research Council 1997, p. 7).

9.10.2.5 Electric Fields in the Body

We now review some of the basic principles that govern the interaction of electric and magnetic fields with the body. One of the important principles is the relationship between the electric field in air and the field within the body, which is a conductor. A simple model that shows how this coupling takes place is the one-dimensional problem shown in Fig. 9.18. An infinite slab of tissue has dielectric constant κ and conductivity σ. In the air perpendicular to the surface of the slab is an external oscillating electric field $E(t) = E_0 \cos \omega t$. We assume that the dielectric constant is independent of frequency and accounts for the polarization

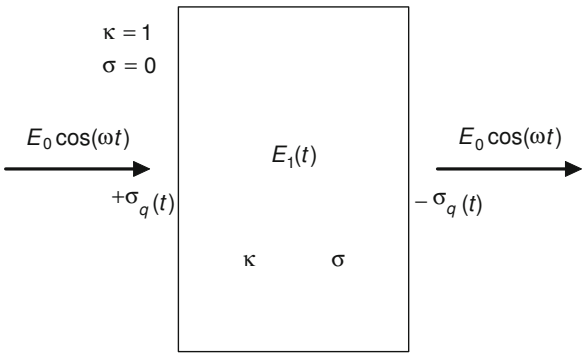


Fig. 9.18 An infinite slab of tissue is immersed in an oscillating electric field of amplitude E_0 in air

of the tissue. An ionic current flows and causes free charge per unit area $\pm\sigma_q$ to accumulate on the surfaces of the slab. Within the slab, the field is $E_1(t)$ and the current density is $j = \sigma E_1$. Gauss's law (Eq. 6.21b) applied to either surface gives

$$-\epsilon_0 E_0 \cos \omega t + \kappa \epsilon_0 E_1(t) = \sigma_q(t). \tag{9.66}$$

Conservation of free charge at the surface requires that[8]

$$\sigma E_1 = j = -\frac{d\sigma_q}{dt}. \tag{9.67}$$

If we differentiate Eq. 9.66 and combine it with Eq. 9.67, we obtain

$$\frac{dE_1}{dt} + \frac{\sigma}{\kappa\epsilon_0} E_1 = -\frac{\omega}{\kappa} E_0 \sin \omega t. \tag{9.68}$$

The factor $\kappa\epsilon_0/\sigma$ is a characteristic of the tissue and has the dimensions of time. We will call it τ_t.[9] Typical tissue conductivity is about 0.1 S m^{-1}. We must be careful about the value of the dielectric constant. We have used a value of 80 for water. However, tissue is much more complex than pure water and there are several effects that alter the dielectric constant (Foster and Schwan 1996). It takes time for both the polarization charges and conducting ions to move. As a result, both the conductivity and the dielectric constant of tissue depend on the frequency of the applied electric field and in fact are not independent of one another (see Foster and Schwan 1996, especially pp. 31–41). Several effects change

[6] Foster (1996) reviewed many of the laboratory studies and described cases where subtle cues meant the observers were not making truly "blind" observations. Though not directly relevant to the issue under discussion here, a classic study by Tucker and Schmitt (1978) at the University of Minnesota is worth noting. They were seeking to detect possible human perception of 60-Hz magnetic fields. There appeared to be an effect. For 5 years they kept providing better and better isolation of the subject from subtle auditory clues. With their final isolation chamber, none of the 200 subjects could reliably perceive whether the field was on or off. Had they been less thorough and persistent, they would have reported a positive effect that does not exist.

[7] That is, the carcinogenic effects are in International Association for Research on Cancer group 2B (possibly carcinogenic), a group that includes coffee and pickled vegetables.

[8] Readers who are familiar with the concepts of reactance and complex impedance must be frustrated because we have not used them. The reason is pedagogic. Because many in our intended audience may have had only one year of calculus, we want to avoid the use of complex numbers. In Chap. 11 we introduce them as a parallel notation. They are widely used in the image reconstruction described in Chap. 12.

[9] Recall that the membrane time constant τ was used in Eq. 6.40. The values of conductivity or resistivity and dielectric constant are different in this case.

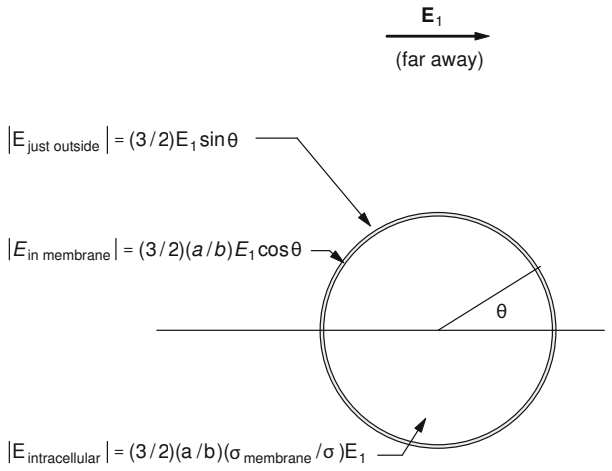


Fig. 9.19 The electric fields in and around a *spherical cell*. The *cell* has radius a and membrane thickness b. The field far from the *cell* has an amplitude E_1

the conductivity and dielectric constant as a function of frequency. At power-line frequencies, the dominant effect is the slight movement of the counterions and charge in the double layer at a cell membrane in response to the applied electric field. As a result, $\kappa \approx 10^6$ and $\tau_t = 9.1 \times 10^{-5}$ s.

We try a solution to Eq. 9.68 of the form $E_1(t) = A \sin \omega t + B \cos \omega t$. It satisfies the equation if

$$A = -\frac{\omega\tau_t}{\kappa(1+\omega^2\tau_t^2)}E_0 \approx -\frac{\omega\epsilon_0}{\sigma}E_0,$$
$$B = -\omega\tau_t A = -\frac{(\omega\tau_t)^2}{\kappa(1+\omega^2\tau_t^2)}E_0 \approx 0. \tag{9.69}$$

For 60 Hz and a dielectric constant of 10^6, $A = 33\times10^{-9}E_0$, $B = 1.1 \times 10^{-9}E_0$. The amplitude of the field in tissue is $E_1 \approx A$:

$$E_1 \approx 33 \times 10^{-9}E_0. \tag{9.70}$$

The field in air is reduced by a factor of about 3×10^{-8} in tissue because the tissue is a good conductor. The total reduction is nearly the same for a dielectric constant of 80, as can be seen from the fact that the approximate form for A does not depend on κ.

9.10.2.6 Electric Fields in a Spherical Cell

Another important factor is the electric fields that exist in and near a cell. Consider a spherical cell with inner radius a and membrane thickness b immersed in an infinite conducting medium in which there is an electric field E_1 far from the cell. We saw above that a field in air of $E_0 = 300$ V m^{-1} is reduced to $E_1 = 10^{-5}$ V m^{-1} in the conducting medium. The potential can be determined analytically by solving Poisson's equation (with zero charge density) in the three regions and matching boundary conditions much as we did to obtain Eq. 9.67. The results, valid for slowly varying applied fields such as a 50 or 60-Hz power line field, are shown in Fig. 9.19.[10] Only the amplitude of the electric field is shown. Assume the conductivities σ of the extracellular and intracellular fluids are the same, that $a = 10$ μm and $b = 6$ nm, and that $\sigma_{\text{membrane}} = 2.4 \times 10^{-8}\sigma$. The important features of this solution are that the field just outside the cell is roughly the same as the field far away, the field inside the membrane is magnified by a large factor (a/b), and the field inside the cell is multiplied by a very small factor $(a\sigma_{\text{membrane}}/b\sigma)$. Thus, the cell membrane shields the intracellular space from extracellular electric fields, so these fields are not likely to directly affect cell organelles and important biomolecules such as DNA. This is reflected in the last line of Table 9.5.

Table 9.5 Comparison of the signal in a cell to thermal noise for an applied electric field in air $E_0 = 300$ V m^{-1}. From Eq. 9.71, $E_1 = 10^{-5}$ V m^{-1}. $T = 300$ K. $z = 10$. $d = 10^{-5}$ m

Model	Outside the cell	In the cell membrane	Inside the cell
E (V m^{-1})	1.0×10^{-5}	1.62×10^{-2}	5.40×10^{-10}
k_BT/eE (m)	2.57×10^{3}	1.59	4.79×10^{7}
$zeEd/k_BT$	3.9×10^{-8}	6.3×10^{-5}	2.1×10^{-12}

9.10.3 Electrical Interactions and Noise

If an organism is affected in some way by an external field, then it can be regarded as a detector of that field. The external field can therefore be thought of as a signal. To be detected, the signal must be greater than the noise. The noise can be either thermal (Johnson) noise, shot noise, or noise from the electric currents that normally flow in the body due to nerve conduction and muscle contraction. To have a signal that is not masked by Johnson noise, we must have an electric field E such that

$$\frac{zev}{k_BT} = \frac{zeEd}{k_BT} > 1, \tag{9.71}$$

where z is the valence of an effective charge that moves a distance d in the electric field E. Table 9.5 shows the result of a calculation using a field in air of 300 V m^{-1}. We use a value $z = 10$. For d, we use the diameter of the cell, $d = 10$ μm (though for the membrane perhaps the much smaller thickness of the cell membrane should be used). The values of $zeEd/k_BT$ are very small.

One proposal to overcome this signal-to-noise problem is that the biological effect is due to the averaging of the field over many cells or over time. This was first proposed by Weaver and Astumian (1990), and a specific model has been

[10] Calculated using equations in Polk (1995), p. 62.

formulated by Astumian et al. (1995). The model applies the Nernst–Planck equation (Eq. 9.37) and shows that if the concentration of some substance outside the cell is much larger than inside, the response to an oscillating v is "rectification" or a net inward current. This would allow an accumulation of the substance within the cell. The averaging times in their model are 13 h. Weaver and Astumian (1995) review the entire causality problem, including the effects of shot noise. Adair (2000) reviews many other aspects of the problem.

9.10.4 Magnetic Interactions and Noise

The magnetic field is not attenuated at the body surface like the electric field is. Kirschvink et al. (1992a) reported that the human brain contains several million magnetosomes per gram. Kobayashi et al. (1995) found that contamination with magnetic particles could affect laboratory experiments with cell cultures, even if the cells being studied do not normally contain magnetosomes. Commercial disposable, presterilized plastic laboratory ware used in tissue culture experiments was found to contain ferromagnetic particles smaller than 100 nm that are readily taken up by white blood cells.

What about the signal-to-noise ratio for magnetic effects? The situation is somewhat more favorable than for the electric case. We saw in Chap. 8 that a single magnetosome has appreciable alignment with the earth's magnetic field, even in the presence of thermal bombardment. The earth's field is about 5×10^{-5} T. For a single magnetosome

$$\frac{m B_{\text{earth}}}{k_B T} = \frac{(6.4 \times 10^{-17})(5 \times 10^{-5})}{(1.38 \times 10^{-23})(300)} = 0.77. \qquad (9.72)$$

For a larger magnetosome of radius 100 μm, $m = 2 \times 10^{-15}$ A m^2 and the energy ratio in the earth's field is 24. The field due to a typical power line is about 100 times smaller: about 2×10^{-7} T.

Kirschvink (1992) proposed a model whereby a magnetosome in a field of 10^{-4}–10^{-3} T could rotate to open a membrane channel. As an example of the debate that continues in this area, Adair (1991, 1992, 1993, 1994) argued that a magnetic interaction cannot overcome thermal noise in a 60-Hz field of 5×10^{-6} T. However, Polk (1994) argued that more biologically realistic parameters, including a large number of magnetosomes in a cell, could allow an interaction at 2×10^{-6} T.

The essential features of all the models are like this. Imagine a particle with magnetic moment **m** in the earth's field. It will tend to align with the field as shown in Fig. 9.20(a). The direction of the magnetic moment with the earth's field is θ. Apply an alternating field $B_0 \cos \omega t$ at right angles to the earth's field, as shown in Fig. 9.20(b). There are three

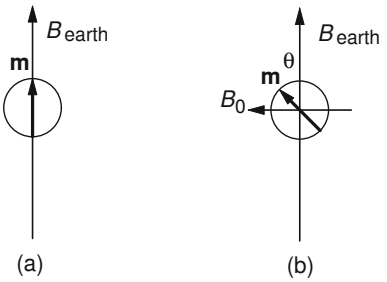


Fig. 9.20 A particle with magnetic moment **m** a aligned with the earth's magnetic field and **b** at an angle θ with the earth's field because of an applied field B_0

torques on the particle. The first is viscous drag, which is proportional to the angular velocity of the particle $d\theta/dt$ but in the opposite direction. The second is the torque tending to align **m** with the earth's field, $-m B_{\text{earth}} \sin\theta$. The third tends to align **m** with the alternating field, $m B_0 \cos \omega t \cos\theta$. Assume that the acceleration is so small that the particle is in rotational equilibrium. (This is not necessary, but it simplifies the math.) Then, from Eq. 1.4,

$$-\beta \frac{d\theta}{dt} + m B_{\text{earth}} \sin\theta - m B_0 \cos \omega t \cos\theta = 0. \qquad (9.73)$$

In order to linearize the equation, assume that θ is small enough so that $\sin\theta \approx \theta$ and $\cos\theta \approx 1$. The linearized equation is

$$\beta \frac{d\theta}{dt} - m B_{\text{earth}} \theta = -m B_0 \cos \omega t. \qquad (9.74)$$

This is a linear differential equation with constant coefficients that can be solved by the techniques of Appendix F. Consider only the particular solution and try a solution of the form

$$\theta = \theta_1 \cos \omega t + \theta_2 \sin \omega t. \qquad (9.75)$$

Substitution of this in the equation shows that

$$\theta_1 = \frac{m^2 B_0 B_{\text{earth}}}{(\omega\beta)^2 + (m B_{\text{earth}})^2},$$

$$\theta_2 = -\frac{\omega\beta m B_0}{(\omega\beta)^2 + (m B_{\text{earth}})^2},$$

and

$$\theta_m = \frac{m B_0}{\left[(\omega\beta)^2 + (m B_{\text{earth}})^2\right]^{1/2}}, \qquad (9.76)$$

where θ_m is the maximum amplitude: $\theta_m^2 = \theta_1^2 + \theta_2^2$. We saw in Chap. 4 (Stokes' law) that the translational viscous drag on a spherical particle is $6\pi\eta a v$. Similarly, the viscous torque on a rotating sphere is $8\pi\eta a^3 (d\theta/dt)$ (Lamb 1932, pp. 588-589). The measured values for viscosity inside a cell range

from 0.003 to 0.015 N s m^{-2} (Polk 1994). Using the average of these, $\beta = 0.009(8\pi)a^3 = 0.23a^3$. The magnetic moment of a single-domain magnetosome is also proportional to volume: $m = 2 \times 10^6 a^3$. This leads to a maximum amplitude that is independent of a:

$$\theta_m = \frac{2 \times 10^6 a^3 B_0}{\left[(377)^2(0.23)^2 a^6 + (2 \times 10^6)^2(5 \times 10^{-5})^2 a^6\right]^{1/2}} = 1.5 \times 10^4 B_0. \quad (9.77)$$

Kirschvink originally argued from data about hair-cell deformation that a deflection of 16° or 0.3 rad is needed. This would require $B_0 = 2 \times 10^{-5}$ T. (He had a slightly different value because he used a different viscosity. He also included the torque due to the force on the channel gate.)

In the absence of the applied field, the thermal fluctuations in angle can be estimated as follows. In the linear approximation, the work required to displace the particle an amount θ from the direction of the earth's field is

$$W = \int \tau \, d\theta = \int m B_{\text{earth}} \theta \, d\theta = m B_{\text{earth}} \frac{\theta^2}{2}. \quad (9.78)$$

Equipartition of energy again gives us

$$\overline{\theta^2}_{\text{thermal}} = \frac{k_B T}{m B_{\text{earth}}} = \frac{k_B T}{(2 \times 10^6)(5 \times 10^{-5}) a^3} = \frac{k_B T}{100 a^3}. \quad (9.79)$$

For a 50-nm magnetosome, this gives $\theta_{\text{rms}} = 0.58$ rad. For a 100-nm magnetosome it is 0.2 rad, comparable to the maximum angles deduced from the model in the preceding paragraph.

9.10.5 Microwaves, Mobile Phones, and Wi-Fi

Many of the concerns about the effects of power-line fields on the body also apply to radio-frequency fields. Sources include radio waves, microwaves, mobile phones, and Wi-Fi devices. A sample of the controversy surrounding this issue can be found in Khurana et al. (2008)[11]

A review by Moulder et al. (2005) concluded that "Overall, a weight-of-evidence evaluation shows that the current evidence for a causal association between cancer and exposure to RF energy is weak and unconvincing." However, they pointed out that there have been only a few epidemiological studies (which overall show no association). Moreover, the energy deposited in a small region of the head by a cell phone may be only an order of magnitude less than the exposure guideline (10 W m^{-2}). While the studies they review did not suggest that RF energy is a primary carcinogen, they could not rule out the possibility that RF energy could enhance the carcinogenicity of other agents.

An exhaustive (390 page) report has been prepared by the International Committee on Non-ionizing Radiation Protection (Vecchia et al. 2009). They point out that heating of tissue by RF energy is well understood. The plausibility of effects by nonthermal mechanisms that have been proposed is very low. Epidemiological studies at the time of publication "give no convincing evidence of a causal relation between RF exposure and any adverse health effect." However, "these studies have too many deficiencies to rule out an association." As for mobile phone use and brain tumors, "overall the studies published to date do not demonstrate a raised risk within approximately 10 years of use for any tumor of the brain or any other head tumor." "For slow-growing tumors... the current observation period is still too short. Currently data are completely lacking on the potential carcinogenic effects of exposures in childhood or adolescence."

Sheppard et al. (2008) evaluated all the proposed mechanisms for radio-frequency interactions with biological molecules and processes They concluded, "an examination of all generally accepted and proposed mechanisms open to quantitative analysis shows that in the frequency range from several megahertz to a few hundred gigahertz, the focus of this paper, the principal mechanism for biological effects, and the only well-established mechanism, is the heating of tissues by dielectric and resistive loss."

In recent years, the use of Wi-Fi to connect computers and household appliances to the Internet has become common. It has become a public concern about possible health effects, particularly in schools. Foster and Moulder (2013) reviewed the current state of research. They concluded that the levels of RF exposure in a house are far below international and US limits. The engineering aspects are well understood. The biological studies are difficult to interpret "but provide no basis to anticipate any biological effects...." They observe

> Finally, it is noted that Wi-Fi and WLANs[12] can raise immediate and urgent safety issues apart from possible RF bioeffects [such as]... privacy invasion and hacking. The Internet... raises a number of safety issues (particularly with children) that have nothing to do with RF exposure. Excessive concern about speculative hazards from RF exposures to Wi-Fi, without concern for these more immediate potential hazards, is comparable to worry about the health effects of using mobile phones without concern for the hazards of texting while driving.

[11] Each issue of the journal *Medical Physics* contains one Point/Counterpoint article, in which a proposition is stated and two prominent medical physicists debate it, one for and one against. You can download all the point/counterpoint articles at http://www.medphys.org. They are a great resource to use when teaching medical physics.

[12] Wireless local area networks.

Symbols Used in Chapter 9

Symbol	Use	Units	First used page
a	Radius	m	245
b	Spacing	m	245
d	Displacement of charge	m	259
e	Electron charge	C	239
f	Force	N	247
f	Fraction of time an ion is in a channel		254
g_{K}	Potassium conductance per unit area	S m^{-2}	252
h	Length of cylindrical channel	m	254
h	Planck's constant	J s	257
j, $\overline{j}$	Electric current density	A m^{-2}	248
j_s	Particle current density	m^{-2} s^{-1}	247
j_v	Volume current density	m s^{-1}	247
k_B	Boltzmann constant	J K^{-1}	239
$\mathbf{m}$	Magnetic moment	A m^2	260
n	Number of ions		254
p_e, $\mathbf{p}_e$, p_{tot}	Electric dipole moment	C m	246
p, p_c, p_o	Probability		253
q	Charge	C	245
r, $\mathbf{r}$	Position	m	241
r	Radius in cylindrical coordinates	m	249
r	Radius in spherical coordinates	m	244
t	Time	s	254
u	$rv(r)$	V m	244
u, u_o	Energy (normalized to k_BT)		248
v, v'	Potential	V	239
v_{Nernst}	Nernst potential	V	248
w	Energy	J or eV	253
x	Position	m	241
x	Distance along cylindrical axis	m	249
z	Valence		239
A, B, A', B'	Constants	V	243
B_{earth}	Earth's magnetic field	T	260
B_0	Amplitude of applied oscillating magnetic field	T	260
C, C'	Concentration	m^{-3}	239
C_i	Concentration of species i	m^{-3}	242
[Cl], [Cl$'$]	Chloride concentration	m^{-3}	240
C	Capacitance	F	255
D, D_{eff}, D_0	Diffusion constant	m^2 s^{-1}	247
E, E_x, E_0, E_1	Electric field	V m^{-1}	241
E_{ext}	External electric field	V m^{-1}	246
E_{pol}	Polarization electric field	V m^{-1}	246
E	Photon energy	J	257
F	Faraday constant	C mol^{-1}	239
F, $\mathbf{F}$	Force	N	247
G	Conductance	S	247
J	Current per unit area of membrane	A m^{-2}	249
[K], [K$'$]	Potassium concentration	m^{-3}	240
L	Separation	m	247
[M$^+$], [M$^{+\prime}$]	Concentration of impermeant cations	m^{-3}	240
[M$^-$], [M$^{-\prime}$]	Concentration of impermeant anions	m^{-3}	240
[M], [M$'$]	Net concentration of impermeant ions	m^{-3}	240
N	Number per unit volume	m^{-3}	246
N_A	Avogadro's number	mol^{-1}	246
[Na], [Na$'$]	Sodium concentration	m^{-3}	240
P	Polarization	C m^{-2}	246
R	Gas constant	J K^{-1} mol^{-1}	239
R	Resistance	Ω	247
R_p	Pore radius	m	249
S	Area	m^2	241
T	Temperature	K	239
U, W	Energy	J	255
V	Particle velocity	m s^{-1}	247
α	Proportionality constant		253
β	Linear viscous drag coefficient	N s m^{-1}	247
β	Rotational viscous drag coefficient	N s m	260
ϵ_0	Electrical permittivity of free space	C^2 N^{-1} m^{-2}	241
κ	Dielectric constant		241
η	Coefficient of viscosity	Pa s	260
λ_D	Debye length	m	242
λ	Characteristic length	m	248
ν	Frequency	Hz or s^{-1}	257
ρ, ρ_{ext}	Charge density	C m^{-3}	241
ρ	Resistivity	Ω m	247
σ_q, σ_q'	Charge per unit area	C m^{-2}	246
σ	Conductivity	S m^{-1}	247
σ_i	Standard deviation of current	A	255
σ_n	Standard deviation of number of ions		255
σ_q	Standard deviation of charge	C	255
σ_q	Charge per unit area	C m^{-2}	258
σ_v	Standard deviation of voltage	V	255
τ	Time constant	s	247
$\boldsymbol{\tau}$	Torque	N m	261
τ_t	Tissue time constant	s	258
θ	Angle		260
ϕ	Angle in cylindrical coordinates		249
χ	Susceptibility		246
ω, ω_s, ω_0	Solute permeability	N^{-1} s^{-1}	249
ω	Angular frequency	s^{-1}	258
ω_t	Characteristic angular frequency of tissue	s^{-1}	
ξ	Energy in units of k_BT		242
Γ	Radial concentration factor		249

Problems

Section 9.1

Problem 1. The chloride ratio between plasma and interstitial fluid is 0.95. Plasma protein has a valence of about -18. In the interstitial fluid, $[\mathrm{Na}'] = [\mathrm{Cl}'] = 155$ mmol l^{-1}. Find the sodium, chloride and protein concentrations in the plasma and the potential difference across the capillary wall, assuming Donnan equilibrium.

Problem 2. Suppose that there are two compartments with equal volume $V = 1$ l, separated by a membrane that is permeable to K and Cl ions. Impermeant positive ions have a concentration 0 on the left and $[\mathrm{M}'] = [\mathrm{M}^{+\prime}] = 10$ mmol l^{-1} on the right. The initial concentration of potassium is $[\mathrm{K}_0] = 30$ mmol l^{-1} on the left. $T = 310$ K.

(a) Find the initial concentrations of potassium and chloride on both sides and the potential difference.

(b) A fixed amount of potassium chloride (10 mmol) is added on the left. After things have come to equilibrium, find the new concentrations and potential difference.

Problem 3. The extracellular space in cartilage contains large, immobile, negatively charged molecules called glycoaminoglycans (GAGs). An early sign of osteoarthritis is the loss of GAGs. The concentration of the GAGs is difficult to measure directly, but Shapiro et al. (2002) measured the sodium ion concentration in cartilage using magnetic resonance imaging (see Chap. 18). Assume the interstitial fluid of the body consists of 150 mM of sodium ions and 150 mM of chloride ions, and that both of these ions can move freely between the body fluid and the extracellular space of cartilage. The cartilage sodium ion concentration is measured to be 250 mM. If Donnan equilibrium holds, what is the concentration of the GAGs? For simplicity, assume the GAGs are monovalent.

Section 9.2

Problem 4. Derive the Poisson equation from Gauss's law in Cartesian coordinates in three dimensions.

Problem 5. Consider ions uniformly dispersed in a solution. Find the average linear separation of the ions for concentrations of 1, 10, 100, and 1000 mmol l^{-1}.

Problem 6. Verify Eq. 9.19.

Problem 7. Verify the parameters presented in Table 9.2. How accurate is the approximation $e^x \approx 1 + x$ in this case?

Problem 8. Consider a solution consisting of an equal concentration, C, of monovalent cations and anions.

(a) Show that $\rho_{\mathrm{ext}} = -2Ce \sinh\left(\frac{ev}{k_B T}\right)$.

(b) Let $\xi = ev/k_B T$ and $\mathbf{r}' = \mathbf{r}/\lambda_D$, where λ_D is given by Eq. 9.13. Show that the nonlinear Poisson–Boltzmann equation (Eq. 9.12) becomes $\nabla'^2 \xi = \sinh \xi$.

Problem 9. Analytical solutions to the nonlinear Poisson–Boltzmann equation are rare but not unknown. Consider the case when the potential varies in one dimension (x), the potential goes to zero at large x, and there exist equal concentrations of monovalent cations and anions. Chandler et al. (1965) showed that the solution to the 1-D Poisson–Boltzmann equation, $d^2\xi/dx'^2 = \sinh \xi$ (see Problem 8), is $\xi(x') = 4 \tanh^{-1}\left[\tanh(\xi_0/4)\, e^{-x'}\right]$, where ξ_0 is a constant and $0 < x' < \infty$.

(a) Verify that this expression satisfies $d^2\xi/dx'^2 = \sinh \xi$. (You may need a math handbook with a collection of hyperbolic function identities.)

(b) Linearize the Poisson–Boltzmann equation and show that its solution is $\xi(x') = \xi_0 e^{-x'}$.

(c) Show that both solutions are equal to ξ_0 at $x' = 0$ and equal to 0 at $x' = \infty$.

(d) Compare the solutions for the linear and nonlinear Poisson–Boltzmann equation at $x' = 0.5$ for the cases $\xi_0 = 0.1$, 1, and 10.

Section 9.3

Problem 10. The value of A used to obtain Eq. 9.29 was determined by saying that as $r \to 0$, the electric field must approach $ze/\kappa 4\pi\epsilon_0 r^2$. An elaboration of the model would be to say that the central ion has radius a and that the electric field at $r = a$ must be the same as the field at the surface of the ion, $ze/\kappa 4\pi\epsilon_0 a^2$. How does this change the expression for $v(r)$?

Problem 11. Using the method in Sect. 9.3, derive the Poisson–Boltzmann equation in cylindrical coordinates (r, ϕ, z; see Appendix L) assuming the electric field is radial and does not depend on ϕ or z. Solutions to the linearized version of this equation are zeroth order modified Bessel functions (see Abramowitz and Stegun 1972).

Section 9.4

Problem 12. A collection of molecular electric dipoles, each of moment $\mathbf{p}$, are in thermal equilibrium at temperature T. If the dipoles experience an electric field of strength E, then determine the average value of $\cos\theta$, where θ is the angle between the dipole and the electric field. Hint: assume the dipole orientations follow the Boltzmann distribution, which in this case is $\exp(pE\cos\theta/k_B T)$, and integrate over all solid angles $d\Omega = 2\pi \sin\theta d\theta$. Show that if $pE \ll k_B T$ the average of $\cos\theta$ is proportional to E, but if $pE \gg k_B T$

the average of $\cos\theta$ saturates at a value of one. Interpret this physically.

Problem 13. If Fig. 9.7 shows the water molecule in its average orientation, is the central ion an anion or a cation?

Section 9.5

Problem 14. Find an expression for the slope of the Nernst–Planck constant-field curve in Fig. 9.10 when v is equal to the Nernst potential, v_0. Hint: expand the exponentials in Eq. 9.45 around v_0.

Problem 15. Show that when $j = 0$, Eq. 9.42 gives $C(x) = C_0 e^{-zev(x)/k_BT}$, as we already know must be true in equilibrium. Hint: solve for dv/dx.

Problem 16. Calculate the conductivity of saline (9 g of NaCl in 1 l of water) at 25 °C.

Problem 17. The discussion surrounding Eqs. 9.34–9.41 was for a model of ions in a pore with constant electric field. It is also possible to write an integral version of the Nernst–Planck equation. Consider a single channel in which the current is the same for all values of x, the distance along the channel. If the diffusion constant and cross-sectional area of the channel are allowed to vary, and with the usual substitution $u(x) = zev(x)/k_BT$, Eq. 9.41 becomes

$$i = j(x)S(x) = -zeD(x)S(x)\left(\frac{dC}{dx} + C(x)\frac{du}{dx}\right).$$

(a) Show that if each term is multiplied by e^u, this can be written as

$$\frac{ie^{u(x)}}{D(x)S(x)} = -ze\left(e^{u(x)}\frac{dC}{dx} + C(x)e^{u(x)}\frac{du}{dx}\right).$$

(b) Show that if the integration is carried from x_1 to x_2, then the current in the channel is

$$i = -\frac{ze\left[C(x_2)e^{u(x_2)} - C(x_1)e^{u(x_1)}\right]}{I},$$

where the integral

$$I = \int_{x_1}^{x_2} \frac{e^{u(x)}\,dx}{S(x)D(x)},$$

contains all the information about the channel.

Problem 18. Cardiac cells have a potassium channel, called "I_{K1}", which shows inward rectification (larger current for potentials more negative than the potassium Nernst potential, v_K, than for potentials more positive than v_K). This channel sometimes is said to show *anomalous rectification*. Why is it anomalous? (The mechanism of anomalous rectification is described by Nichols et al. 1996.)

Section 9.6

Problem 19. Consider a channel that is 100 times more permeable to potassium than to sodium (ignore all other ions).

(a) Write an equation for the reversal potential as a function of the intracellular and extracellular sodium and potassium ion concentrations.

(b) Assume $[K_i] = 150$, $[Na_i] = 50$, and $[Na_e] = 150$ mM. Plot v_r versus $[K_e]$ using semilog paper. On the same plot, draw the potassium Nernst potential as a function of $[K_e]$.

Problem 20. Calculation of the permeability ratios from measurement of the reversal potential is difficult because the concentrations inside the axon are not known. One can overcome this by measuring how the reversal potential (Eq. 9.55) changes as outside concentrations are varied. Obtain an equation for the shift of reversal potential if two measurements are made: one in which $[Na_1] = 0$, and the other with $[K_1] = 0$ (assume $\omega_{Cl} = 0$).

Section 9.7

Problem 21. A patch–clamp experiment shows that the conductance of a single Ca^{2+} channel is $G = 25$ pS. The membrane thickness is $b = 6$ nm. Use $v = 50$ mV.

(a) Assuming that the resistivity of the fluid in the channel is $\rho = 0.5\ \Omega$ m, find an expression and numerical value for the channel radius a.

(b) If the conductance per unit area is 1200 S m^{-2}, find the number of pores per unit area.

(c) The current is $i = Gv$, where v is the applied voltage. Find an expression for n, the number of calcium ions per second passing through the channel, in terms of whichever of parameters G, v, b, and a are necessary.

(d) How many calcium ions are in the channel at one time, if the calcium concentration is C mmol l^{-1}?

Problem 22. A potassium channel might have a radius of 0.2 nm and a length of 6 nm. If it contained potassium at a concentration of 150 mmol l^{-1}, how many potassium ions on average would be in the channel?

Problem 23. How long does it take for a sodium ion to drift in the electric field (assumed constant) through a membrane of thickness L and applied potential v? How long does it take to move by pure diffusion? Find numerical values when the membrane is 6 nm thick and potential difference is 70 mV.

Problem 24. Suppose that a sodium pore when open passes 10 pA and $j_{Na} = 0.2$ A m^{-2}. Calculate the number of open pores per unit area and the average linear spacing between them.

Problem 25. Calculate the current density of sodium ions in a region of length 6 nm due to (a) pure diffusion when

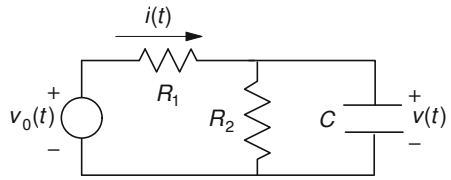
there is no potential difference and the concentrations are 145 and 15 mmol l^{-1}, (b) pure drift when the concentration is 145 mmol l^{-1} and the potential difference is 70 mV, and (c) both diffusion and drift if the electric field is constant.

Problem 26. Patch-clamp recording is done with a micropipette of radius 1 μm.

(a) If the pipette encircles a single channel with conductance 20 pS, what is the channel current when the channel is open and the voltage across the membrane is 20 mV away from the Nernst potential for the ion in question? Make a simple estimate using Ohm's law.

(b) Assuming a capacitance of 0.01 F m^{-2}, what current charges the capacitance of the membrane patch under the micropipette if a 20-mV change occurs linearly in 5 μs?

Problem 27. The following circuit illustrates the effects that must be considered when an electrode is used to measure the properties of a patch of membrane. R_1 is the resistance of the electrode. R_2 and C are properties of the membrane. The applied voltage $v_0(t)$ is a step at $t = 0$. The electrode current is $i(t)$. The voltage across the membrane patch is $v(t)$.



(a) Show that

$$v_0(t) = R_1 C \frac{dv}{dt} + \frac{R_1 + R_2}{R_2} v(t).$$

(b) Show that the time constant is $\tau = R_1 R_2 C/(R_1 + R_2)$ and that $\tau \to R_1 C$ if $R_1 \ll R_2$, $\tau \to R_2 C$ if $R_1 \gg R_2$.

(c) If $v_0(t)$ is a step of height v_0 at $t = 0$, show that

$$v(t) = v_0 \frac{R_2}{R_1 + R_2} \left(1 - e^{-t/\tau}\right)$$

and

$$i(t) = \frac{v_0}{R_1 + R_2} \left(1 + \frac{R_2}{R_1} e^{-t/\tau}\right).$$

(d) Plot $v(t)$ and $i(t)$.

(e) The case $R_1 \ll R_2$ is called *voltage-clamped*. Find expressions for $v(t)$ and $i(t)$ in that case and plot them. Where does the transient current flow? For fixed R_2, what is the time constant?

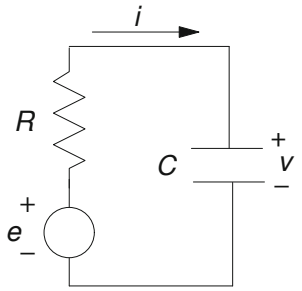
(f) In the *current-clamped* case, $R_1 \gg R_2$ and $i_0 = v_0/R_1$. Find expressions for $v(t)$ and $i(t)$ and plot them. For fixed R_2, what is the time constant?

(g) Make numerical plots of $v(t)$ and $i(t)$ when $v_0 = 150$ mV, $R_1 = 10^6\ \Omega$, $C = 5$ pF, and $R_2 = 10^{11}\ \Omega$.

Problem 28. A patch-clamp experiment is done with a micropipette having a resistance of $10^6\ \Omega$. When 150 mV is applied across the membrane, the current is 0 when the pores are closed and 1 pA when one channel is open. The membrane capacitance is 4×10^{-3} F m^{-2}. The microelectrode tip has an inner radius of 20 μm. What is the time constant for voltage changes? Does it depend on whether the channel has opened or closed?

Section 9.8

Problem 29. Weaver and Astumian (1990) derived Eq. 9.62a for the thermal noise of the transmembrane potential using a different method than in Sect. 9.8. A resistor has a voltage noise spectral density, $\sigma_e^2(f)$ (in units of V^2 Hz^{-1}), such that $\sigma_e^2(f) = 4k_B T R$, where f is the frequency (Sect. 11.16). It corresponds to voltage e in the figure. Weaver and Astumian represented the membrane as a parallel combination of membrane capacitance C and membrane resistance R (which is always in series with its noise source, e). The voltage across the capacitor, v, is the transmembrane potential.



(a) For a particular frequency f, derive a relationship between the spectral density of the voltage fluctuations of the transmembrane potential, $\sigma_v^2(f)$, and $\sigma_e^2(f)$. (Hint: derive an equation governing the voltage in an RC circuit, and then solve it using the methods described in Appendix F.)

(b) Integrate $\sigma_v^2(f)$ over all frequencies to get the voltage fluctuations σ_v^2.

(c) Estimate $\sqrt{\sigma_v^2}$ for a spherical cell of radius 10 μm, having a membrane capacitance per unit area of 0.01 F m^{-2}.

Section 9.9

Problem 30. In some nerve membranes a region of *negative resistance* is found, in which the current decreases as the voltage is increased.

(a) Where have we seen this behavior before?

(b) To see why it happens, consider two cases. The current through the membrane is given by $j = g(v)(v - v_0)$, where $g(v)$ is a property of the membrane, and the Nernst potential v_0 depends on the ion concentration on either side of the membrane. For this problem let

$v_0 = +50$ mV. Calculate j as a function of v for two cases: (a) $g(v) = 1$; (b) the conductance increases rapidly with voltage: $g(v) = (5.6 \times 10^{-7})e^{0.288v}$ (v in mV).

(c) Negative resistance increases the sensitivity of the ampullae of Lorenzini, as measured by Lu and Fishman (1994). To see why, calculate the output voltage in a two-resistance voltage divider network (as in Fig. 6.23) and discuss what happens if R_2 is negative.

Section 9.10

Problem 31. Estimate the transmembrane potential that corresponds to the threshold for electroporation. Compare it to the normal cell resting potential.

Problem 32. Here is one way that signal-to-noise ratio can be improved. Suppose that there are N receptors, connected in the nervous system in such a way that an output response requires a logical AND between all N receptors. The output is sampled every t seconds to determine whether or not there is a response. If the signal exists, all N receptors respond. If the signal does not exist, each receptor responds to thermal noise with a probability p (which might be $p = e^{-U/k_BT}$, where U is an activation energy). Assume that p is the same for each receptor, and that whether a receptor has responded to thermal noise is independent of the response of all other receptors and also independent of its response at any other time.

(a)What is the signal-to-noise (S/N) ratio as a function of N? Suppose that $N = 8$. Plot S/N as a function of p.

(b) Find U/k_BT vs. N for $S/N = 4$.

Problem 33. Here is another way to look at the signal-to-noise ratio.

(a) Show that the energy of a charged parallel-plate capacitor can be written as $\kappa\epsilon_0 E^2 V/2$, where $V = Sd$ is the volume between the plates. This is a special case of a general relationship that the energy per unit volume associated with an electric field is $\kappa\epsilon_0 E^2/2$.

(b) Use the information about the magnitude of the electric field in the cell membrane from Fig. 9.19 to calculate the total electrostatic energy in the membrane.

(c) Compare the ratio of the total electrostatic energy to k_BT when the air field is 300 V m^{-1}. This overestimates the ratio, because the energy is spread over the entire membrane and is not available to interact in one place.

Problem 34. Obtain Eq. 9.79 from the expression $U = -mB\cos\theta$ that was derived in Problem 8.35, by making a suitable expansion for small angles.

Problem 35. Electric fields in the body caused by exposure to power lines are produced by two mechanisms: direct coupling to the power line electric field, and Faraday induction from the power line magnetic field. Consider a high-voltage power line that produces an electric field of 10 kV m^{-1} and a magnetic field of 50 μT (Barnes 1995). Estimate the electric field induced in the human body by these two mechanisms. Which is larger? Compare the strength of these powerline-induced electric fields to the strength of naturally-occurring electric field produced in the body by the heart (estimate the strength of this endogenous field using the data in Fig. 7.23).

Problem 36. Derive the equations for the electric field shown in Fig. 9.19. Use the following method. Let the potentials be $v_{\text{outside}} = A\cos\theta/r^2 - E_1 r\cos\theta$ and $v_{\text{inside}} = Br\cos\theta$, where A and B are unknown constants. At the cell surface, the following boundary condition applies when the cell membrane is thin and obeys Ohm's law:

$$\sigma_{\text{outside}}\left(\frac{\partial v_{\text{outside}}}{\partial r}\right)\Bigg|_{r=a} = \sigma_{\text{inside}}\left(\frac{\partial v_{\text{inside}}}{\partial r}\right)\Bigg|_{r=a} = (v_{\text{outside}} - v_{\text{inside}})\frac{\sigma_{\text{membrane}}}{b}\Bigg|_{r=a}$$

(a) Verify that the expressions for v_{outside} and v_{inside} obey Laplace's equation and behave properly at $r = 0$ and $r = \infty$.

(b) Use the boundary condition to determine A and B.

(c) Use your expressions for the potential to determine the electric fields given in Fig. 9.19.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions with formulas, graphs and mathematical tables. U. S. Government Printing Office, Washington, DC

Adair RK (1991) Constraints on biological effects of weak extremely-low-frequency electromagnetic fields. Phys Rev A 43:1039–1048

Adair RK (1992) Reply to "Comment on 'Constraints on biological effects of weak extremely-low-frequency electromagnetic fields.'" Phys Rev A 46:2185–2187

Adair RK (1993) Effects of ELF magnetic fields on biological magnetite. Bioelectromagnetics 14:1–4

Adair RK (1994) Constraints of thermal noise on the effects of weak 60-Hz magnetic fields acting on biological magnetite. Proc Natl Acad Sci U S A 91:2925–2929

Adair RK (2000) Static and low-frequency magnetic field effects: health risks and therapies. Rep Prog Phys 63:415–454

Adair RK, Astumian RD, Weaver JC (1998) Detection of weak electric fields by sharks, rays and skates. Chaos 8:576–587

Ashcroft F (2012) The spark of life: electricity in the human body. Norton, New York

Astumian RD, Weaver JC, Adair RK (1995) Rectification and signal averaging of weak electric fields by biological cells. Proc Natl Acad Sci U S A 92:3740–3743

Barnes FS (1995) Typical electric and magnetic field exposures at power-line frequencies and their coupling to biological systems. In: Blank M (ed) Electromagnetic fields: biological interactions

and mechanisms. American Chemical Society, Washington, DC, pp 37–55
Bastian J (1994) Electrosensory organisms. Phys Today 47(2):30–37
Bockris JO'M, Reddy AKN (1970) Modern electrochemistry, vol 1. Plenum, New York
Bren SPA (1995) 60 Hz EMF health effects—a scientific uncertainty. IEEE Eng Med Biol 14:370–374
Carstensen EL (1995) Magnetic fields and cancer. IEEE Eng Med Biol 14:362–369
Chandler WK, Hodgkin AL, Meves H (1965) The effect of changing the internal solution on sodium inactivation and related phenomena in giant axons. J Physiol 180:821–836
Crouzy SC, Sigworth FJ (1993) Fluctuations in ion channel gating currents: analysis of nonstationary shot noise. Biophys J 64:68–76
DeFelice LJ (1981) Introduction to membrane noise. Plenum, New York
Denk W, Webb WW (1989) Thermal-noise-limited transduction observed in mechanosensory receptors of the inner ear. Phys Rev Lett 63(2):207–210
Doyle DA, Cabral JM, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. Science 280: 69–77
Foster KR (1996) Electromagnetic field effects and mechanisms: in search of an anchor. IEEE Eng Med Biol 15(4):50–56
Foster KR, Moulder JE (2013) Wi-Fi and health: review of current status and research. Health Phys 105(6):561–575
Foster KR, Schwan HP (1996) Dielectric properties of tissues. In: Polk C, Postow E (eds) Handbook of biological effects of electromagnetic fields. CRC, Boca Raton, pp 25–102
Grosberg AY, Nguyen TT, Shklovskii BI (2002) Colloquium: the physics of charge inversion in chemical and biological systems. Rev Mod Phys 74: 329–345
Hafemeister D (1996) Resource letter BELFEF-1: biological effects of low-frequency electromagnetic fields. Am J Phys 64:974–981
Hamill OP, Marty A, Neher E, Sakmann B, Sigworth FJ (1981) Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. Pflügers Arch 391:85–100
Hille B (2001) Ion channels of excitable membranes, 3rd ed. Sinauer Associates, Sunderland
Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268:1144–1149
Jiang Y, Le A, Chen J, Ruta V, Cadene M, Chait BT, MacKinnon R (2003) X-ray structure of a voltage-dependent K^+ channel. Nature 423:33–41
Kalmijn AJ (1988) Detection of weak electric fields. In: Aetma J et al. (eds) Sensory biology of auatic animals. Springer, New York, pp 151–186
Keynes RD (1994) The kinetics of voltage-gated ion channels. Q Rev Biophys 27(4):339–434
Khurana VG, Moulder JE, Orton CG (2008) There is currently enough evidence and technology available to warrant taking immediate steps to reduce exposure of consumers to cell-phone-related electromagnetic radiation. Med Phys 35:5203–5206
Kirschvink JL (1992) Comment on "Constraints on biological effects of weak extremely-low-frequency electromagnetic fields." Phys Rev A 46:2178–2184
Kirschvink JL, Kobayashi-Kirschvink A, Woodford BJ (1992a) Magnetite biomineralization in the human brain. Proc Natl Acad Sci U S A 89:7683–7687
Kobayashi AK, Kirschvink JL, Nesson MH (1995) Ferromagnetism and EMFs. Nature 374:123
Lamb H (1932) Hydrodynamics. Cambridge University Press, Cambridge
Leuchtag HR, Swihart JC (1977) Steady-state electrodiffusion. Scaling, exact solutions for ions of one charge, and the phase plane. Biophys J 17:27–46
Lewis CA (1979) Ion-concentration dependence of the reversal potential and the single channel conductance of ion channels at the frog neuromuscular junction. J Physiol 286:417–445
Lu J, Fishman HM (1994) Interaction of apical and basal membrane ion channels underlies electroreception in ampullary epithelia of skates. Biophys J 67:1525–1533
Mauro A (1962) Space charge regions in fixed charge membranes and the associated property of capacitance. Biophys J 2:179–198
Moulder JE, Foster KR (1995) Biological effects of power-frequency fields as they relate to carcinogenesis. Proc Soc Expt Biol Med 209:309–324
Moulder JE, Foster KR (1999) Is there a link between exposure to power-frequency electric fields and cancer? IEEE Eng Med Biol 18(2):109–116
Moulder JE, Foster KR, Erdreich LS, McNamee JP (2005) Mobile phones, mobile phone base stations and cancer: a review. Int J Radiat Biol 81(3):189–203
Moy G, Corry B, Kuyucak S, Chung S-H (2000) Tests of continuum theories as models of ion channels. I. Poisson-Boltzmann theory versus Brownian dynamics. Biophys J 78:2349–2362
National Research Council (1997) Committee on the possible effects of electromagnetic fields on biologic systems. Possible health effects of exposure to residential electric and magnetic fields. National Academy Press, Washington, DC
Neher E, Sakmann B (1976) Single-channel currents recorded from membrane of denervated frog muscle fibers. Nature 260:799–802
Nichols CG, Makhina EN, Pearson WL, Sha Q, Lopatin AN (1996) Inward rectification and implications for cardiac excitability. Circ Res 78:1–7
Pallotta BS, Magleby KL, Barrett JN (1981) Single channel recordings of Ca^{2+}-activated K^+ currents in rat muscle cell culture. Nature 293:471–474
Payandeh J, Scheuer T, Zheng N, Catterall WA (2011) The crystal structure of a voltage-gated sodium channel. Nature 475:353–358
Pickard WF (1988) A model for the acute electrosensitivity of cartilaginous fishes. IEEE Trans Biomed Eng 35(4):243–249
Polk C (1994) Effects of extremely-low-frequency magnetic fields on biological magnetite. Bioelectromagnetics 15:261–270
Polk C (1995) Bioelectromagnetic dosimetry. In: Blank M (ed) Electromagnetic fields: biological interactions and mechanisms. American Chemical Society, Washington, DC, pp 57–78
Polk C (1996) Introduction. In: Polk C, Postow E (eds) Handbook of biological effects of electromagnetic fields. CRC, Boca Raton, pp 1–23
Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B (2009) The role of DNA shape in protein-DNA recognition. Nature 461:1248–1253
Shapiro EM, Borthakur A, Gougoutas A, Reddy R (2002) ^{23}Na MRI accurately measures fixed charge density in articular cartilage. Magn Reson Med 47:284–291
Sheppard AR, Swicord ML, Balzano Q (2008) Quantitative evaluations of mechanisims of radiofrequency interactions with biological molecules and processes. Health Phys 95(4):365–396
Sigworth FJ (1993) Voltage gating of ion channels. Q Rev Biophys 27(1):1–40
Tenforde TS (1995) Spectrum and intensity of environmental electromagnetic fields from natural and man-made sources. In: Blank M (ed) Electromagnetic fields: biological interactions and mechanisms. American Chemical Society, Washington, DC, pp 13–35
Tucker RD, Schmitt OH (1978) Tests for human perception of 60 Hz moderate strength magnetic fields. IEEE Trans Biomed Eng 25:509–518

Uehara M, Sakane KK, Maciel HS, Urruchi WI (2000) Physics and biology: bio-plasma physics. Am J Phys 68:450–455

Vecchia P, Matthes R, Ziegelberger G, Lin J, Saunders R, Swerdlow A (eds) (2009) Exposure to high frequency electromagnetic fields, biological effects and health consequences (100 kHz–300 GHz). ICNRP 16/2009. Oberschleißheim, Germany, International Commssion on Non-Ionizing Radiation Protection

Weaver JC (2000) Electroporation of cells and tissues. IEEE Trans Plasma Sci 28:24–33

Weaver JC, Astumian RD (1990) The response of living cells to very weak electric fields: the thermal noise limit. Science 247:459–462

Weaver JC, Astumian RD (1995) Issues relating to causality of bioelectromagnetic fields. In: Blank M (ed) Electromagnetic fields: biological interactions and mechanisms. American Chemical Society, Washington, DC, pp 79–96

Yi M, Nymeyer H, Zhou H-X (2008) Test of the Gouy-Chapman theory for a charged lipid membrane against explicit-solvent molecular dynamics simulations. Phys Rev Lett 101:038103

Feedback and Control

10

We now turn to the way in which the body regulates such things as temperature, oxygen concentration in the blood, cardiac output, number of red or white blood cells, and blood concentrations of substances like calcium, sodium, potassium and glucose. Each of these is regulated by a *feedback loop*. A feedback loop exists if variable x determines the value of variable y, and variable y in turn determines the value of variable x.

Suppose that x is the deviation of a bullet from its desired path. A bullet has no feedback; after it has left the gun, its deviation from the desired path is determined by the initial aim of the gun, fall due to gravity, drift caused by the wind, and air turbulence. An accuracy of one part in 10^4 (about a tenth of an inch in 50 ft) is quite good. A car, on the other hand, is steered by the driver. If deviation from the center of the lane x becomes appreciable, the driver changes y, the position of the steering wheel. The value of y determines x through the steering mechanism and the tires. It is possible to have a car deviate less than 1 ft from the desired position within a lane after driving 3000 miles, an accuracy of one part in 10^7. This is an example of *negative feedback*. If x gets too large, the factors in the feedback loop tend to reduce it.

Negative-feedback systems can generate oscillations of their variables. We see oscillations in physiological systems on many different time scales, from the rhythmic activity of the heart, to changes in the rate of breathing, to daily variations in body temperature, blood pressure, and hormone levels, to monthly variations such as the menstrual cycle, to annual variations such as hibernation, coloring, fur growth, and reproduction.

It is also possible to have *positive* feedback. Two bickering children can goad each other to new heights of anger. Positive feedback initiates the action potential described in Chap. 6: depolarization of the axon leads to increased sodium permeability, which further speeds depolarization. Blood pressure is regulated in part by sensors in the kidney. A patient with high blood pressure may suffer damage to the blood vessels, including those feeding the kidneys, which reduces the blood pressure at the sensors. The sensors then ask for still higher blood pressure, which accelerates the damage, which leads to still higher blood pressure, and so on.

The simplest feedback loop consists of two processes: one in which y depends on x and another in which x depends on y. The loop can have many more variables. Steering the car, in addition to the variables of lane position and steering-wheel position, involves vision, neuromuscular processes, all of the variables in the automobile's steering mechanism, and the Newtonian mechanics of the car's motion—with external variables such as the behavior of other drivers continually bombarding the system.

Sections 10.1–10.3 deal with the relationships between the feedback variables when the system is in *equilibrium* or in the *steady state,* and none of the variables are changing with time. The techniques for determining the operating point—the steady-state values of the variables—are graphical and can be applied to any system if the relationship among the variables is known.

When the system is not at equilibrium, it returns to the equilibrium point if the system is stable. Although the equations describing this return to equilibrium are usually not linear, Sects. 10.4–10.7 discuss how linear systems behave when they are not at the operating point. A linear system may " decay" exponentially to the steady-state values or it may exhibit oscillations.

Most systems are not linear. Section 10.8 discusses systems described by nonlinear equations in one or two dimensions, introducing some of the vocabulary and graphical techniques of nonlinear systems analysis. It closes with an example of resetting the phase of a biological oscillator. Section 10.9 introduces the ideas of *period doubling* and *chaotic behavior* through difference equations and the *logistic map*. It then describes a *linear map* that appears to be chaotic but is not. Section 10.10 shows how a linear differential equation that depends on a fixed delay in the variable can exhibit

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_10, © Springer International Publishing Switzerland 2015

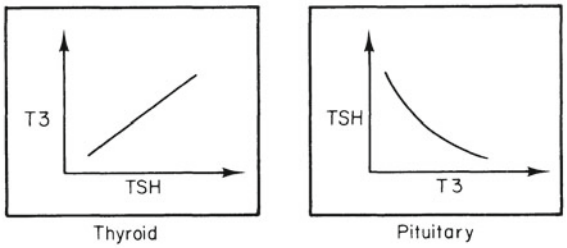


Fig. 10.1 Schematic curves of the relationship between thyroid hormone (*T3*) and thyroid stimulating hormone (*TSH*) in the thyroid gland and in the pituitary

either damped or continuous oscillations. Section 10.11 summarizes the earlier sections, and Sect. 10.12 gives several biological examples.

A great deal of work was done on modeling physiological feedback systems between 1950 and 1975. Books from that era include Riggs (1970) and Stark (1968). More information about mathematical modeling in biology and medicine can be found in Murray (2007, 2008) and Keener and Sneyd (2008a, 2008b).

10.1 Steady-State Relationships Among Variables

Any feedback loop can be broken down, conceptually at least, into separate processes that relate a dependent variable to an independent variable and possibly to some other parameters. Figure 10.1 shows an example. In the first process the thyroid gland, in response to thyroid-stimulating hormone (TSH) from the pituitary, produces the thyroid hormones thyroxine (T4) and tri-iodothyronine (T3). An increase of TSH increases production of T3 and T4. These processes depend on other parameters, such as the amount of iodine available in the body to incorporate into the T3 and T4. In the second process, the pituitary increases the production of TSH if the concentration of T3 in the blood falls. It may also respond to T4 and other variables as well. (This is an oversimplification. The pituitary actually responds to hormones secreted by the hypothalamus. The hypothalamus is responding to the levels of T3 and T4.)

For a quantitative example, consider a simple model relating the amount of carbon dioxide in the alveoli (air sacs of the lung) and the rate of breathing (ventilation rate). If the body is producing CO_2 at a constant rate, a given ventilation rate corresponds to a definite value of P_{CO_2}, the partial pressure of carbon dioxide in the alveoli. In the steady state the amount of CO_2 exhaled (the volume of gas leaving the lungs per minute times P_{CO_2}) is just equal to the amount produced

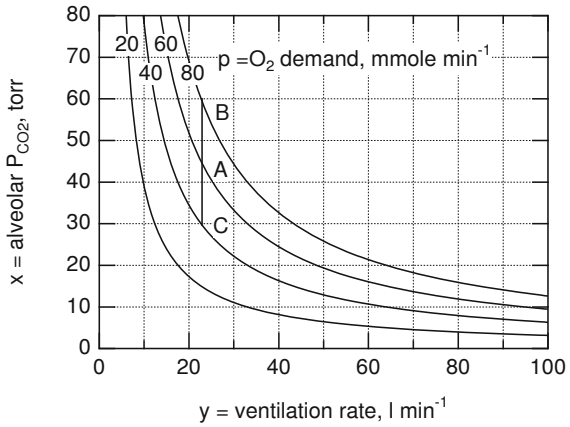


Fig. 10.2 The pressure of CO_2 in the alveoli of the lungs decreases as the ventilation rate is increased. The different curves correspond to different total metabolic rates

in the body. Figure 10.2 shows this relationship when the pH and P_{O_2} of the blood are fixed. As ventilation rate rises, P_{CO_2} falls. We are ignoring several other feedback loops (Riggs 1970, pp. 401–418). If the metabolic rate rises, P_{CO_2} also rises. Experiments show that ventilation rate y and alveolar P_{CO_2} (which we will call x) are related by (Riggs 1970)

$$x = \frac{15.47p}{y - 2.07}. \tag{10.1}$$

In these equations y (the independent variable) is measured in l min^{-1}, x (the dependent variable) is in torr, and parameter p is the body's oxygen consumption in mmol min^{-1}. A typical resting person requires $p = 15$ mmol min^{-1}.

Equation 10.1 can be derived using a simple model for respiration. Let the metabolic rate of the body be described by o, the rate of oxygen consumption in mol s^{-1}. The respiratory quotient F relates o to the rate of CO_2 production, so

$$(\text{rate of } CO_2 \text{ production}) = Fo. \tag{10.2}$$

A typical value of F is 0.8.

Carbon dioxide is removed from the body by breathing. If the rate at which air flows through the alveoli is[1] $(dV/dt)_{\text{alveoli}}$ in m^3 s^{-1}, then the rate of removal is obtained from the ideal-gas law:

$$(\text{rate of } CO_2 \text{ removal}) = \frac{x(dV/dt)_{\text{alveoli}}}{RT}.$$

The rate $(dV/dt)_{\text{alveoli}}$ is less than the ventilation rate y because air in the trachea and bronchi does not exchange

[1] Strictly speaking, $(dV/dt)_{\text{alveoli}}$ is not the derivative of a function V. (It always has a positive value, and the lungs are not expanding without limit!) We use the notation to remind ourselves that it is the rate of air exchange in the alveoli.

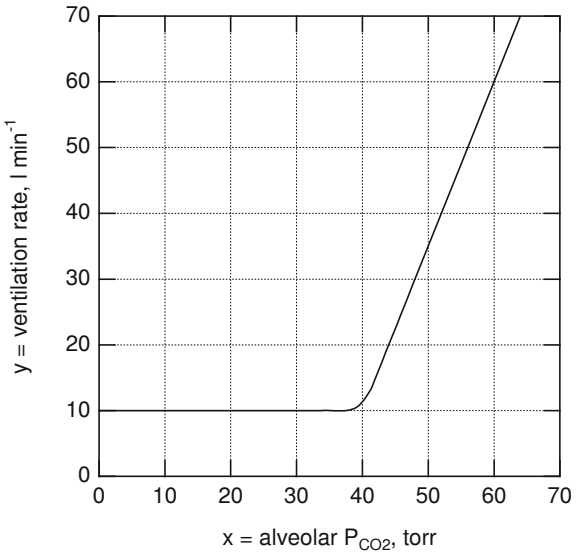


Fig. 10.3 When P_{CO_2} in the blood rises above 40 torr, the breathing rate increases

oxygen or carbon dioxide with the blood: $(dV/dt)_{\text{alveoli}} = y - b$. Therefore

$$(\text{rate of } CO_2 \text{ removal}) = \frac{x(y-b)}{RT}. \qquad (10.3)$$

In equilibrium the rate of production is equal to the rate of removal, so

$$Fo = \frac{x(y-b)}{RT}$$

or

$$x = \frac{RTFo}{y-b}. \qquad (10.4)$$

With the proper conversion of units from p to o, this is Eq. 10.1.

If the metabolic rate were to change without a change in breathing rate, $x = P_{CO_2}$ would change drastically. Suppose that someone exercises moderately so that $p = 60$ mmol min^{-1}, $y = 23$ l min^{-1}, and $x = 44$ torr, point A in Fig. 10.2. If ventilation rate y remained constant while p rose to 80 mmol min^{-1}, $x = P_{CO_2}$ would soar to about 60; if p fell to 40, x would drop to 30. Feedback ensures that this does not happen. One of the feedback mechanisms consists of an area of the brain stem that senses the value of $x = P_{CO_2}$ and causes y to change. Figure 10.3 shows a typical curve for a 70-kg male (Patton 1989, p. 1034). (The concentration of CO_2 in blood is nearly the same as in the alveoli.)

10.2 Determining the Operating Point

We now have two processes relating the steady-state values of x and y. For alveolar gas exchange, we know x as a function of y: $x = g(y, p)$. For the regulatory mechanism,

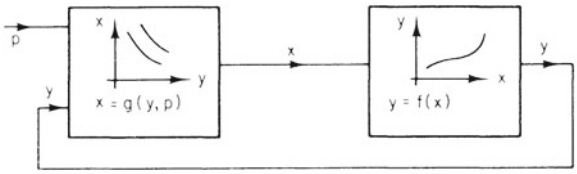


Fig. 10.4 A general feedback loop. Either box may involve some parameters

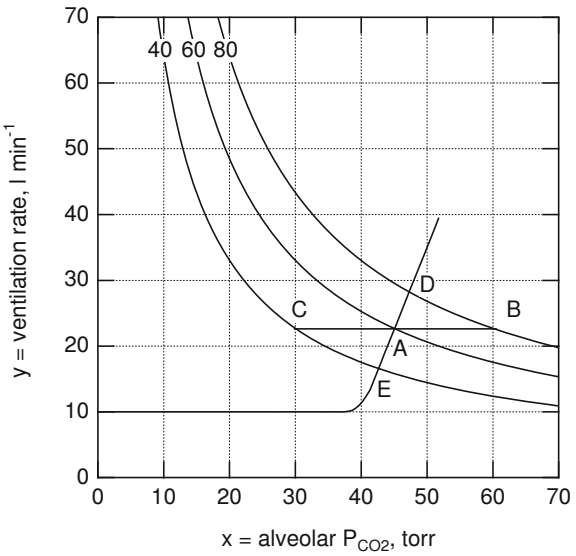


Fig. 10.5 Regulation of the breathing rate. A change of metabolic rate (parameter p) causes a change in ventilation rate y, so that $x = P_{CO_2}$ does not change as much

we know $y = f(x)$. Together, these constitute a feedback loop, Fig. 10.4. To find the *operating point*, these two equations must be solved simultaneously. The easiest way to do this is to plot them on the same graph as in Fig. 10.5. When $p = 60$ mmol min^{-1}, the operating point is at A. In a plot like this the horizontal axis represents the independent variable for one process and the dependent variable for the other.

If the feedback loop includes several variables, for example

$$x = f(w), \quad y = g(x), \quad z = h(y), \quad w = i(z),$$

we can combine three of these equations to get $x = F(y)$ and plot it with $y = g(x)$.

10.3 Regulation of a Variable and Open-Loop Gain

We can also see from Fig. 10.5 how feedback causes y to change in response to a change in parameter p to reduce the change in x. If y does not change, a change of p from 60 to

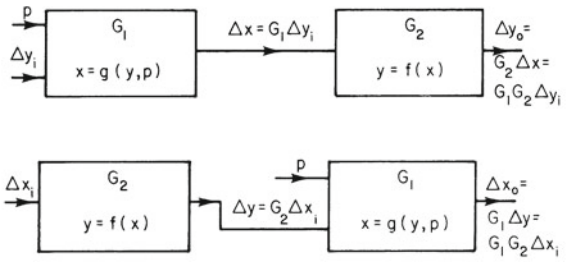


Fig. 10.6 The open loop gain is calculated by opening the loop at any point. **a** Loop opened in y. **b** Loop opened in x

80 causes the operating point to go from A to B. In fact, y increases so that the new operating point is at D. The feedback loop is said to regulate the value of x.

The *gain* of each box in Fig. 10.4 is the ratio of the change in the output variable to the change in the input variable. For the first box

$$G_1 = \left(\frac{\Delta x}{\Delta y}\right)_{\text{box } g,\ p \text{ fixed}} = \left(\frac{\partial x}{\partial y}\right)_{\text{box } g,\ p \text{ fixed}} = \left(\frac{\partial g}{\partial y}\right)_p. \tag{10.5}$$

For the second box,

$$G_2 = \left(\frac{\Delta y}{\Delta x}\right)_{\text{box } f} = \left(\frac{\partial y}{\partial x}\right)_{\text{box } f} = \frac{\partial f}{\partial x}. \tag{10.6}$$

The product G_1G_2 is called the *open-loop gain* (OLG). Its name comes from the fact that if the feedback loop is opened at any point and a small change is made in the input variable at the opening, the change in the output variable is the open-loop gain times the change in the input variable:

$$\text{OLG} = G_1G_2 = \left(\frac{\partial x}{\partial y}\right)_{\text{box } g}\left(\frac{\partial y}{\partial x}\right)_{\text{box } f} = \frac{\partial g}{\partial y}\frac{\partial f}{\partial x}. \tag{10.7}$$

The open-loop gain can be calculated by taking the derivatives in either order, which corresponds to breaking the loop after either box (Fig. 10.6).

If the relationships between the derivatives have been plotted as in Fig. 10.5, it may be easiest to evaluate the derivatives graphically. In that case, it is easiest to work with $\partial y/\partial x$ for box g. But $\partial y/\partial x = 1/(\partial x/\partial y)$. Therefore,

$$\text{OLG} = G_1G_2 = \frac{(\partial y/\partial x)_{\text{box } f}}{(\partial y/\partial x)_{\text{box } g}}. \tag{10.8}$$

It is important to calculate the gain in the direction that causality operates. Going around the loop the wrong way gives the reciprocal of the open-loop gain.

We can now calculate how much feedback reduces the change in x, compared to the case in which there is no feedback and the value of y going into box g is held fixed. For box g, where $x = g(y, p)$, we can write for small changes in p and y

$$\begin{aligned}\Delta x &= \left(\frac{\partial x}{\partial p}\right)_{y,\text{ box } g}\Delta p + \left(\frac{\partial x}{\partial y}\right)_{p,\text{ box } g}\Delta y \\ &= \left(\frac{\partial x}{\partial p}\right)_{y,\text{ box } g}\Delta p + G_1\Delta y.\end{aligned} \tag{10.9}$$

When there is no feedback, Δy is zero and

$$\Delta x = \left(\frac{\partial x}{\partial p}\right)_{y,\text{ box } g}\Delta p.$$

When there is feedback, there is a value of Δy to be included. If the change in x with feedback is $\Delta x'$, the change in y can be calculated from the second box:

$$\Delta y = \left(\frac{\partial f}{\partial x}\right)\Delta x' = G_2\Delta x'. \tag{10.10}$$

This can be combined with Eq. 10.9:

$$\Delta x' = \left(\frac{\partial x}{\partial p}\right)_y \Delta p + G_1(G_2\Delta x') = \Delta x + G_1G_2\Delta x'$$

and solved for $\Delta x'$:

$$\Delta x' = \frac{\Delta x}{1 - G_1G_2} = \frac{\Delta x}{1 - \text{OLG}}. \tag{10.11}$$

The effect of feedback is to cause a change in y that reduces the change in x by the factor $1 - \text{OLG}$. When the feedback is negative, the open-loop gain is negative, $1 - \text{OLG}$ is greater than one, and there is a reduction in Δx. If the feedback is positive and the open-loop gain is less than one, $\Delta x'$ is larger than Δx.

For the respiration example, the equations for each box are

$$\begin{aligned} x = g(y, p) &= \frac{15.47p}{y - 2.07}, \\ y = f(x) &= \begin{cases} 10, & x \le 40, \\ 10 + 2.5(x - 40), & x > 40. \end{cases}\end{aligned} \tag{10.12}$$

The derivatives are

$$\begin{aligned}\left(\frac{\partial g}{\partial p}\right)_y &= \frac{15.47}{y - 2.07}, \\ G_1 = \left(\frac{\partial g}{\partial y}\right)_p &= -\frac{15.47p}{(y - 2.07)^2}, \\ G_2 = \left(\frac{\partial f}{\partial x}\right) &= 2.5.\end{aligned}$$

At operating point A in Fig. 10.5, the values are

$$x = 45.07, \quad p = 60, \quad y = 22.67,$$
$$\left(\frac{\partial g}{\partial p}\right)_y = 0.757, \tag{10.13}$$
$$G_1 = -2.19, \quad G_2 = 2.5, \quad \text{OLG} = -5.48.$$

If p changes from 60 to 62, then without feedback $\Delta x = (0.757)(2) = 1.5$. With feedback, $\Delta x' = 1.5/(1 + 5.48) = 0.23$.

10.4 Approach to Equilibrium without Feedback

The technique described in the preceding section allows us to determine the equilibrium state or operating point of a system if we can measure the functions f and g. It does not tell us how the system behaves when it is not at the equilibrium point, nor does it tell us how the system moves from one point to another when parameter p is changed. To learn that, we need an equation of motion for each process or box in the feedback loop. The equation of motion is usually a differential equation. In real systems the differential equation is often nonlinear and difficult to solve. We first consider models described by linear differential equations, and then we consider some of the behaviors of nonlinear systems.

The response of a system cannot be instantaneous. At equilibrium, the rate of exhaling carbon dioxide is the same as the rate of production throughout the body. If the rate of production rises in a certain muscle group, the extra carbon dioxide enters the blood and is distributed throughout the body, and the carbon dioxide concentrations in the blood and alveoli rise gradually.

To develop a quantitative model, assume that all the carbon dioxide in the body is stored in a single well-stirred compartment of volume V_c. This assumption of uniform concentration is certainly an oversimplification. The total number of moles is n and the concentration is n/V_c. The concentration in the blood is related to the partial pressure in the alveoli by a solubility constant α: $n/V_c = \alpha x$. Therefore $dn/dt = \alpha V_c dx/dt$. Moreover, dn/dt is equal to the rate of production (Eq. 10.2) minus the rate of removal (Eq. 10.3):

$$\frac{dx}{dt} = \frac{Fo}{\alpha V_c} - \frac{x(y-b)}{\alpha V_c RT}.$$

We change the definition of F to take account of the fact that o and p are both the rate of oxygen consumption in slightly different units (o is in mol s^{-1} and p is in mmol min^{-1}):

$$\frac{dx}{dt} = \frac{Fp}{\alpha V_c} - \frac{x(y-b)}{\alpha V_c RT}. \tag{10.14}$$

This differential equation depends on both x and y and in fact is nonlinear since the variables are multiplied together in the last term. At equilibrium $dx/dt = 0$ and Eq. 10.14 gives Eq. 10.4.

If y is constant (a constant breathing rate, which could be accomplished by placing the subject on a respirator), then there is no feedback and Eq. 10.14 is a linear differential equation with constant coefficients:

$$\frac{dx}{dt} + \frac{y_0 - b}{\alpha V_c RT} x = \frac{Fp}{\alpha V_c}.$$

It can be solved using the techniques of Appendix F. Suppose that for $t \le 0$, $p = p_0$, $x = x_0$, and $y = y_0$. For $t > 0$ the subject exercises, so that $p = p_0 + \Delta p$, $x = x_0 + \xi$, and y is unchanged. The equation then becomes

$$\frac{d\xi}{dt} + \frac{y_0 - b}{\alpha V_c RT}\xi = \frac{F\Delta p}{\alpha V_c}. \tag{10.15}$$

The homogeneous equation is

$$\frac{d\xi}{dt} + \frac{1}{\tau_1}\xi = 0, \tag{10.16}$$

where the time constant is

$$\tau_1 = \frac{RT\alpha V_c}{y_0 - b}. \tag{10.17}$$

The homogeneous solution is $\xi = Ae^{-t/\tau_1}$. The particular solution is

$$\xi = \frac{FRT}{y_0 - b}\Delta p = a\,\Delta p,$$

so the complete solution is $\xi = a\,\Delta p + Ae^{-t/\tau_1}$. We now use the initial condition to determine A. At $t = 0$ $\xi = 0$, so $A = -a\,\Delta p$. The complete solution without feedback that matches the initial condition is

$$x - x_0 = a\,\Delta p(1 - e^{-t/\tau_1}). \tag{10.18}$$

Figure 10.7 shows how x changes with time on a plot of x vs. t and a plot of y vs. x. The dots are spaced at equal times.

10.5 Approach to Equilibrium in a Feedback Loop with One Time Constant

Suppose now that y is allowed to change and that $\eta = y - y_0$. We can write the equation for the change in x, Eq. 10.14 as

$$\begin{aligned}\frac{d\xi}{dt} = \frac{dx}{dt} &= \frac{Fp_0}{\alpha V_c} + \frac{F\Delta p}{\alpha V_c} - \frac{(x_0+\xi)(y_0 - b + \eta)}{\alpha V_c RT} \\ &= \underbrace{\frac{Fp_0}{\alpha V_c} - \frac{x_0(y_0-b)}{\alpha V_c RT}}_{=0} + \frac{F\Delta p}{\alpha V_c} - \frac{\xi(y_0-b)}{\alpha V_c RT} \\ &\quad - \frac{x_0\eta}{\alpha V_c RT} - \frac{\xi\eta}{\alpha V_c RT}.\end{aligned}$$

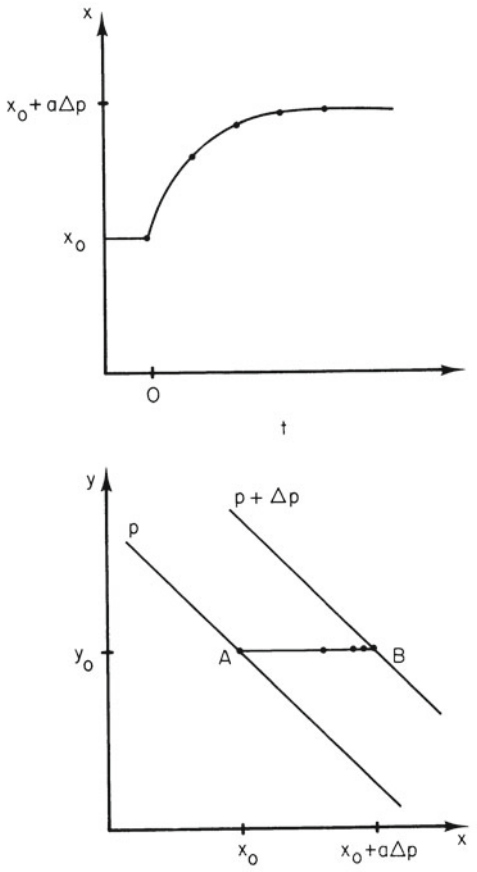


Fig. 10.7 The change in x without feedback in response to a step change in parameter p. **a** Plot of x vs. t. **b** Plot of y vs. x

Multiplying all terms by τ_1 as defined in Eq. 10.17, and identifying

$$G_1 = \left(\frac{\partial g}{\partial y}\right)_p = -\frac{x_0}{y_0 - b},$$

we obtain

$$\tau_1 \frac{d\xi}{dt} = a\,\Delta p - \xi + G_1\eta - \frac{\xi\eta}{y_0 - b}. \tag{10.19}$$

The product $\xi\eta$ in the last term makes the equation nonlinear. If we assume that the last term can be neglected, we have a linear differential equation

$$\tau_1 \frac{d\xi}{dt} = a\,\Delta p - \xi + G_1\eta. \tag{10.20}$$

Now assume that the response of the second box is linear and instantaneous, so that

$$\eta = G_2\xi. \tag{10.21}$$

If this is substituted in the linearized equation, Eq. 10.20, the result is

$$\tau_1 \frac{d\xi}{dt} + (1 - G_1G_2)\xi = a\,\Delta p. \tag{10.22}$$

The steady-state solution before $t = 0$ is $x_0 = a\,p_0/(1 - G_1G_2)$. At $t = 0$ the oxygen demand is changed to $p_0 + \Delta p$. The new steady-state (inhomogeneous) solution is $\xi = a\,\Delta p/(1 - G_1G_2)$ and the homogeneous solution is $\xi = Ae^{-t/\tau}$ where the time constant is

$$\tau = \frac{\tau_1}{1 - G_1G_2}. \tag{10.23}$$

(You can show this by dividing each term in Eq. 10.22 by τ_1 and comparing it to the equation for exponential decay.) After combining the homogeneous and inhomogeneous solutions and using the initial condition $\xi(0) = 0$ to determine A, we obtain the final result:

$$\xi = x - x_0 = \frac{a\,\Delta p}{1 - G_1G_2}\left(1 - e^{-t/\tau}\right). \tag{10.24}$$

This solution has the same form as Eq. 10.18. Both the total change in x and the time constant have been reduced by the factor $1/(1 - G_1G_2)$. The change in y can be determined from $\eta = G_2\xi$. The new solution is plotted along with the old solution in Fig. 10.8. This plot is for a system in which the OLG is $G_1G_2 = -1.3$. The time constant and the change in x are both reduced by $1/2.3$.

It is important to realize that although the feedback reduced the time constant, it has not made x change faster. The curve of $x(t)$ with feedback has always changed less than the curve without feedback, and it has always changed more slowly. The reduction in time constant occurs because x does not change as much with feedback present, so it reaches its asymptotic value more quickly.

This result assumes that box f has a negligible time constant. Applied to the respiratory example, it means that the carbon dioxide-sensing system responds rapidly compared to the time it takes for carbon dioxide levels within the blood to change after a change in p. Figure 10.9a repeats Fig. 10.8 and shows the changes in x and y resulting from a step change in p. When the second time constant is negligible, y is always proportional to x and the system moves back and forth along line AB.

The CO_2 sensors actually take a while to respond. To see what effect this might have, imagine the extreme case where the sensors are very slow compared to the change of carbon dioxide concentration in the blood. In that case, when p changes, y does not change right away. The system behaves at first as if there were no feedback, moving from point A to point C in Fig. 10.9b. As the feedback slowly takes effect, the system moves from C to B. When the exercise ends, the system moves to point D because the subject is breathing too hard. Then it finally moves from D back to A. The actual

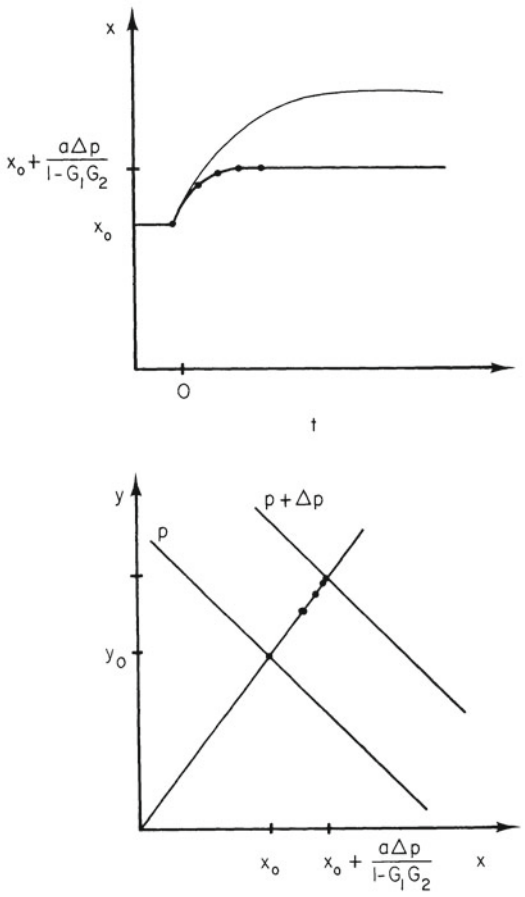


Fig. 10.8 The change in x with feedback in response to a step change in parameter p. **a** Plot of x vs. t. The change in x without feedback is shown for comparison. **b** Plot of y vs. x

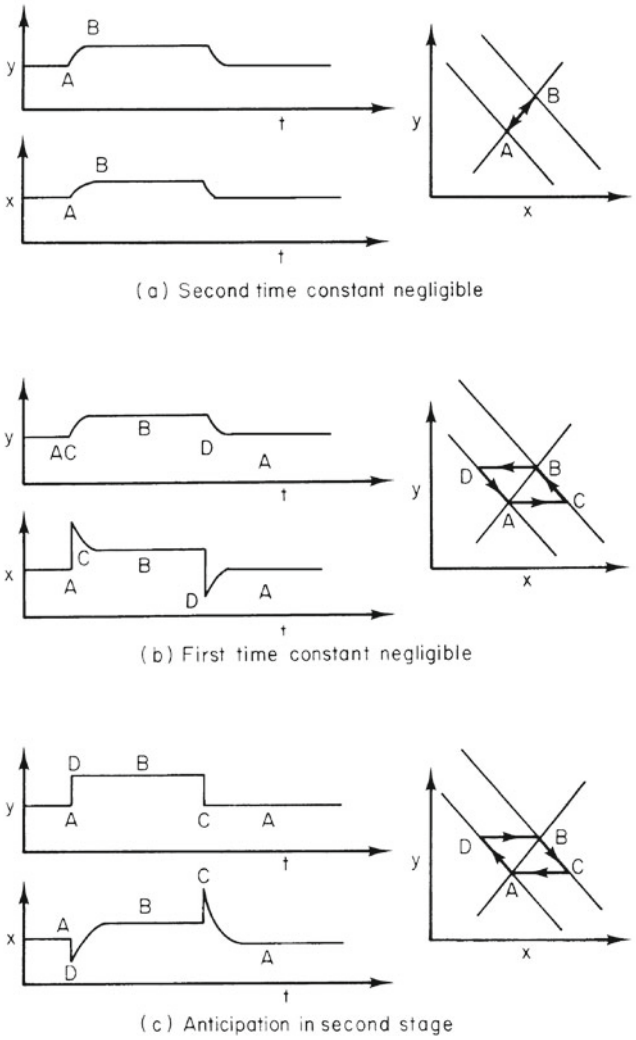


Fig. 10.9 Changes in x and y after a step change in parameter p. **a** The second time constant is negligible compared to the first; x and y move exponentially to their new equilibrium values. **b** The first time constant is negligible; the slow change in y means that there is no feedback at first. **c** The second stage anticipates the change in y that will be required; there is too much feedback at first

system behaves in a manner somewhere between these two extremes, as we will see in the next section.

Consider a third possibility, that a regulatory mechanism anticipates the increased metabolic demand. This might happen if we took deep breaths before we began to exercise, or if additional muscle movement signaled the respiratory control center before the carbon dioxide concentration had a chance to change. Suppose that such anticipation is the only feedback mechanism. With the initiation of exercise, y changes to its final value. The level of carbon dioxide has not yet built up, so the increased ventilation reduces x below its normal value. We can approximate this by point D in Fig. 10.9c. As the increased activity drives x up, the system moves at constant y to point B. When the exercise stops, y drops immediately to the resting value, though carbon dioxide is still coming out of the muscles. The result is that x rises to point C before finally falling back to point A.

Figure 10.10 shows what actually happens in the control of respiration. There is a fast neurological control and a slower chemical control. The result is a combination of the processes in Figs. 10.9a–10.9c.

If we had not made the linear approximation we would not have been able to solve the equation, but the behavior would have been very similar. The nonlinear equation is obtained by substituting the equation for the second box, Eq. 10.21, in Eq. 10.19 instead of Eq. 10.20. The resulting equation is

$$\tau_1 \frac{d\xi}{dt} = a\,\Delta p - (1 - G_1G_2)\xi - \frac{G_2\xi^2}{y_0 - b}. \qquad (10.25)$$

Both this and the linear version are plotted in Fig. 10.11 for $a\,\Delta p = 0$. In each case $d\xi/dt$ is positive when $\xi < 0$ and negative when $\xi > 0$, so ξ approaches zero as time goes on.

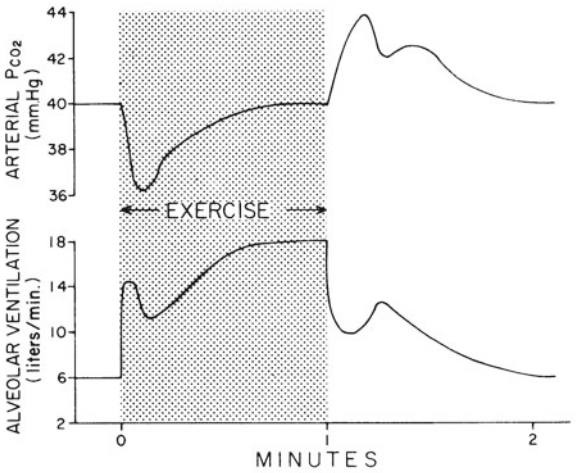


Fig. 10.10 Change of arterial P_{CO_2} and alveolar ventilation in response to exercise. Note that $x = P_{CO_2}$ is in the upper graph and the ventilation rate y is in the lower graph, the opposite of Fig. 10.9. (Reprinted from Guyton (1995) with the permission of Elsevier. Data are extrapolated to humans from dogs. The dog experiments are described in C. R. Bainton (1972). J Appl Physiol 33: 778–787)

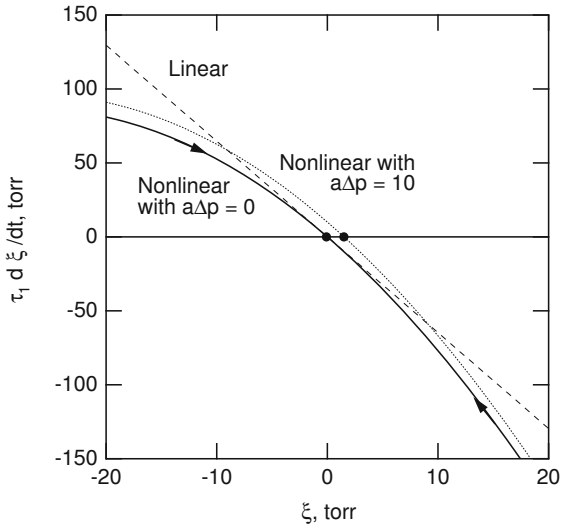


Fig. 10.11 Plots of $\tau_1(d\xi/dt)$ vs. ξ. The straight dashed line is the linear approximation, Eq. 10.22. The parabolas are plots of the nonlinear equation, Eq. 10.25, for two different values of parameter $a\Delta p$. The closed circles show stable fixed points

The direction of evolution of ξ is shown by the arrows on the nonlinear curve. This is often called a *one-dimensional flow*. The variable ξ "flows" to the origin, which is called a *stable fixed point* of the flow. If we change $a\ \Delta p$ to 10, the curve shifts as indicated by the dotted line, and the fixed point moves to a slightly different value of ξ.

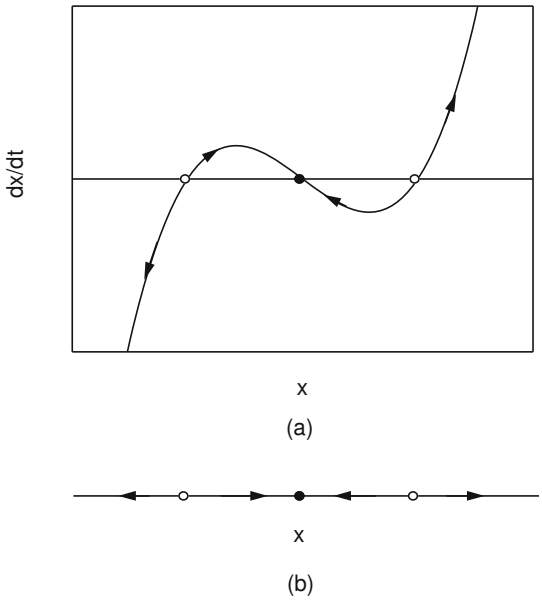


Fig. 10.12 Graphical analysis of the solution to a more complicated differential equation $dx/dt = f(x)$. **a** Plot of dx/dt vs. x. The arrows show the direction that x changes. The open circles show unstable fixed points, and the filled circle is a stable fixed point. **b** The fixed points and the direction of flow are shown on the x axis

This is a particular case of a differential equation in one dependent variable of the form $dx/dt = f(x)$. A great deal about the solution to the general equation can be learned by graphing it as we have done above. When the derivative is positive the function increases with time, and when it is negative it decreases. Figure 10.12a shows a more complicated function, with arrows showing the direction of the flow. The stable fixed point is indicated by a solid circle. There are two *unstable fixed points*, indicated by open circles. If x has precisely the value of an unstable fixed point, it remains there because $dx/dt = 0$. However, if it is displaced even a small amount, it flows away from the unstable point. Figure 10.12b shows just the x axis with the fixed points and the arrows. Stable fixed points are often called *attractors* or *sinks*. The unstable fixed points are called *repellers* or *sources*. Chapter 2 of Strogatz (1994) has an excellent and detailed discussion of one-dimensional flows.

10.6 A Feedback Loop with Two Time Constants

In the preceding section we considered a feedback loop in which only one process had a significant time constant. The other process responded "instantaneously;" its time constant was much shorter. Here we consider the case in which both processes have comparable time constants. We will see that it

is possible for such a (linear) system to exhibit damped sinusoidal behavior in response to an abrupt change in one of the parameters. Whether it does or not depends on the relative values of the two time constants and the open-loop gain. We consider both graphical and analytical techniques for solving this problem.

In earlier sections we discussed control of breathing. Equation 10.20 was the linear model for the departure of one variable from equilibrium:

$$\tau_1 \frac{d\xi}{dt} = -\xi + G_1\eta + a\,\Delta p.$$

For the second process, instead of $\eta = G_2\xi$ we assume that the behavior is given by an analogous equation

$$\tau_2 \frac{d\eta}{dt} = -\eta + G_2\xi. \tag{10.26}$$

For negative feedback either G_1 or G_2 must be negative.

We have a special case of a pair of first-order differential equations

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2), \\ \frac{dx_2}{dt} &= f_2(x_1, x_2). \end{aligned} \tag{10.27}$$

(Here x_1 and x_2 are general variables and have no relationship with the breathing problem considered earlier.)

We first combine the two first-order equations to make a second-order equation which, because we are using linear equations, can be solved exactly. To do this, differentiate Eq. 10.20:

$$\tau_1 \frac{d^2\xi}{dt^2} + \frac{d\xi}{dt} = a\frac{dp}{dt} + G_1\frac{d\eta}{dt}.$$

Substitute Eq. 10.26 in this and obtain

$$\tau_1 \frac{d^2\xi}{dt^2} + \frac{d\xi}{dt} = -\frac{G_1}{\tau_2}\eta + \frac{G_1G_2}{\tau_2}\xi + a\frac{dp}{dt}.$$

To eliminate η, solve Eq. 10.20 for $G_1\eta$ and substitute it in this equation:

$$\tau_1 \frac{d^2\xi}{dt^2} + \frac{d\xi}{dt} = -\frac{\tau_1}{\tau_2}\frac{d\xi}{dt} - \frac{1}{\tau_2}\xi + \frac{a}{\tau_2}p + \frac{G_1G_2}{\tau_2}\xi + a\frac{dp}{dt}.$$

After like terms are combined, the result is

$$\frac{d^2\xi}{dt^2} + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)\frac{d\xi}{dt} + \frac{1 - G_1G_2}{\tau_1\tau_2}\xi = \frac{a}{\tau_1\tau_2}p(t) + \frac{a}{\tau_1}\frac{dp}{dt}. \tag{10.28}$$

This is another linear differential equation with constant coefficients. The right-hand side is a known function of time, since $p(t)$ is known. The homogeneous equation is very common in physics and is called the *harmonic oscillator equation*. It is usually written in the form

$$\frac{d^2\xi}{dt^2} + 2\alpha\frac{d\xi}{dt} + \omega_0^2\xi = 0, \tag{10.29a}$$

with the identifications

$$2\alpha = \frac{1}{\tau_1} + \frac{1}{\tau_2} = \frac{\tau_1 + \tau_2}{\tau_1\tau_2} \tag{10.29b}$$

and

$$\omega_0^2 = \frac{1 - G_1G_2}{\tau_1\tau_2}. \tag{10.29c}$$

Appendix F shows that as long as $\alpha \geq \omega_0$, the system is critically damped or overdamped and there will be no oscillation or *ringing*. This will be the case if

$$\frac{(\tau_1 + \tau_2)^2}{4\tau_1^2\tau_2^2} \geq \frac{1 - G_1G_2}{\tau_1\tau_2}$$

or

$$\frac{(\tau_1 + \tau_2)^2}{4\tau_1\tau_2} \geq 1 - G_1G_2. \tag{10.30}$$

This equation is symmetric in τ_1 and τ_2. The important parameter is $r = \tau_1/\tau_2$. There is no ringing when

$$\frac{(1 + r)^2}{4r} \geq 1 - G_1G_2.$$

Since the feedback is negative, $G_1G_2 = -|G_1G_2|$. Then there is no ringing if

$$|G_1G_2| < \frac{r}{4} + \frac{1}{4r} - \frac{1}{2}, \quad G_1G_2 < 0. \tag{10.31}$$

If the two time constants are equal ($r = 1$), the right-hand side of Eq. 10.31 is zero. There will be ringing if the open-loop gain has a magnitude greater than zero. For large values of r (say $r > 10$), the equation is approximately $|G_1G_2| < r/4$. If the magnitude of the open-loop gain is larger than this, there will be ringing.

We can see the general behavior of Eqs. 10.20 and 10.26 by examining the behavior of the derivatives. Both derivatives are zero and there is a fixed point when

$$\xi = \frac{a\,\Delta p}{1 - G_1G_2}, \quad \eta = \frac{G_2 a\,\Delta p}{1 - G_1G_2}.$$

For $\Delta p = 0$ the fixed point is at the origin. Figures 10.13 and 10.14 show plots of ξ and η for different values of the gain and damping. The plots of η vs. ξ are called *state-space* plots or *phase-space* or *phase-plane* plots. The plots shown here are spiral to the fixed point. Depending on the values of the gains and time constants (try positive feedback) there can also be exponentially growing solutions. An extensive literature exists analyzing stability for both Eqs. 10.27 and their linearized versions. See Chaps. 5 and 6 of Strogatz (1994) or Chap. 3 of Hilborn (2000). For a visual but nonmathematical analysis, see Abraham and Shaw (1992).

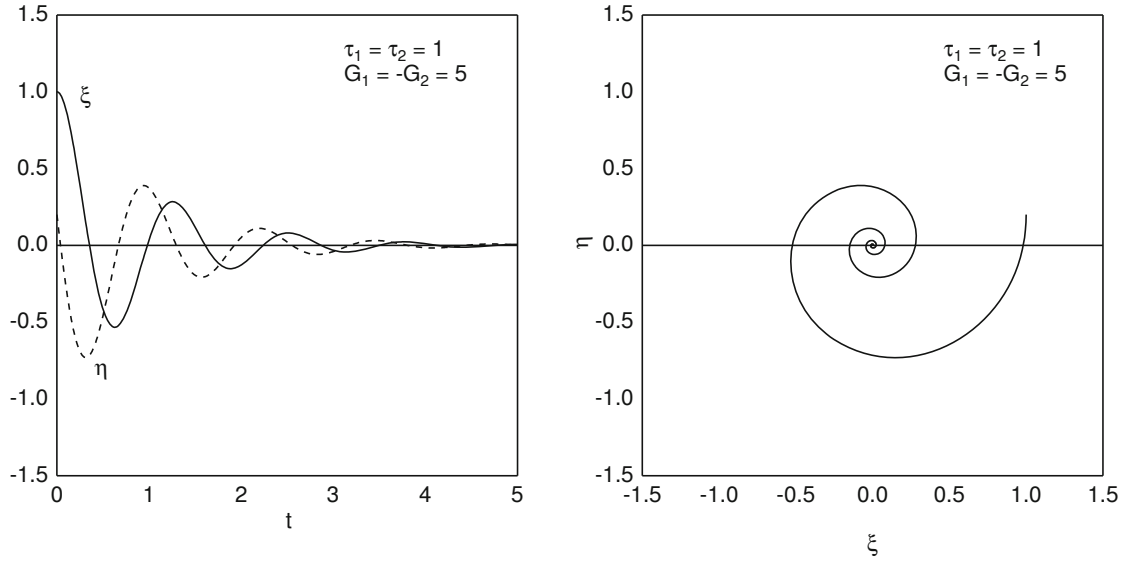


Fig. 10.13 A solution to Eq. 10.28 is plotted that has a value 1 and time derivative 0 when $t = 0$. The variable η is obtained from ξ by using Eq. 10.26. Plots of ξ and η vs. t are shown on the left. A state-space plot of η vs ξ is shown on the right

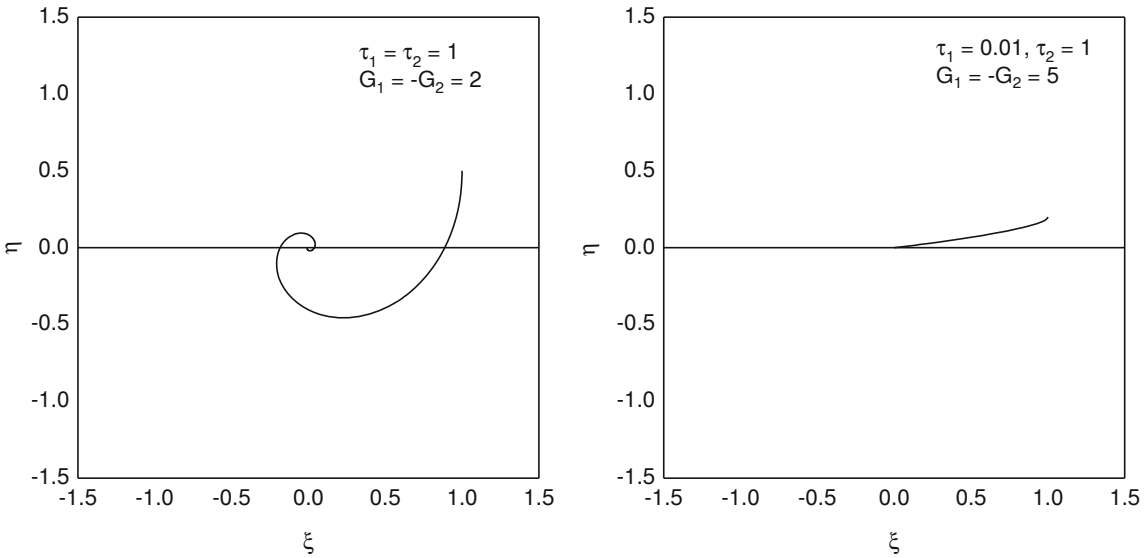


Fig. 10.14 Additional state-space plots for the same initial conditions as in Fig. 10.13, but with different values of the parameters

10.7 Proportional, Derivative, and Integral Control

We have been looking at a particular example of a *control system*. In Sect. 10.4 and 10.5 we considered a system for which, for $t \leq 0$, $p = p_0$, $x = x_0$, and $y = y_0$. For $t > 0$ the subject exercised, so that $p = p_0 + \Delta p$, $x = x_0 + \xi$, and $y = y_0 + \eta$. There were two equations of motion, Eqs. 10.20 and 10.21:

$$\tau_1 \left(\frac{d\xi}{dt} \right) + \xi = a\Delta p + G_1 \eta. \tag{10.32}$$

$$\eta = G_2 \xi. \tag{10.33}$$

Combining these we find the steady state solution is

$$\xi = \frac{a\Delta p}{1 - G_1 G_2}. \tag{10.34}$$

The goal of a control system is to make $\xi = x - x_0$ as small as possible as p is varied. In the language of control theory x_0 is called the *set point* or *reference quantity*, x is called the *present value* or *actual output*, and $e = x_0 - x = -\xi$ is called the *controller error*.

At $t = 0$ there is a step change Δp. Without feedback (G_1 or $G_2 = 0$), this causes a change in the steady-state

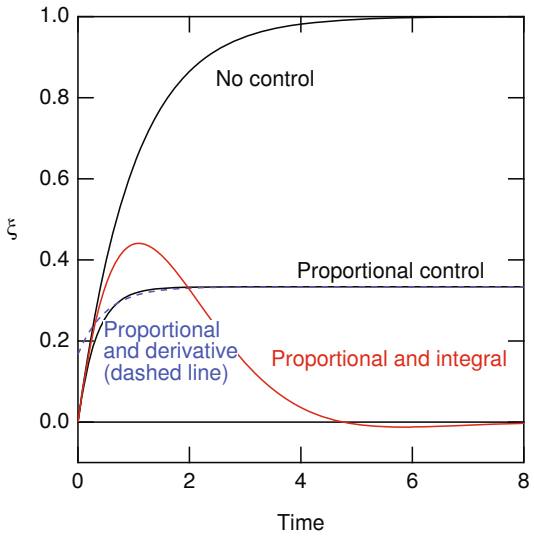


Fig. 10.15 Examples of the different kinds of control: none, proportional, proportional with derivative, and proportional with integral. Derivative reaches the steady-state value more quickly. Integral control allows the error, ξ, to become zero

value $\xi = a\Delta p$. Feedback reduces this to $\xi = a\Delta p/(1 - G_1G_2)$. This is called *proportional control*. The input to the controller (the second box in Fig. 10.4) is ξ.

In *derivative control* an additional signal is introduced on the right hand side of Eq. 10.33, which is proportional to the derivative of the error signal. Usually a combination of proportional and derivative control is used:

$$\eta = G_2\xi + G_2\tau_d\frac{d\xi}{dt}.$$

Factor τ_d has the dimensions of time and gives the relative strength of the derivative control term. Combining this with Eq. 10.32 gives us

$$\tau_1\left(\frac{d\xi}{dt}\right) + \xi = a\Delta p + G_1G_2\xi + \tau_dG_1G_2\left(\frac{d\xi}{dt}\right).$$

Regrouping gives

$$\frac{d\xi}{dt} + \frac{(1 - G_1G_2)}{(\tau_1 - \tau_dG_1G_2)}\xi = \frac{a\Delta p}{(\tau_1 - \tau_dG_1G_2)}.$$

In *integral control* a term

$$\frac{G_2}{\tau_i}\int_0^t \xi(t)dt$$

is added to Eq. 10.33. For a combination of proportional and integral control we have

$$\tau_1\left(\frac{d\xi}{dt}\right) + \xi = a\Delta p + G_1G_2\xi + G_1G_2\frac{1}{\tau_i}\int_0^t \xi(t)dt.$$

This can be rearranged as

$$\tau_1\left(\frac{d\xi}{dt}\right) + (1 - G_1G_2)\xi + \frac{G_1G_2}{\tau_i}\int_0^t \xi(t)dt = a\Delta p.$$

Differentiating gives us the harmonic oscillator equation:

$$\frac{d^2\xi}{dt^2} + \frac{(1 - G_1G_2)}{\tau_1}\frac{d\xi}{dt} + \frac{G_1G_2}{\tau_1\tau_i}\xi = 0.$$

Proportional, proportional plus derivative, and proportional plus integral control are compared in Fig. 10.15. Derivative control causes the abrupt jump in ξ at $t = 0$. However, the offset or steady-state value remains the same as in proportional control. Integral control can reduce the offset to zero, but oscillations are likely. El-Samad et al. (2002) and Khammash and El-Samad (2004) describe proportional and integral control of calcium regulation in cows that have just given birth. LeDuc et al. (2011) review control techniques and how they can be used to manipulate cells in the laboratory.

10.8 Models Using Nonlinear Differential Equations

We have used many models in this book. In Chap. 2 we introduced a linear differential equation that leads to exponential growth or decay, and we used it to model tumor and bacterial growth and the movement of drugs through the body. We briefly examined some nonlinear extensions of this model. In Chap. 4 we modeled diffusion processes with linear equations—Fick's first and second laws—and we used a linear model to describe solvent drag. In Chap. 5 we used the model of a right-cylindrical pore. In Chap. 6 we used both a linear model—electrotonus—and a nonlinear model—the Hodgkin–Huxley equations. In this chapter we introduced a linear model for feedback, and we saw how two linear processes in a feedback loop could lead to oscillations, the linear harmonic oscillator.

Linear models have one advantage: they can be solved exactly. But most processes in nature are not linear. Jules Henri Poincaré realized around 1900 that systems described exactly by the completely deterministic equations of Newton's laws could exhibit wild behavior. Poincaré was studying the three-body problem in astronomy (such as Sun–Earth–Moon). While we are all familiar with the fact that the motion of the Sun–Earth–Moon system is evolving smoothly with time and that eclipses can be predicted centuries in advance, this smooth behavior does not happen for all systems. For certain ranges of parameters (such as the masses of the objects and initial positions and velocities) the solutions can exhibit behavior that is now termed *chaotic*. If we consider

the motion that results from two sets of initial conditions that differ from each other only by an infinitesimal amount in one of the variables, we find that in chaotic behavior there can be solutions that diverge exponentially from each other as time goes on, even though the solutions remain bounded. Poincaré developed some geometrical techniques for studying the behavior of such systems. Thorough study of nonlinear systems requires the use of a digital computer. As a result, it has only been since the 1970s that we have realized how often chaotic behavior can occur in a system governed by deterministic equations. With computers we have gained more insight into the properties of chaotic behavior.

Just as the harmonic oscillator provides a model for behavior seen in many contexts from electric circuits to shock absorbers in automobiles to the endocrine system, certain features of nonlinear models have wide applicability. These include period doubling, the ability to reset the phase (timing) of a nonlinear oscillator, and deterministic chaos.

Some have said that Newtonian physics has been overthrown by chaos. This is not true. The same equations hold; predictable motions with which we have long been familiar still take place. Much of our current technology is based on them. We build television sets and send a spacecraft to explore several planets in succession. With chaos, we have come to understand a rich set of solutions to these same equations that we were not equipped to study before.

Many books about nonlinear systems have been written. A particularly interesting one for this audience is by Kaplan and Glass (1995). It is written for biologists and has many clear and relevant examples. Others are by Glass and Mackey (1988), by Hilborn (2000), and by Strogatz (1994).

Space limitations prevent more than a brief hint at some of the features of nonlinear dynamics, here and in Chap. 11. In this section we will discuss some one- and two-dimensional nonlinear differential equations. These will not lead to chaos, but will allow us to describe a very simple model for *phase resetting*. In Sect. 10.9 we will discuss equations that exhibit chaotic behavior.

10.8.1 Describing a Nonlinear System

Suppose that a nonlinear system with N variables can be described by a set of N first-order differential equations:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, \ldots, x_N), \\ \frac{dx_2}{dt} &= f_2(x_1, x_2, \ldots, x_N), \\ &\ldots, \\ \frac{dx_N}{dt} &= f_N(x_1, x_2, \ldots, x_N). \end{aligned} \tag{10.35}$$

(These are an extension of the pair of differential equations we saw as Eqs. 10.27. Our model of breathing had two variables. It would be more realistic to use a breathing model with more variables, since alveolar ventilation also depends on arterial pH, weakly on oxygen partial pressure, and on the nervous factors that were described earlier.)

If the equations are cast in this form with N variables, then N initial conditions are required, corresponding to the constant of integration required for each equation. It is customary to say that there are N *degrees of freedom*. This is the language of *system dynamics*. This definition of degrees of freedom is different from what we used in Chap. 3, where each degree of freedom was represented by a second-order differential equation ($d^2x/dt^2 = F_x/m$, for example) and two initial conditions were required for each degree of freedom.

We can put Newton's second law in this form by writing two first-order differential equations instead of one second-order equation. For motion in one dimension, instead of

$$m\frac{d^2x}{dt^2} = F(x, v),$$

we write a pair of first-order equations:

$$\frac{dv}{dt} = \frac{F(x, v)}{m}, \quad \frac{dx}{dt} = v.$$

This system has two degrees of freedom in our new terminology. In either description, two initial conditions are required.

In many situations the force (or more generally the functions on the right-hand side of Eqs. 10.35) is time dependent. In standard form, the functions on the right do not depend on time. This is remedied by introducing one more variable, $x_{N+1} = t$. The additional differential equation is $dx_{N+1}/dt = 1$.

The evolution or " motion" of the system can be thought of as a trajectory in N-dimensional space, starting from the point that represents the initial conditions. Time is a parameter. We have seen an example of this for two dimensions in Figs. 10.13 and 10.14. It is possible to prove that two distinct trajectories cannot intersect in a finite period of time and that a single trajectory cannot cross itself at a later time (see Hilborn 2000, p. 77 or Strogatz 1994, p. 149). This is true in the full N-dimensional space; if we were to measure only two variables, we could see apparent intersections in the state plane that we were observing. This means that chaotic behavior, in which variables appear to change wildly and two-dimensional trajectories appear to cross, does not occur for a pair of differential equations of the form in Eqs. 10.35. At least three variables are required. A system with two degrees of freedom that is externally driven[2] can

[2] That is, one of the functions on the right-hand side of the set of equations depends on time.

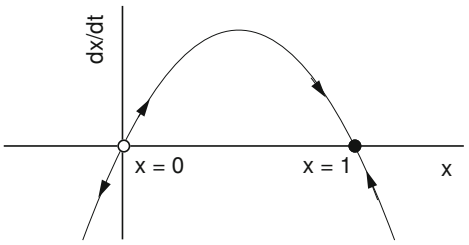


Fig. 10.16 Plot of dx/dt vs. x for the logistic differential equation

exhibit chaotic behavior because of the additional variable x_{N+1} that is introduced.

10.8.2 An Example of Phase Resetting: The Radial Isochron Clock

In Chap. 2 we studied the logistic differential equation

$$\frac{dy}{dt} = by\left(1 - \frac{y}{y_\infty}\right).$$

It is convenient to rewrite the logistic equation in terms of the dimensionless variable $x = y/y_\infty$:

$$\frac{dx}{dt} = bx(1 - x). \tag{10.36}$$

This separates the scale factor y_∞ from the dynamic factor b that tells how rapidly y and x are changing.[3] A plot of dx/dt vs. x is shown in Fig. 10.16. There is an unstable fixed point at $x = 0$ and a stable fixed point at $x = 1$. The logistic equation is one of a whole class of nonlinear first-order differential equations for which dx/dt as a function of x has a maximum. It has been studied extensively because of its relative simplicity, and it has been used for population modeling. (Better population models are available.[4] The logistic model assumes that the population is independent of the populations of other species, that the growth of the species does not affect the carrying capacity y_∞, and that the population increases smoothly with time.)

Many of the important features of nonlinear systems do not occur with one degree of freedom. We can make a very simple model system that displays the properties of systems with two degrees of freedom by combining the logistic equation for variable r with an angle variable θ that increases at a constant rate:

$$\frac{dr}{dt} = ar(1 - r), \quad \frac{d\theta}{dt} = 2\pi. \tag{10.37}$$

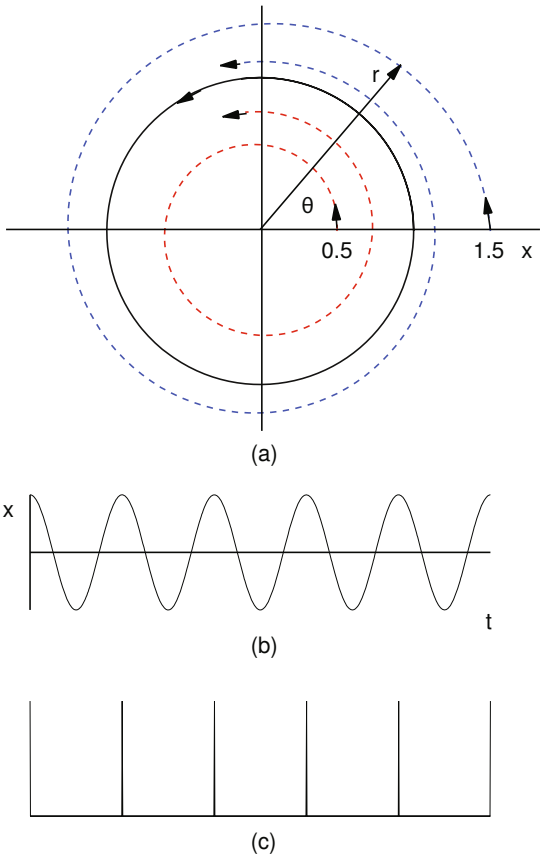


Fig. 10.17 A system with two degrees of freedom. **a** The limit cycle is represented by the solid circle. Systems starting elsewhere in the plane have trajectories that approach the limit cycle as $t \to \infty$, as shown by the dashed lines. **b** The value of $x = r\cos\theta$ is plotted as a function of time. **c** A timing pulse is generated every time θ is a multiple of 2π

We can interpret (r, θ) as the polar coordinates of a point in the xy plane. When t has increased from 0 to 1 the angle has increased from 0 to 2π, which is equivalent to starting again with $\theta = 0$. This model system has been used by many authors. Glass and Mackey (1988) have proposed that it be called the *radial isochron clock*. Typical behavior is shown in Fig. 10.17a. If $r = 1$, there is a circular orbit corresponding to the stable fixed point of Eq. 10.36. Such a stable orbit is called a *stable limit cycle*.[5] There is an unstable limit cycle, $r = 0$, corresponding to the unstable fixed point of Eq. 10.36. Any initial conditions except $r = 0$ give trajectories that move toward the stable circular limit cycle as time progresses. The set of points in the xy plane lying on orbits that move to the limit cycle as $t \to \infty$ is called the *basin of attraction* for the limit cycle. In this case the basin of attraction

[3] We could also, if we wish, define a new time scale, $t' = bt$, and deal with the completely dimensionless equation $dx/dt' = x(1 - x)$.

[4] See Begon et al. (1996).

[5] A *stable limit cycle* is an oscillation in the solutions to a set of differential equations that is always reestablished following any small perturbation.

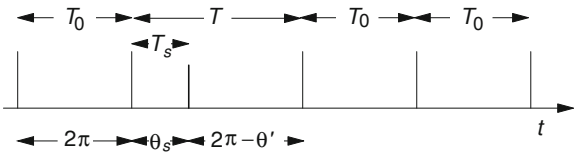


Fig. 10.18 Resetting the phase of an oscillation. The oscillator fires regularly with period T_0. A stimulus at time T_s after it has fired causes a period of length T, after which the periods are again T_0

includes all points except the origin. If we look at the time behavior, Fig. 10.17b shows the behavior of $x = r\cos\theta$ on the limit cycle. The oscillator might provide timing information as the phase moves through some value. Figure 10.17c shows a series of pulses every time θ is a multiple of 2π.

In many cases the differential equations contain one or more parameters that can be varied, and the number and shape of the limit cycles change as the parameters are changed. A point in parameter space at which the number of limit cycles changes or their stability changes is called a *bifurcation*. We will see examples of bifurcations in the next section. See the references for a much more extensive discussion.

One important characteristic of nonlinear oscillators is that a single pulse can reset their phase. If they are subject to a series of periodic pulses they can be *entrained* to oscillate at the driving frequency. (The nonlinear oscillators that sweep the electron beam across the screen of a television tube are entrained by synchronization pulses in the television signal.) Our simple two-dimensional model exhibits phase resetting that is very similar to that exhibited by cardiac tissue.[6]

Suppose that a cardiac pacemaker depolarizes every T_0 seconds and that it can be modeled by our radial isochron clock. Assume that depolarization occurs when $\theta = 0$ or a multiple of 2π. A stimulus is applied at time T_s after the beginning of the cycle, as shown in Fig. 10.18. As a result, the time from the previous depolarization to the next one is changed to T, after which the period reverts to T_0. (In a real experiment, it may be necessary to wait several cycles before measuring so that any transient behavior has time to decay, and then extrapolate back to find the value of T.) Often a stimulus early in the cycle is found to delay the next depolarization, while a stimulus late in the cycle advances it. Our model provides a simple geometric interpretation of this behavior, independent of any knowledge of the detailed dynamics.

[6] This discussion is based on Glass and Mackey (1988), p. 104 ff. See also the works by Winfree (1987, 2001). Strogatz (2003) discusses phase resetting and other nonlinear phenomena in an engaging and nonmathematical manner.

A delayed depolarization is shown in Fig. 10.18. Pulses are occurring every T_0 seconds when the phase is a multiple of 2π (that is, 0). A stimulus is applied at a time T_s after the previous pulse, at which time the phase is θ_s. Since the phase advances linearly, we have the proportion

$$\frac{T_s}{T_0} = \frac{\theta_s}{2\pi}.$$

Suppose the stimulus causes the system to move to a new state with a phase θ', which we do not yet know. Since in our model $d\theta/dt$ is constant, the phase advances after the stimulus at the same rate as it would have without the stimulus. The next pulse occurs when the phase again reaches 2π. This occurs at a time T after the previous pulse, or a time $T - T_s$ after the stimulus, when the phase has increased from θ' to 2π. Therefore

$$\frac{T - T_s}{T_0} = \frac{2\pi - \theta'}{2\pi}$$

and

$$\frac{T}{T_0} = \frac{2\pi + \theta_s - \theta'}{2\pi}. \tag{10.38}$$

We use our limit cycle model to relate θ_s and θ' as shown in Fig. 10.19. The system has been moving on a circle of unit radius representing the stable limit cycle. Assume that the only effect of the stimulus is to shift the value of x by a distance b along the $+x$ axis. For the angles shown in Fig. 10.19a this results in a point with $\theta' < \theta_s$, a delay in the phase or $T > T_0$. For an initial angle in the lower half plane (Fig. 10.19b) it results in $\theta' > \theta_s$ and $T < T_0$. The relation between the two angles can be obtained from the triangles:

$$\cos\theta' = \frac{\cos\theta_s + b}{\left[(\cos\theta_s + b)^2 + \sin^2\theta_s\right]^{1/2}} = \frac{\cos\theta_s + b}{(1 + b^2 + 2b\cos\theta_s)^{1/2}}.$$

The stimulus changed both θ and r. After the stimulus each evolves independently according to its own differential equation. The trajectory returns to the limit cycle as r returns to its attractor, but the phase is forever altered. Figure 10.20a is a plot of θ' vs. θ_s for two values of b. When $b < 1$, θ' takes on all values, while for $b > 1$, θ' is restricted to values near 0 and 2π. The first case is called a *type-1 phase resetting*, and the second is called a *type-0 phase resetting*. Figure 10.20b combines these results with Eq. 10.38 to determine T/T_0 as a function of T_s/T_0.

Figure 10.21 shows experimental data for electrotonically stimulated Purkinje fibers from the conduction system of a dog. The fibers were undergoing spontaneous oscillation with $T_0 = 1.575$ s. Stimuli of two different amplitudes were applied at different parts of the cycle, T_s/T_0. Two different curves were obtained. The one with larger current looks like

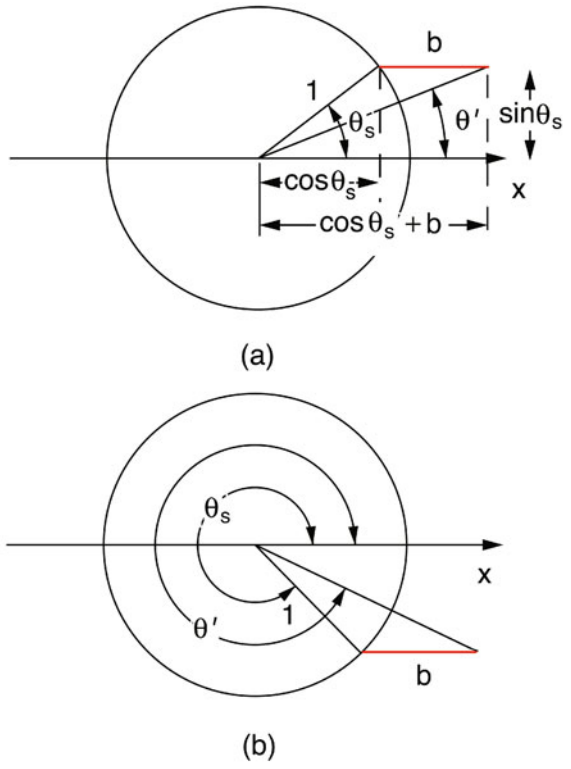


Fig. 10.19 The limit-cycle oscillator model for resetting the phase of an oscillation. At phase θ_s a stimulus changes the value of x by an amount $+b$. **a** For angles $0 < \theta_s < \pi$, this places the system on a trajectory with a smaller phase θ', delaying the next pulse. **b** For angles $\pi < \theta_s < 2\pi$, the stimulus results in a larger phase and the next pulse occurs earlier. The system returns to the limit cycle while θ continues to increase at a constant rate

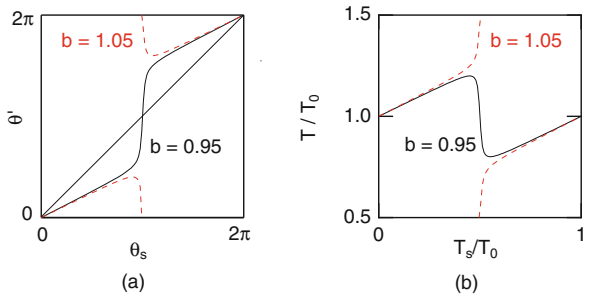


Fig. 10.20 Plots of **a** the new phase vs. the old phase and **b** the length of the period vs. the time when the stimulus is applied

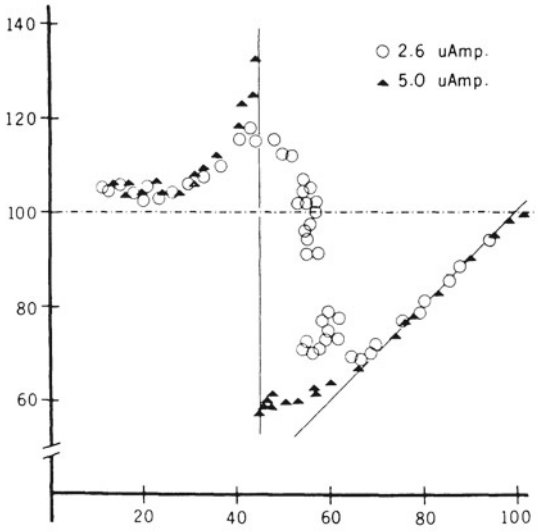


Fig. 10.21 Phase resetting of a spontaneously oscillating Purkinje fiber by stimulation with an electrical impulse. The abscissa is T_s/T_0 expressed as a percentage. The ordinate is T/T_0 expressed as a percentage. Two different stimulus strengths were used. Compare the smaller stimulus (the open circles) to $b = 0.95$ and the larger stimulus (solid triangles) to $b = 1.05$ in Fig. 10.20b. (Reproduced with permission from Jalife and Moe (1976). Copyright 1976 American Heart Association)

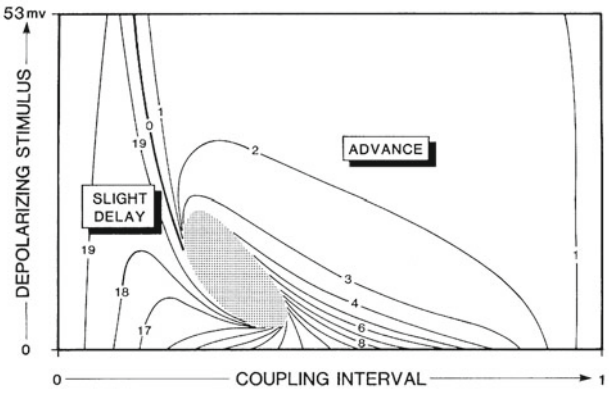


Fig. 10.22 Phase resetting in a Hodgkin–Huxley model. The coupling interval is the delay from the previous pulse to the stimulation pulse in fractions of a period. The ordinate shows the size of the stimulus pulse in mV. The contours show the latency or time from the stimulus to the next pulse, measured in twentieths of a period. (From Winfree (1987). Copyright ©1987. Reproduced by permission of Prof. Arthur Winfree)

the curve with $b = 1.05$ in Fig. 10.20b, while the one with smaller current looks like the curve with $b = 0.95$.

10.8.3 Stopping an Oscillator

It is theoretically possible to apply a stimulus that would put the system at the point $r = 0$ in the state space. In that case it would not oscillate, though for this model $r = 0$ is an unstable equilibrium point and any slight perturbation would lead the system back to the stable limit cycle. In more complicated models it is possible to have a region of state space corresponding to no oscillation and a basin of attraction that leads to it. Figure 10.21 shows the results of a calculation by Winfree (1987) of the effect of stimuli on resetting the phase

of the Hodgkin–Huxley equations adjusted to oscillate spontaneously. The abscissa is the coupling interval or the time after the previous pulse at which the stimulus is delivered. The ordinate is the height of the depolarizing pulse in mV. The contour lines show different values of the latency—the time in twentieths of a cycle period from the stimulus to the next pulse. Winfree called the shaded region of state space where annihilation occurs a "black hole."[7]

10.9 Difference Equations and Chaotic Behavior

We have alluded to the possibility of chaotic behavior, but we have not yet seen it. Chaotic behavior of nonlinear differential equations requires three or more degrees of freedom. It is possible to see chaotic behavior in difference equations with a single degree of freedom because the restriction that the trajectory cannot cross itself or another trajectory no longer applies. It arose from the continuous nature of the trajectories for a system of differential equations.

10.9.1 The Logistic Map: Period Doubling and Deterministic Chaos

We considered the logistic differential equation as a model for population growth. The differential equation assumes that the population changes continuously. For some species each generation is distinct, and a difference equation is a better model of the population than a differential equation. An example might be an insect population where one generation lays eggs and dies, and the next year a new generation emerges. A model that has been used for this case is the logistic difference equation or *logistic map*

$$y_{j+1} = a y_j \left(1 - \frac{y_j}{y_\infty}\right)$$

with $a > 0$ and j the generation number. It can again be cast in dimensionless form by defining $x_j = y_j/y_\infty$:

$$x_{j+1} = a x_j (1 - x_j). \tag{10.39}$$

While superficially this looks like the logistic differential equation, it leads to very different behavior. The stable points are not even the same. A plot of x_{j+1} vs. x_j is a parabola,

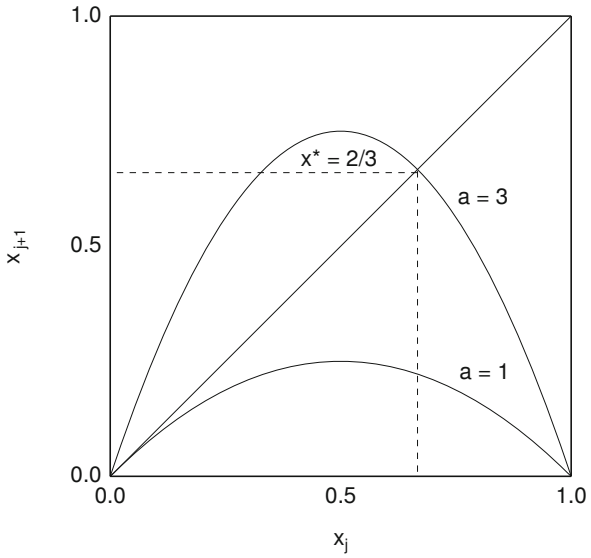


Fig. 10.23 Plot of x_{j+1} vs. x_j for the logistic difference equation or logistic map, for two values of parameter a

from which we can immediately see the following properties of the logistic map:

$$\begin{aligned} x_j < 0, &\quad x_{j+1} < 0, \\ x_j = 0, &\quad x_{j+1} = 0, \\ 0 < x_j < 1, &\quad x_{j+1} > 0, \\ x_j = 1, &\quad x_{j+1} = 0, \\ x_j > 1, &\quad x_{j+1} < 0. \end{aligned}$$

If we are to use this as a population model, we must restrict x to values between 0 and 1 so the values do not go to $-\infty$. In order to keep successive values of the map within the interval (0, 1) we also make the restriction $a < 4$.

For the logistic differential equation, $x = 1$ was a point of stable equilibrium. However, for the logistic map, if $x_j = 1$ the next value is $x_{j+1} = 0$. The equilibrium value x^* can be obtained by solving Eq. 10.39 with $x_{j+1} = x_j = x^*$:

$$x^* = a x^*(1 - x^*) = 1 - 1/a. \tag{10.40}$$

Point x^* can be interpreted graphically as the intersection of Eq. 10.39 with the equation $x_{j+1} = x_j$ as shown in Fig. 10.23. You can see from either the graph or from Eq. 10.40 that there is no solution for positive x if $a < 1$. For $a = 1$ the solution occurs at $x^* = 0$. For $a = 3$ the equilibrium solution is $x^* = 2/3$. Figure 10.24 shows how, for $a = 2.9$ and an initial value $x_0 = 0.2$, the values of x_j approach the equilibrium value $x^* = 0.655$. This equilibrium point is called an *attractor*.

[7] See Winfree (1987), especially Chaps. 3 and 4, or Glass and Mackey (1988), pp. 93–97.

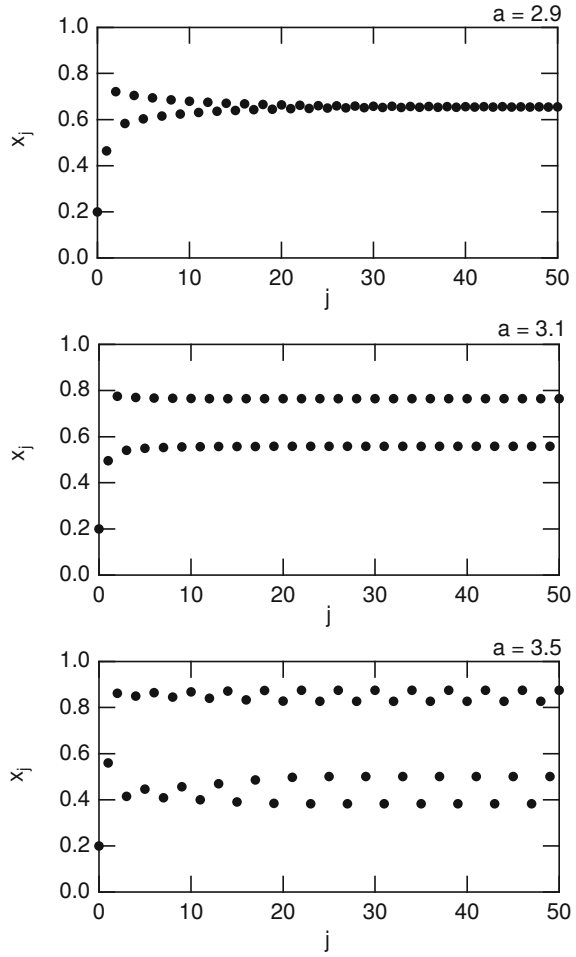


Fig. 10.24 Plots of x_j vs. j for different values of a, showing how the sequence of values converges to one, two, or four values of x called the *attractors*

Figure 10.24 also shows the remarkable behavior that results when a is increased to 3.1. The values of x_j do not come to equilibrium. Rather, they oscillate about the former equilibrium value, taking on first a larger value and then a smaller value. This is called a *period-2 cycle*. The behavior of the map has undergone *period doubling*. What is different about this value of a? Nothing looks strange about Fig. 10.23. But it turns out that if we consider the slope of the graph of x_{j+1} vs. x_j at x^*, we find that for $a > 3$ the slope of the curve at the intersection has a magnitude greater than 1. Many books explore the implications of this.

The period doubling continues with increasing a. For $a > 3.449$ there is a cycle of period 4. A plot of the period-4 cycle for $a = 3.5$ is also shown in Fig. 10.24. For $a > 3.54409$ there is a cycle of period 8. The period doubling continues, with periods 2^N occurring at more and more closely spaced values of a. When $a > 3.569946$, for many values of a the behavior is aperiodic, and the values of x_j never form a repeating sequence. Remarkably, there are ranges of a in this

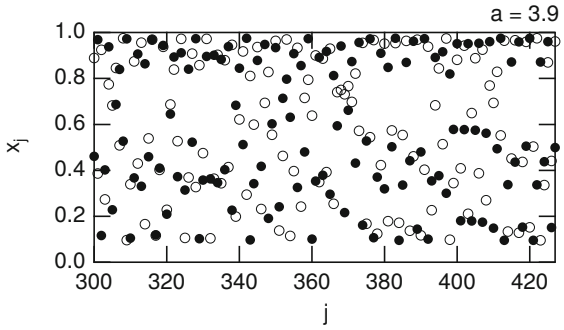


Fig. 10.25 For this value of a the solution is aperiodic. There is no attractor

region for which a repeating sequence again occurs, but they are very narrow. The details of this behavior are found in many texts. In the context of ecology they are reviewed in a classic paper by May (1976).

For $a < 3.569946$, starting from different initial values x_0 leads after a number of iterations to the same set of values for the x_j. For values of a larger than this, starting from slightly different values of x_0 usually leads to very different values of x_j, and the differences become greater and greater for larger values of j. This is shown in Fig. 10.25 for $a = 3.9$. The sequence is plotted from $j = 301$ to $j = 425$. The solid circles represent the sequence starting with $x_0 = 0.20$; the open circles represent the sequence for $x_0 = 0.21$.

This is an example of chaotic behavior or *deterministic chaos*. Deterministic chaos has four important characteristics:

1. The system is deterministic, governed by a set of equations that define the evolution of the system.
2. The behavior is bounded. It does not go off to infinity.
3. The behavior of the variables is aperiodic in the chaotic regime. The values never repeat.
4. The behavior depends very sensitively on the initial conditions.

10.9.2 The Bifurcation Diagram

Figure 10.26 shows the values of x_j that occur after any transients have died away for different values of parameter a. The diagram was made by picking a value of a. A value of x_0 was selected and the iterations were made. After 50 iterations, the next 300 values of x_j were plotted. Then a was incremented slightly and the process was repeated. This is called a *bifurcation diagram*. The figure shows the range $1 < a < 4$. The asymptotic value of x_j rises according to $x^* = 1 - 1/a$ until period doubling occurs at $a = 3$. A four-cycle appears for $a > 3.449$, and for $a > 3.569946$ chaos sets in. Within the chaotic region are very narrow bands of finite periodicity.

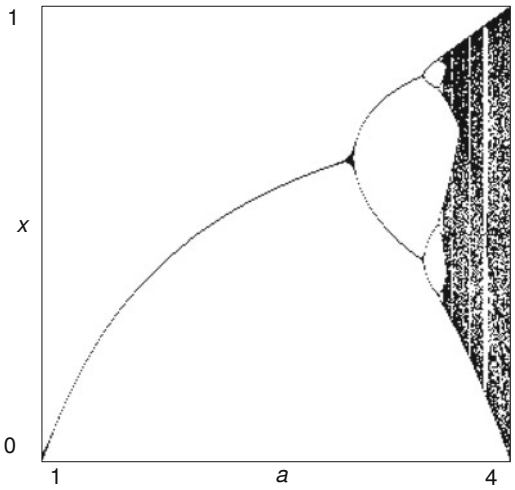


Fig. 10.26 A bifurcation diagram for the logistic map, showing 300 values of x_j for values of a between 1 and 4. The plot was made using the Macintosh software *A Dimension of Chaos* by Matthew A. Hall.

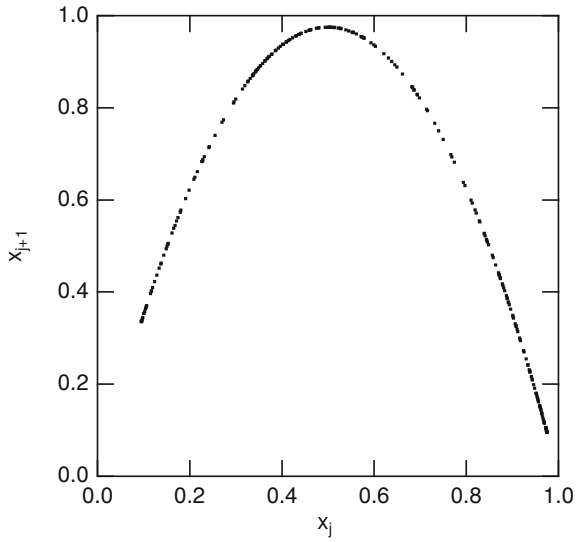


Fig. 10.28 A plot of x_{j+1} vs x_j for the data of Fig. 10.25 recovers the logistic map

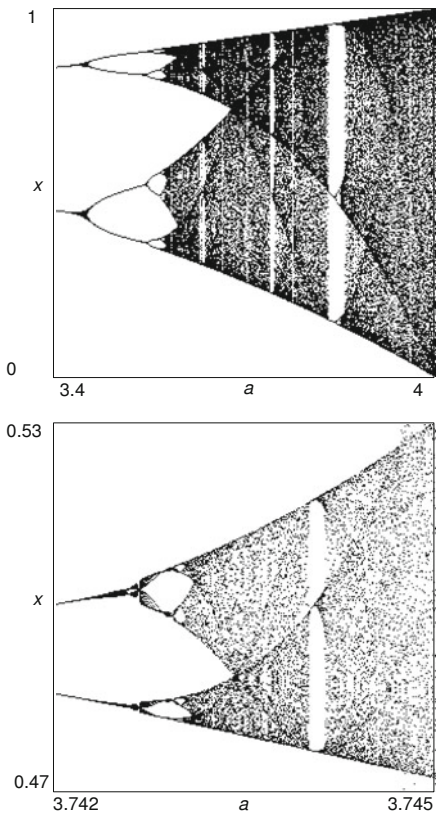


Fig. 10.27 An example of self-similarity. The top curve shows $3.4 < a < 4$. The bottom curve shows $3.742 < a < 3.745$. Note its similarity to the top curve. The plot was made using the Macintosh software *A Dimension of Chaos* by Matthew A. Hall

Figure 10.27 shows a feature of many chaotic systems called *self-similarity*. The bifurcation diagram is plotted for two ranges of a: 3.4–4.0 and 3.743–3.745. The x scale is expanded in the second diagram. Note the similarity of the two bifurcation diagrams.

Even though the plot of x_j vs. j in Fig. 10.25 has no obvious pattern, the values of x_j were obtained from the logistic map. When we plot x_{j+1} vs. x_j the points fall on the map (Fig. 10.28).

The simplest systems in which chaotic behavior can be seen are first-order difference equations in which x_{j+1} is a function of x_j. The function is peaked and "tunable" by some parameter. Chaotic behavior occurs for some values of the parameter. In fact, it appears that the ratios of the parameter values involved in the period doubling and approach to chaos may be independent of the particular shape of the curve.[8]

10.9.3 Quasiperiodicity

Some systems exhibit *quasiperiodicity*. Consider the map $x_{j+1} = x_j + b$ where b is a fixed parameter. Wrap the function back on itself so that x remains in the interval (0, 1). This is done by using the modulo or remainder function.[9] The map is

$$x_{j+1} = x_j + b \,(\text{mod}\, 1)\,. \tag{10.41}$$

[8] See Hilborn (1995), Chap. 2, Kaplan and Glass (1995), p. 30, or Strogatz (1994), p. 370.

[9] The function $x \bmod n$ gives the number that remains after subtracting n from x enough times so that the result is less than 1. For example, $1.5742 \bmod 1 = 0.5742$; $7.5 \bmod 1 = 0.5$.

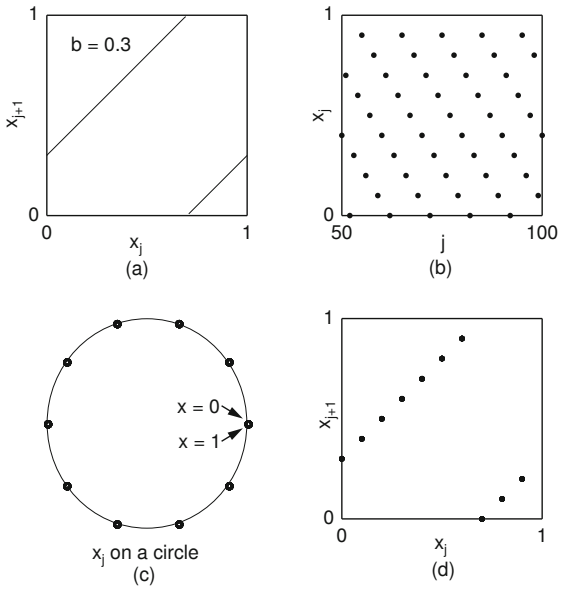


Fig. 10.29 The linear map $x_{j+1} = x_j + 0.3 \pmod 1$. **a** Plot of the map. **b** Plot of x_j vs. j for 50 points. **c** The map plotted on a *circle* for 128 values of j, which lie on only 10 points. **d** A plot of 128 values of x_{j+1} vs. x_j falls on 10 points on the map

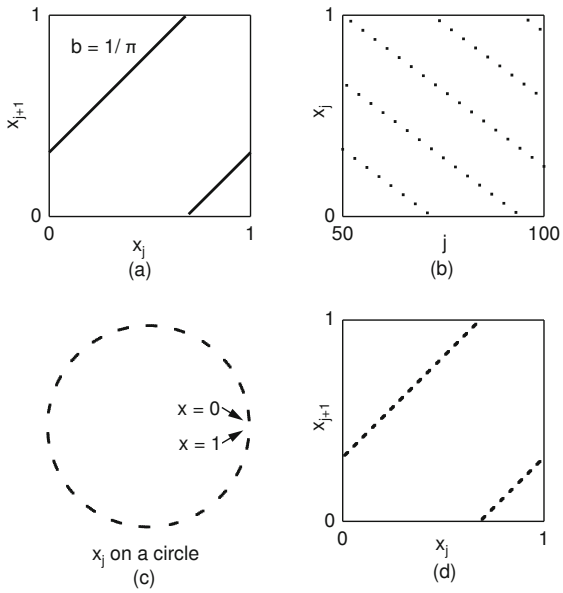


Fig. 10.30 The linear map $x_{j+1} = x_j + 1/\pi \pmod 1$. **a** Plot of the map. **b** Plot of x_j vs j for 50 points. **c** The map plotted on a *circle* for 128 points. **d** Plot of x_{j+1} vs. x_j gives 128 points on the map

The function is plotted in Fig. 10.29 for $b = 0.3$. The map is plotted in (a). The apparent discontinuities are due to the wrapping. A sequence of 50 points is plotted in Fig. 10.29b. Because $b = 3/10$ is a rational fraction, the points repeat themselves exactly every 10 steps. This can be seen in Fig. 10.29c, which plots 128 consecutive points on a circle. The angle counterclockwise from the horizontal axis is $\theta_j = 2\pi x_j$. The 128 values all fall at 10 points on the circle. The plot of x_{j+1} vs. x_j in Fig. 10.29d has 10 points that fall on the map.

Compare this with Fig. 10.30, which is a plot of the same map for an irrational value of the parameter, $b = 1/\pi$. The curve in Fig. 10.30a looks very similar. However, the values of x_j never repeat. This is difficult to see from Fig. 10.30b, but can be seen in c, where the 128 points are all at different values of θ. If more points were plotted, the circle would be completely filled. All of the points plotted in d are also different, but of course they lie on the map function. If we were to make a bifurcation diagram the values of x_j would fill all points on the graph, unless b were a rational fraction, when there would be a finite number of points. This appears at first sight to be chaotic behavior, but it is not. The function is deterministic, it is bounded, and the values of x never repeat. But it does not satisfy the last criterion: sensitive dependence on initial conditions. In chaotic behavior, two trajectories that start from initial points that are very close diverge in time. If a slightly different value of x_0 is used for this map, all of the values in the new sequence are shifted from the original sequence by the same amount. There is no divergence of the two solutions. In quasiperiodicity, the trajectories for two points that are initially close remain close.

10.10 A Feedback Loop with a Time Constant and a Fixed Delay

In Sect. 10.6 we saw that if both processes in a two-stage feedback system had comparable time constants, there was the possibility for damped oscillations or "ringing." Another possibility is that a portion of the system may respond to values of a state variable at some earlier time. The fixed time delay could be the time it takes a signal to travel along a nerve or the time it takes for a chemical to pass through a blood vessel.

We will consider a linear model for such a system, as shown in Fig. 10.31:

$$\begin{aligned} \tau_1 \frac{dy}{dt} + y &= G_1 x + p_1, \\ x &= G_2 y(t - t_d) + p_2. \end{aligned} \tag{10.42}$$

The first equation is like those in Sect. 10.6, except that the factor a multiplying p_1 is set equal to unity. The second equation says that $x(t)$ is proportional to the value of y at

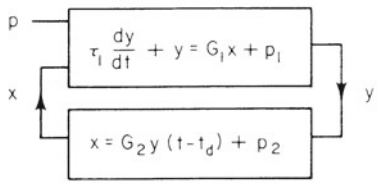


Fig. 10.31 A two-stage feedback loop. The upper process is described by a single time constant; the lower one introduces a fixed time delay

the earlier time $t - t_d$, plus some other parameter p_2. These can be combined to give a *delay-differential equation*:

$$\tau_1 \frac{dy}{dt} = -y + G_1 G_2 y(t - t_d) + p_1 + G_1 p_2$$

or, defining $p = p_1 + G_1 p_2$ to eliminate clutter,

$$\tau_1 \frac{dy}{dt} = -y + G_1 G_2 y(t - t_d) + p. \tag{10.43}$$

This equation can give rise to sustained as well as damped oscillations. It is not hard to see why. Suppose that y is above some equilibrium value and that $G_1 G_2 < 0$. The first term on the right causes y to decrease toward equilibrium. But when it is nearly at equilibrium the second term, responding to an earlier positive value of y, continues to make y decrease so y goes negative. Now y is below the equilibrium value and the same arguments can be applied as y increases. This paragraph could go on for a long time.

Why do we now have oscillations for a system with apparently only one degree of freedom? The reason is the delay term. In order to specify the initial state of the system at $t = 0$, we must specify the value of y for all times $-t_d < t < 0$. This is effectively an infinite number of values of y. Delay differential equations have an infinite number of degrees of freedom.

The mathematics for such a system become quite involved (even for the linear system we discuss here). The techniques for solving the equation were first described by Hayes (1950). The equation has been considered for biological examples by Glass and Mackey (1988).

The derivative is zero and the equation has a fixed point y_f when $y_f = p/(1 - G_1 G_2)$. It is convenient to work with the new variable $w = y - y_f$ and rewrite Eq. 10.43 as

$$\tau_1 \frac{dw}{dt} = -w + G_1 G_2 w(t - t_d).$$

We make another simplifying assumption; that the magnitude of the open-loop gain $G_1 G_2$ is so much greater than 1 that the $-w$ term can be neglected.[10] Then the equation becomes

$$\frac{dw}{dt} = \frac{G_1 G_2}{\tau_1} w(t - t_d).$$

Now recall that since $G_1 G_2 \ll -1$, this coefficient is approximately the negative of the reciprocal of the time constant with no delay and *with* feedback (see Eq. 10.23). Therefore the equation we will solve is

$$\frac{dw}{dt} = -\frac{1}{\tau} w(t - t_d). \tag{10.44}$$

If the delay time is zero, this is the familiar equation for exponential decay. As we argued above, a delay can allow oscillation. One can show by substitution that for certain values of the parameters one possible solution has the form $w(t) = w_0 e^{-\gamma t} \cos \omega t$. We will find the conditions for a steady oscillation of the form $w(t) = w_0 \cos \omega t$. The left-hand side of Eq. 10.44 is $dw/dt = -\omega w_0 \sin \omega t$. The right-hand side is

$$-(1/\tau) w_0 \cos(\omega t - \omega t_d) = -(1/\tau) w_0 \cos \omega t \cos \omega t_d - (1/\tau) w_0 \sin \omega t \sin \omega t_d.$$

Therefore the proposed solution will satisfy Eq. 10.44 only if

$$-\omega w_0 \sin \omega t = -(1/\tau) w_0 \sin \omega t_d \sin \omega t$$

and

$$0 = -(1/\tau) w_0 \cos \omega t_d \cos \omega t,$$

from which we get $\omega = 1/\tau$ and $\cos \omega t_d = 0$ or $\omega t_d = \pi/2$. Combining these gives $t_d/\tau = \pi/2$. From these we see exactly how the sustained oscillation occurs. The delay time and frequency are such that the shift is exactly one-quarter cycle. This is the same shift that would be obtained by taking the second time derivative of the undelayed function, which would lead to the undamped harmonic oscillator equation.

10.11 Negative Feedback Loops: A Summary

The last several sections have been mathematically complex. However, you do not need to memorize a large number of equations to carry away the heart of what is in them. The essential features are as follows:

1. If the equations relating the input and output variables of each process of a negative feedback loop are known, then their simultaneous solution gives the equilibrium or steady-state values of the variables. (In a biological system it may be very difficult to get these equations.) The

[10] If you are considering a problem where this is not a reasonable assumption, see the Appendix of Glass and Mackey (1988).

solution is called the operating point or a fixed point of the system of equations.
2. If a single process in the negative feedback loop determines the time behavior, and the rate of return of a variable to equilibrium is proportional to the distance of that variable from equilibrium, then the return to equilibrium is an exponential decay and the system can be characterized by a time constant.
3. In a negative feedback system one variable changes to stabilize another variable. The amount of stabilization and the accompanying decrease in time constant depend on the open-loop gain.
4. It is possible to have oscillatory behavior with damped or constant amplitude if the two processes have comparable time constants and sufficient open-loop gain, or if one of the processes depends on the value of its input variable at an earlier time, or if the process has three or more degrees of freedom.
5. A nonlinear system oscillating on a limit cycle can have its phase reset by an external stimulus.
6. Nonlinear systems of difference equations with one or more degrees of freedom or nonlinear systems of differential equations with three or more degrees of freedom may exhibit bifurcations and chaotic behavior.

10.12 Additional Examples

This section provides some additional examples of the principles we have seen above. The details of the experiments and modeling are given in the references.

10.12.1 Cheyne–Stokes Respiration

We have seen how the body responds to CO_2 levels in the blood by controlling the rate and amplitude of breathing to maintain the CO_2 concentration within a narrow range. The frequency and amplitude of breathing can also undergo oscillation. Some patients almost stop breathing for a minute or so and then breathe with much greater amplitude than normal. This is called *Cheyne–Stokes breathing*. Guyton et al. (1956) showed that diverting carotid artery blood in dogs through a long length of tubing increased the transit time between heart and brain and caused Cheyne–Stokes respirations. Cheyne–Stokes respirations have been modeled with a nonlinear delay-differential equation by Mackey and Glass (1977). Their results are shown in Fig. 10.32.

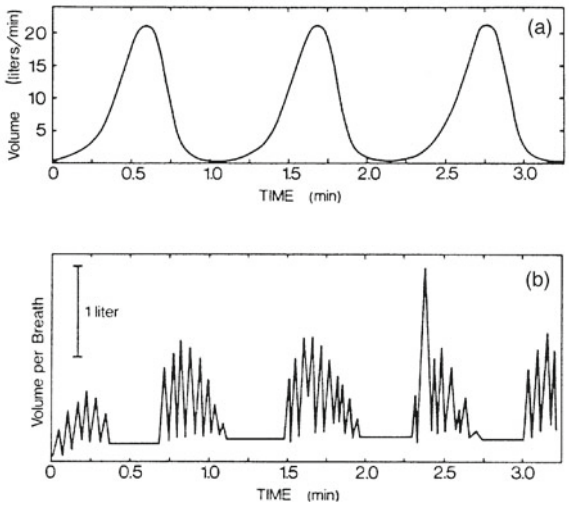


Fig. 10.32 Cheyne–Stokes respirations. **a** The results of the model calculation $y = g(x(t - t_d))$. **b** Ventilation during Cheyne–Stokes respiration. (Reprinted with permission from Mackey and Glass 1977. Copyright 1977 AAAS)

10.12.2 Hot Tubs and Heat Stroke

Problems 10.10 and 10.11 discuss how the body perspires in order to prevent increases in body temperature. At the same time blood flows through vessels near the surface of the skin, giving the flushed appearance of an overheated person. The cooling comes from the evaporation of the perspiration from the skin. If the perspiration cannot evaporate or is wiped off, the feedback loop is broken and the cooling does not occur. If a subject in a hot tub overheats, the same blood flow pattern and perspiration occur, but now heat flows into the body from the hot water in the tub. The feedback has become positive instead of negative, and heat stroke and possibly death occurs. This has been described in the physics literature by Bartlett and Braun (1983).

10.12.3 Pupil Size

The pupil changes diameter in response to the amount of light entering the eye. This is one of the most easily studied feedback systems in the body, because it is possible to break the loop and to change the gain of the system. Let the variables be as follows: x is the amount of light striking the retina, p is the light intensity, and y the pupil area. In the normal case, x is proportional to y and p: $x = Apy$. The body responds to increasing x by decreasing y so $y = f(x)$. These processes are shown in Fig. 10.33.

The reason this system can be studied so easily is that shining a very narrow beam of light into the pupil means that the change of pupil radius no longer affects x; the loop is

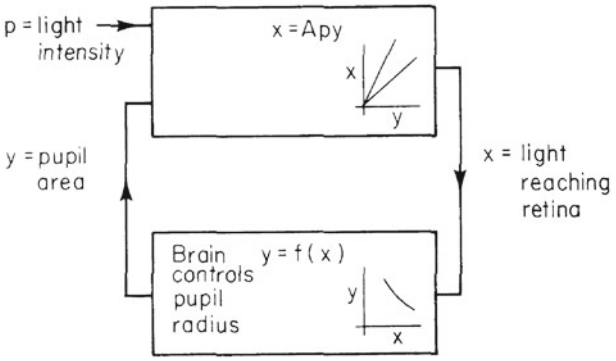


Fig. 10.33 The feedback system for controlling the size of the pupil

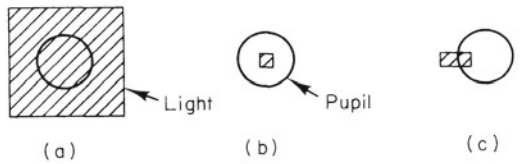


Fig. 10.34 The feedback loop for pupil size can be changed by changing the way in which light strikes the eye. **a** In the normal situation $x = Apy$. **b** When the spot of light is smaller than the pupil, the feedback loop is broken. **c** When the spot of light strikes the edge of the pupil, the gain is increased

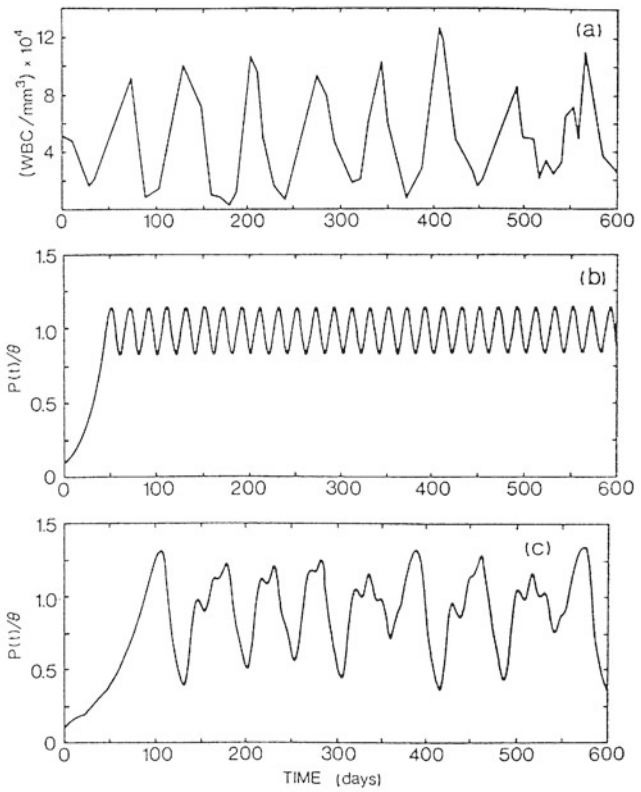


Fig. 10.35 A nonlinear model for white-blood-cell production. **a** White-blood-cell count from a patient with chronic granulocytic leukemia. **b** The results of a nonlinear delay-differential equation model with a delay time of 6 days are an oscillation with a period of 20 days. **c** The results of using the same model with a delay time of 20 days are aperiodic. (Reprinted with permission from Mackey and Glass 1977. Copyright 1977 AAAS)

broken in the upper box of Fig. 10.33. Shining a light into the eye so that it is on the edge of the pupil increases the gain in the upper box. These schemes are shown in Fig. 10.34. Furthermore, it has been discovered experimentally that the process in the lower box controls the size of both pupils, even though light is directed at only one eye.

The properties of $y = f(x)$ have been studied extensively by Stark (see Stark 1957, 1968, 1984). The results are consistent with a feedback loop having several time constants and also a fixed delay. Increasing the open-loop gain as in Fig. 10.34c causes the pupil to oscillate at a frequency of about 1.3 Hz (cycles per second). Stark (1984) reviews this work, including the use of noise to analyze the system and nonlinearities.

10.12.4 Oscillating White-Blood-Cell Counts

A delay-differential equation has been used to model the production of red and white blood cells. Figure 10.35 shows the actual white count for a patient with chronic granulocytic leukemia as well as the results of a model calculation. The striking feature of the model is the emergence of an aperiodic pattern when the delay time is increased from 6 to 20 days.

10.12.5 Waves in Excitable Media

The propagation of an action potential is one example of the propagation of a wave in excitable media. We saw in Chap. 7 that waves of depolarization sweep through cardiac tissue. The circulation of a wave of contraction in a ring of cardiac tissue was demonstrated by Mines in 1914. It was first thought that such a wave had to circulate around an anatomic obstacle, but it is now recognized that no obstacle is needed.

Waves in thin slices of cardiac tissue often have the shape of spirals, very similar to simulations produced by a model similar to a two-dimensional Hodgkin–Huxley model (Gray 2009). These waves occur in many contexts beside the heart. They have also been seen in the Belousov–Zhabotinsky

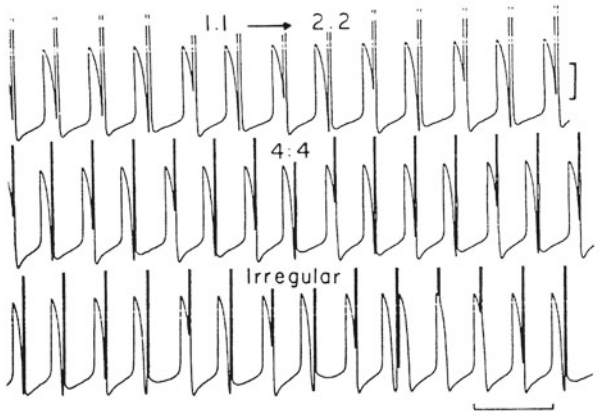


Fig. 10.36 An aggregate of chick heart cells was periodically stimulated. Follow the bottom of the beginning of each sharp spike. The left part of the top strip shows phase locking. The right-hand portion of the top strip shows period doubling. The middle strip shows a period-4 behavior. The bottom strip shows irregular behavior consistent with deterministic chaos. (Reprinted with permission from Guevara et al. 1981. Copyright 1981 AAAS)

chemical reaction,[11] in social amoebae, in the retina of the eye, and as calcium waves in oocytes. Beautiful photographs of all of these are found in Winfree (1987). A simple reaction–diffusion model that leads to a propagating chemical wave is found in Problem 4.28

These *spiral waves* seem to be another ubiquitous phenomenon (like period doubling) that depends primarily on the coarse features of the model. They can be generated with simple computer models called *cellular automata*. The rules for such an automaton and photographs of the resulting spiral waves are shown in Chap. 2 of Kaplan and Glass (1995).

The study of spiral waves in the heart is currently an active field (Glass et al. 2002; Keener and Sneyd 2008a; Panfilov 2009; Gray 2009; Luther et al. 2011; Clayton et al. (2011)). They can lead to ventricular tachycardia, they can meander, much as a tornado does, and their breakup into a pattern resembling turbulence is a possible mechanism for the development of ventricular fibrillation (see the next example).

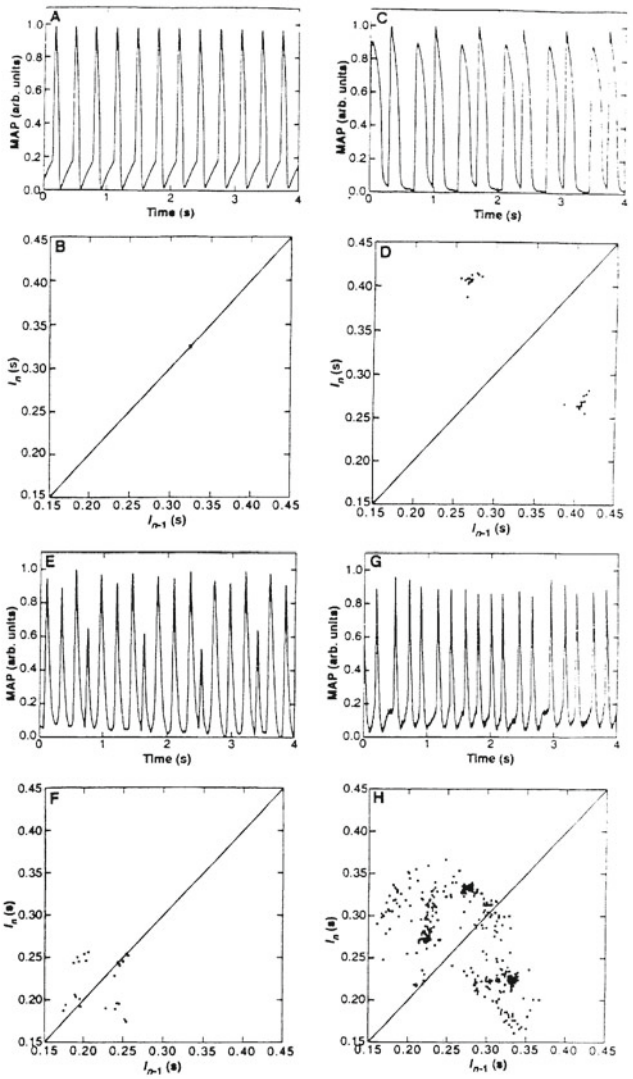


Fig. 10.37 The results of experiments on a preparation consisting of intraventricular septum from a rabbit heart. Plots show the recorded action potentials and the map of I_n vs. I_{n-1} where I is the interval between beats. In *A* and *B* there is a constant interbeat interval and one point on the map. Panels *C* and *D* show period doubling. Panels *E* and *F* show a period-4 pattern. Panels *G* and *H* are completely aperiodic. (Reprinted with permission from Garfinkel et al. 1992. Copyright 1992 AAAS)

10.12.6 Period Doubling and Chaos in Heart Cells

Guevara et al. (1981) have subjected small aggregates of chick heart cells to periodic stimulation. The stimulation frequency was slightly greater than the natural frequency of oscillation. The behavior of the preparation is shown in Fig. 10.36 and can best be seen by examining the *bottom* of the leading edge of the sharp positive pulse. The top strip on the left is phase locking. This is followed on the right in the top strip by an alternation characteristic of period doubling. The middle strip shows a variation of period 4. The bottom strip shows irregularity that is consistent with deterministic chaos.

Garfinkel et al. (1992) have also observed period doubling in a stimulated preparation of rabbit heart. Arrhythmias were induced by adding drugs to the solution perfusing the preparation. Figure 10.37 shows plots of the recorded action potentials and a plot of the map of I_n vs. I_{n-1}, where I is the interval between beats. In panels A and B there is a constant

[11] There are many references. See Mielczarek et al. (1983); Epstein et al. (1983); and Winfree (1987).

interbeat interval and one point on the map. Panels C and D show period doubling. Panels E and F show a period-4 pattern. Panels G and H are completely aperiodic.

Ventricular fibrillation is "the rapid, disorganized, and asynchronous contraction of ventricular muscle. ... it represents the final common pathway for death in most patients who experience out-of-hospital cardiac arrest, and its rate of recurrence is on the order of 30 % in the first year in successfully resuscitated patients." (Epstein and Ideker 1995). It appears to be due to meandering waves, and it does not occur unless the heart exceeds some minimum size (Winfree 2001).

Witkowski et al. (1993, 1994) have made electrode arrays with a spacing of about 200 μm that can be placed directly on the myocardium. The membrane current i_m can be estimated from the spatial derivatives of the extracellular (interstitial) potential. This technique has provided evidence that ventricular fibrillation has a component with simpler dynamics than had previously been thought (Witkowski et al. 1995). More recent studies point toward more complex behavior (Fox et al. 2002).

Symbols Used in Chapter 10

Symbol	Use	Units	First used page
a, b	Arbitrary parameter		277
a	Parameter in logistic map		284
a, b	Constant in logistic equation		281
b	Reduction in ventilation rate because of dead space in lungs	l min^{-1}	271
b	Amplitude of stimulus		282
f, g, h, i	Functions		271
j	Index for successive values in difference equation		284
m	Mass	kg	280
n	Number of moles of dissolved carbon dioxide		273
o	Rate of oxygen consumption	mol s^{-1}	270
p	Rate of oxygen consumption	mmol min^{-1}	270
p	Light intensity		289
r	Variable		281
t	Time	s	273
t_d	Delay time	s	288
v	Velocity	m s^{-1}	280
w, x, y, z	General variables		271
x, y	General variables in a feedback system		269
x	Partial pressure of carbon dioxide	torr	270
x	Amount of light striking the retina		289
x^*	Equilibrium value of x		284
y	Ventilation rate	l min^{-1}	270
y	Pupil area	m^2	289
y_∞	Constant (carrying capacity) in logistic equation		281
A	Proportionality constant		273
F	Respiratory quotient		270
F	General function		271
F_x	x component of force	N	280
G_1, G_2	Gain		272
I	Interbeat interval	s	291
N	Number of variables		280
P_{CO_2}	Partial pressure of carbon dioxide	torr	270
R	Gas constant	J K^{-1} mol^{-1}	270
T	Temperature	K	270
T, T_0, T_s	Time	s	282
V_c	Compartment in which carbon dioxide is distributed throughout the body	m^3 or l	273
α	Solubility constant	mol l^{-1} torr^{-1}	273
α	Damping constant	s^{-1}	277
$\theta, \theta', \theta_s$	Angle		281
τ, τ_1, τ_2	Time constant	s	273
ω, ω_0	Angular frequency	radian s^{-1}	277
ξ, η	Variables		273

Problems

Section 10.1

Problem 1. Make the unit conversions to show that Eq. 10.4 is equivalent to Eq. 10.1.

Section 10.2

Problem 2. The level of the thyroid hormone thyroxine (T4) in the blood is regulated by a feedback system. TSH is released by the pituitary. The thyroid responds to increased levels of TSH by producing more T4. The T4 then acts through the hypothalamus and pituitary to reduce the amount of TSH.

(a) On a graph of T4 vs. TSH, plot hypothetical curves showing these two processes and indicate the equilibrium or operating point.

(b) T4 contains four iodine atoms. If the body has an insufficient supply of dietary iodine, the thyroid cannot make enough T4. What changes in the graphs will result? (This causes iodine deficiency goiter or thyroid hyperplasia. With the advent of iodized table salt and the use of iodine by bakers in bread dough to make their equipment easier to clean, the disease has almost disappeared.)

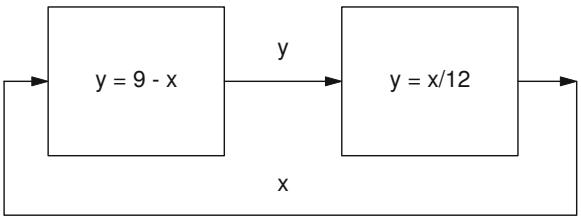
Problem 3. For the feedback system $x = \left[(y-p)/3\right]^{1/2}$, $y = 4 - x^2$ assume that the variable on the right in each equation controls the variable on the left.

(a) Plot y vs. x for each process.
(b) Find the operating point when $p = 0$.
(c) Find the operating point when $p = 1$.

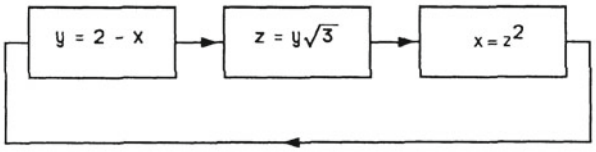
Section 10.3

Problem 4. Find the open-loop gain for the system described in Problem 10.3.

Problem 5. Find the open-loop gain for the system shown.



Problem 6. A feedback loop has the three stages shown. Find the operating point and the open-loop gain if these variables are all positive.



Problem 7. Consider how thyroid hormone is removed from the body by the kidneys. The variables are V, the total plasma volume (l); C, the plasma concentration of thyroid hormone (mol l^{-1}); y, the total amount of hormone (mol); and R, the rate of hormone production (mol s^{-1}). The rate of change is $dy/dt = R - KC$. In the steady state $R = KC$, and y is not changing with time (see Chap. 2). The clearance K is a measure of the kidneys' ability to remove hormone, since the removal process depends on the concentration.

(a) Plot K vs. C for two different values of R. Show on your graph what happens if K remains fixed as R changes.
(b) It has been found experimentally (Riggs 1952. *Pharmacol Rev* **4**: 284–370) that K increases as C increases: $K = aC$. Plot this on your graph, too.
(c) Draw a block diagram showing the proper cause and effect relationship between C and K.
(d) Calculate the open-loop gain. Show how changes in C are altered by the feedback mechanism.

Problem 8. A substance is produced in the body and removed at rate R. The concentration is C. The clearance is defined to be K. In the steady state $0 = dy/dt = R - KC$, or $K = R/C$. It is found experimentally that the clearance depends on the concentration as $K = aC^n$, where C is the independent variable. Find the open-loop gain, eliminating K and a from your answer.

Problem 9. The kidney excretes phosphate in the following way. The total plasma volume V_p contains phosphate at concentration C_p: $Q_p = C_p V_p$. A volume of plasma $(dV/dt)_f$ is filtered through the renal glomeruli into the nephrons each second. Within the nephron, phosphate is either reabsorbed into the plasma or excreted into the urine. Experiments show that virtually all phosphate is reabsorbed up to some rate $(dQ/dt)_{\max}$:

$$\left(\frac{dQ}{dt}\right)_{\text{reabs}} = \begin{cases} C_p\,(dV/dt)_f\,, & C_p\,(dV/dt)_f < (dQ/dt)_{\max} \\ (dQ/dt)_{\max}\,, & C_p\,(dV/dt)_f \geq (dQ/dt)_{\max}\,. \end{cases}$$

As in Problem 7a, at equilibrium the clearance of phosphate from the plasma is defined as

$$K = \frac{(dQ/dt)_{\text{excreted into urine}}}{C_p}.$$

Suppose that exogenous phosphate is entering the plasma at a fixed rate R and that steady state has been reached so that $R = (dQ/dt)_{\text{excreted into urine}}$.

(a) What value for reabsorption does this imply?
(b) Determine two equations relating K and C_p and plot them.
(c) Calculate the open-loop gain of the feedback loop.

Problem 10. With considerable simplification, consider the body to have a constant temperature T throughout and a total heat capacity C. The total amount of thermal energy in the body is U. The heat capacity is defined so that $dU = CdT$. The source of the thermal energy is the body's metabolism: $(dU/dt)_{\text{in}} = M$. If sweating is ignored, the rate of loss of energy by convection and radiation is approximately proportional to the amount by which the body temperature exceeds the ambient or surrounding temperature: $(dU/dt)_{\text{loss}} = K(T - T_a)$.

(a) What is the steady-state temperature as a function of M and T_a?
(b) Write a differential equation for T as a function of time. Suppose that M suddenly jumps by a fixed amount. What is the time constant?

Problem 11. When the body temperature is above 37 °C, sweating becomes important. The rate of energy loss is proportional to the amount of water evaporated. If all the perspiration evaporates, sweating loss can be approximated by $(dU/dt)_{\text{sweat}} = L(T - 37)$.

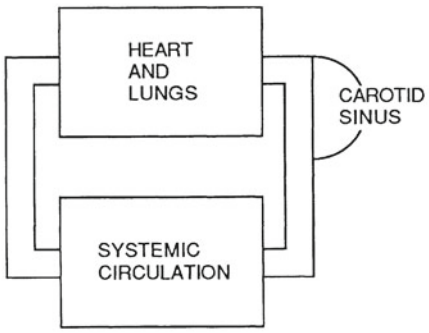
(a) Modify the differential equation of the previous problem to include $(dU/dt)_{\text{sweat}}$ as the input variable with T as the output variable. Combine it with this new equation

to make a feedback loop. Determine the new equilibrium temperature and the time constant.

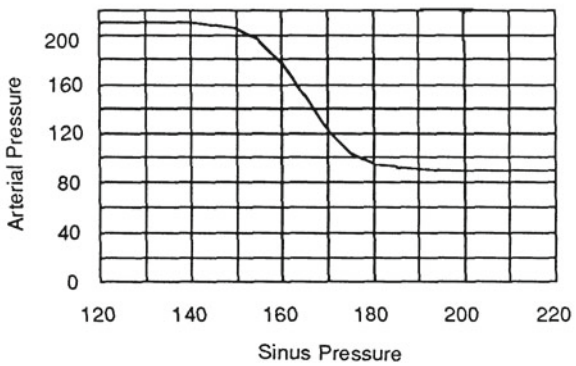
(b) Make numerical comparisons for the previous problem and this one when $M = 71$ kcal h^{-1}, $C = 70$ kcal °C^{-1}, $K = 25$ kcal h^{-1} °C^{-1}, $L = 750$ kcal h^{-1} °C^{-1}, $T_a = 38$ °C (high enough to ensure sweating).

Problem 12. A simplified model of the circulation is shown. Normally the arterial pressure is the same as that in the carotid sinus: $p_{\text{art}} = p_{\text{sinus}}$. In experiments on dogs whose vagus nerves were cut, the carotid arteries were isolated and perfused by a separate pump. This broke the feedback loop and allowed the curve on the accompanying graph to be obtained. The empirical equation shown (based on the work of Scher and Young (1963), summarized in Riggs (1970)) is (with pressures in torr)

$$p_{\text{art}} = 90 + \frac{120}{1 + \exp\left[(p_{\text{sinus}} - 165)/5\right]}.$$



(a)



(b)

(a) Draw a block diagram of the complete feedback system. Label the blocks, show the functional relationship for each one, and indicate the proper cause-and-effect relationship.
(b) Find the operating point.
(c) Find the open-loop gain.

Problem 13. Consider the following special case of linear feedback: $\Delta x = G_1(\Delta p + \Delta y)$, $\Delta y = G_2 \Delta x$. Find the ratio $\Delta x/\Delta p$ when $G_1 \ll -1$, $G_2 < 1$.

Problem 14. Differentiate Eq. 10.4 and show that the expression for G_1 is the same as in Eq. 10.19.

Problem 15. For the thyroid problem, Problem 10.7, write a differential equation that can be solved to give C as a function of time. Suppose that at $t = 0$, R suddenly becomes 0. What is the differential equation then? Solve the equation; note that it is not linear.

Problem 16. The osmolarity of plasma (C, in mosmole) is regulated by the concentration of *antidiuretic hormone* (ADH, in pg ml^{-1}, also known as *vasopressin*). As antidiuretic hormone increases, the kidney reabsorbs more water and the plasma osmolarity decreases, $C = 700/ADH$. When osmoreceptors in the hypothalamus detect an increase of plasma osmolarity, they stimulate the pituitary gland to produce more antidiuretic hormone, $ADH = C - 280$ for C greater than 280, and zero otherwise.

(a) Draw a block diagram of the feedback loop, including accurate plots of the two relationships.
(b) Calculate the operating point and the open-loop gain (you may need to use four to six significant figures to determine the operating point accurately).
(c) Suppose the behavior of the kidney changed so now $C = 750/ADH$. First determine the new value of C if the regulation of ADH is not functioning (ADH is equal to that found in part (b), and then determine the value of C taking regulation of ADH by the hypothalamus into account.

Problem 17. The concentration of potassium ions in plasma K (in mM) is regulated by the concentration of the hormone aldosterone A (in ng per 100 ml). As aldosterone increases, the kidney excretes more potassium in the urine and the plasma concentration of potassium ions decreases: $K = 17.4/A$. When the extracellular potassium ion concentration rises, the adrenal gland produces more aldosterone: $A = (1/200)\exp(1.6K)$.

(a) Draw a block diagram of the feedback loop, including accurate plots of the two relationships.
(b) Calculate the operating point and the open-loop gain. (Note: you may have to find these numerically or graphically.)
(c) Suppose the behavior of the kidney changed so that $K = 20/A$. Determine the new value of K if the regulation of aldosterone is not functioning (A is equal to that found in part (b)), and then determine the value of K taking regulation of aldosterone into account.

Section 10.5

Problem 18. The following is a vastly oversimplified model of calcium regulation in dogs. Calcium is stored in body fluids and bones. Experiments show that the calcium concentration in the blood of a dog obeys approximately the equation (Riggs 1970, p. 491)

$$3.9\frac{dC}{dt} + 1.4C = 81.2 + \left(\frac{dQ}{dt}\right)_{\text{iv}} + \left(\frac{dQ}{dt}\right)_r (t),$$

where C is the plasma concentration in mg l^{-1}, t is the time in h, $(dQ/dt)_{\text{iv}}$ is the rate of intravenous infusion of calcium in mg h^{-1}, and $(dQ/dt)_r$ is the rate of reabsorption of calcium from bone into the blood in mg h^{-1}. (The numerical constants are consistent with these units.) The rate of reabsorption depends on the level of parathyroid hormone (PTH) concentration in the blood, which in turn depends on the calcium concentration. Instead of measuring the PTH concentration, experimenters found that $(dQ/dt)_r$ and C are related empirically by $(dQ/dt)_r = 188 - 1.34C$, where C is the independent variable.

(a) Draw a block diagram with variables $(dQ/dt)_r$ and C.
(b) Write equations to describe the steady state and find steady-state values of $(dQ/dt)_r$ and C when $(dQ/dt)_{\text{iv}} = 0$.
(c) Find the open-loop gain.
(d) Find the time constant for the change of C when the parathyroid glands have been removed, in response to a step change in $(dQ/dt)_{\text{iv}}$.
(e) Find the time constant for the change in C in response to a step change in calcium infusion when the parathyroid glands are intact, so that the feedback loop is closed.

Problem 19. This problem is a simplification by the authors suggested by the data of Chick et al. (1977) . Experimental data on diabetic rats show that the insulin level is 0 and the glucose level is 500. When an artificial pancreas is installed, a new operating point is reached for which $i = 40$ and $g = 100$.

(a) Make the simplest assumption possible: glucose level responds to insulin level according to $g = A + G_1 i$, while insulin responds to glucose as $i = G_2 g$. Find the open-loop gain.
(b) The same series of experiments showed that when the feedback loop is closed, the time constant for glucose to fall is 1.67 h. When the artificial pancreas is removed, the glucose level rises with a time constant of 10.67 h. Estimate the open-loop gain, assuming that the insulin level changes instantaneously.

Section 10.6

Problem 20. Multiply Eq. 10.28 by τ_2 and show that it reduces to Eq. 10.22 when $\tau_2 \ll \tau_1$.

Problem 21. For the two-stage feedback loop with equal time constants τ, show that oscillation results with a frequency $\omega = (|OLG|)^{1/2}/\tau$.

Problem 22. Consider two substances in the plasma with concentrations X and Y. (They might be glucose and insulin.) Assume that experiment has established the following facts.

(i) The steady-state values of each concentration are X_0 and Y_0. Departures from them are $x = X - X_0$ and $y = Y - Y_0$.
(ii) When $y = 0$, X is removed from the body at a rate proportional to x. This is true for both positive and negative values of x: $dx/dt = -(1/\tau_1)x$.
(iii) When $x = 0$, Y influences the rate at which X changes in an approximately linear fashion. An increase of Y above Y_0 ($y > 0$) increases the rate of disappearance of x.
(iv) When $x = 0$, y is cleared at a rate proportional to y: $dy/dt = -(1/\tau_2)y$.
(v) When $y = 0$ and x is nonzero, a positive value of x stimulates the production of Y, while a negative value of x inhibits the production of Y.

Assume that the rate of production is a linear function of x. Write down two linear differential equations to model these observations. That is, add a term to each of the equations given that describes observations (iii) and (v).

Problem 23. Combine the two equations obtained in the previous problem into a single differential equation in x. Show that it has the form

$$\frac{d^2x}{dt^2} + \left(\frac{1}{\tau_1} + \frac{1}{\tau_2}\right)\frac{dx}{dt} + \frac{1 - \text{OLG}}{\tau_1 \tau_2}x = 0.$$

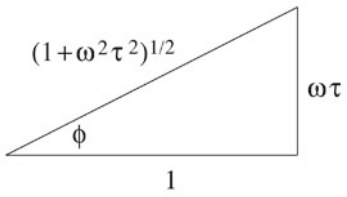
Use the result of Problem 19 to obtain $1 - \text{OLG}$ and suppose that $\tau_1 = 50$ min. For what value of τ_2 will critical damping occur? (If you find two values of τ_2, which seems more reasonable?) If τ_2 is greater than the value you select, will the system be overdamped or underdamped? (Do not take these results too seriously.)

Problem 24. This problem explores the behavior of a simple linear system from the point of view of the system's response to sinusoidal signals of various frequencies.

(a) The differential equation describing a system with time constant τ and gain G is

$$\frac{dx}{dt} = -\frac{1}{\tau}x + \frac{G}{\tau}y.$$

Show by substitution that if $y = Y \sin \omega t$, then $x = X \sin(\omega t + \phi)$, where $\tan\phi = \omega\tau$ and $X(\omega\tau \sin\phi + \cos\phi) = GY$.



(b) Use the relation $\tan\phi = \omega\tau$ to establish the triangle shown, and use it to show that $X = GY/(1+\omega^2\tau^2)^{1/2}$. These two relations give the response of the system in the frequency domain.

Problem 25. The following model for the attrition of troops in battle was developed by F. W. Lanchester and has been found to work reasonably well in several battles. The number of "friendly" troops is $F(t)$ and the number of "enemy" troops is $E(t)$. The rates of change are given by $dF/dt = -aE$, $dE/dt = -bF$, where a and b are the "effectiveness" of each side. The initial number of troops on each side is F_0 and E_0.

(a) What are the initial values of dF/dt and dE/dt?
(b) Obtain a differential equation for F.
(c) Find the most general solution to this differential equation and determine the coefficients from the initial conditions.
(d) Plot F and E for $a = b = 0.05$ and $E_0 = 2F_0$.

Problem 26. The equation $dF/dt = -aE$ of Problem 25 could also be thought of as describing a predator–prey situation if a represents the number of animals that the enemy eats per unit time. Ignoring latent periods such as gestation and infancy, what is the simplest way the equation could be modified to take account of reproduction and other ways of dying?

Section 10.8

Problem 27. Make a phase-space plot and discuss stability for $dy/dt = by$, $dy/dt = -by$, and $dy/dt = a - by$.

Problem 28. Make a drawing similar to Fig. 10.16 for the differential equation $dx/dt = x(c - x^2)$ for different values of c (positive and negative) and describe the stability of the fixed points as a function of c.

Problem 29. (a) Make drawings of the tip of the vector that defines θ' in Fig. 10.19 to show that when $b < 1$, θ' takes on all values, while for $b > 1$, θ' is restricted to values near 0 and 2π.

(b) Redraw Fig. 10.20a in the case that the angles are not reset to zero when they reach 2π.

Problem 30. Consider the undamped harmonic oscillator in the form $dx/dt = v$, $dv/dt = -\omega_0^2 x$.

(a) Make a phase-plane plot.
(b) Is the closed trajectory a limit cycle? Why or why not?
(c) Add a damping force proportional to $-v$ and redraw the phase-plane plot.

Problem 31. In Fig. 10.20, the phase behavior changes dramatically between $b = 0.95$ and $b = 1.05$. This change is most apparent for $\theta_s = \pi$, where $\theta' = 0$ for $b = 0.95$ and $\theta' = \pi$ for $b = 1.05$. What happens for $\theta_s = \pi$ and $b = 1$ exactly?

Problem 32. Reproduce qualitatively plots like Fig. 10.20 for $b = -0.95$ and $b = -1.05$. This corresponds to a hyperpolarizing stimulus.

Problem 33. Write a simple computer program to solve the two differential equations in Eq. 10.37 for $r(t)$ and $\theta(t)$. (See Sect. 6.14 for some guidance on how to solve differential equations numerically).

(a) Make plots of $x(t) = r(t)\cos(\theta(t))$ as a function of time for different stimuli.
(b) Reproduce a few points in the plots of Fig. 10.20 using your program. In particular, examine stimuli given at or near $\theta_s = \pi$ with b approximately equal to 1.
(c) Try varying the parameter a, and see how it affects the solution.

Problem 34. Use the program written in Problem 33 to examine entrainment. Stimulate the radial isochron clock periodically, with a frequency near but not exactly equal to the natural frequency of oscillation. Find examples where the clock is entrained to the stimulus (the oscillation has the same frequency as the stimulus).

Problem 35. A simple model for excitation of cardiac tissue is the FitzHugh–Nagumo model

$$\frac{dv}{dt} = \frac{1}{\varepsilon}\left(v - \frac{v^3}{3} - u\right)$$
$$\frac{du}{dt} = \varepsilon\,(v + \beta - \gamma u)\,,$$

where $\varepsilon = 0.2$, $\gamma = 0.5$ and $\beta = 0.8$.

(a) Make a plot in phase space (v versus u) of the *nullclines* (the curves obtained when $dv/dt = 0$ and $du/dt = 0$).
(b) Determine the steady-state solution (fixed point). You may have to do this numerically or graphically.
(c) Write a simple program to solve these equations on the computer. (See Sect. 6.14 for some guidance on how to solve differential equations numerically.) Plot $v(t)$ and $u(t)$ for the initial conditions $v(0) = -0.70$, $u(0) = -0.65$.

Problem 36. Edward Lorenz (1963) published a simple, three-variable (x, y, z) model of Rayleigh–Bénard convection:

$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y$$
$$\frac{dz}{dt} = xy - \beta z$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$.

(a) Which terms are nonlinear?
(b) Find the three equilibrium points for this system of equations.
(c) Write a simple program to solve these equations on the computer (see Sect. 6.14 for some guidance on how to solve differential equations numerically). Calculate and plot $x(t)$ as a function of t for different initial conditions. Consider two initial equations that are very similar, and compute how the solutions diverge as time goes by.
(d) Plot $z(t)$ vs. $x(t)$, with t acting as a parameter of the curve.

Problem 37. Consider the difference equation

$$x_{n+1} = \begin{cases} ax_n, & 0 < x_n < 0.5 \\ a(1 - x_n), & 0.5 < x < 1 \end{cases}$$

(a) Plot x_{n+1} vs. x_n for the case $a = 3/2$, producing a figure analogous to Fig. 10.23.
(b) Find the range of values of a for which the solution for large n does not diverge to infinity or decay to zero. You can do this using either arguments based on plots such as that in part (a) or numerical examples.
(c) Find the equilibrium value x^* as a function of a, using a method similar to that in Eq. 10.40.
(d) Determine if this value is stable or unstable, based on the magnitude of the slope of the x_{n+1} vs. x_n curve.
(e) For $a = 3/2$, calculate the first 20 values of x_n using 0.250 and 0.251 as initial conditions. Be sure to carry your calculations out to at least five significant figures. Do the results appear to be chaotic? Are the results sensitive to the initial conditions?
(f) For one of the data sets generated in part (e), plot x_{n+1} vs x_n for 20 values of n to create a plot analogous to Fig. 10.28. Explain how you could use this plot to distinguish chaotic data from a random list of numbers between zero and one.

Section 10.9

Problem 38. Show that for the logistic difference equation, the slope dx_{n+1}/dx_n at x^* is given by $2-a$, so that for $a > 3$ the slope has magnitude > 1.

Use a spreadsheet to plot x_n for different values of a and explore the period doubling.

Plot x_{j+1} vs. x_j, and show why we restricted a to values less than four.

For the logistic map with $a = 3.9$, evaluate x_j using the two initial conditions $x_1 = 0.2000$ and $x_1 = 0.2001$. Carry out the calculation for at least 20 iterations.

Problem 39. Cyclic variations in the population of a species are often studied with a predator–prey model such as the Lotka–Voltera equations (Chap. 2 Problem 38). It is also possible to have cyclic variations of a single species. This problem explores one such model and is based on Appendix B of Ginzburg and Colyvan (2004).

Let N_t represent the population at generation t. Let X represent the quality of the resources available to that species. R is the maximum growth rate, and $f(X)$ is a monotonically increasing function that asymptotically approaches unity for large X. The population in the next generation is

$$N_{t+1} = N_t R f(X_t).$$

If f is constant, we have exponential growth or decay, depending on whether Rf is greater or less than 1. We will model f by $f(X_t) = X_t/(k + X_t)$, where parameter k determines how rapidly f approaches its asymptotic value.

Now assume the total amount of food, S, does not change with time and that X depends on the per capita food supply S/N through a monotonically increasing function g:

$$X_{t+1} = X_t g(S/N_{t+1}).$$

A crucial assumption is that the current quality depends on both the present per capita food supply and the quality in the previous generation. When there is more food available to the mother, it increases the reproductive rate of the mother. Ginzburg and Colyvan call this the *maternal effect*.

Model functions f and g by

$$N_{t+1} = N_t R \frac{X_t}{k + X_t}$$
$$X_{t+1} = X_t M \frac{S/N_{t+1}}{p + S/N_{t+1}}.$$

(a) Show that with the change of variables $n = pN/S$ and $x = X/k$ the equations reduce to

$$n_{t+1} = n_t R \frac{x_t}{1 + x_t}$$
$$x_{t+1} = x_t M \frac{1}{1 + n_{t+1}}.$$

(b) Use a spread sheet to model the behavior for a range of values of R and M, starting with $R = 20$ and $M = 10$. Use initial conditions $n_0 = x_0 = 1$. If $M > 1$, explore what values of R lead to oscillations.
(c) Use the spread sheet to construct phase-plane plots of $\ln(n)$ vs $\ln(x)$.

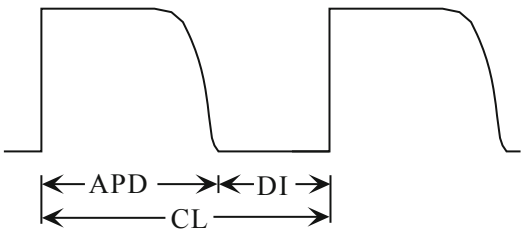
Problem 40. Consider the two sets of data below, one produced by the logistic map and the other produced from a table of random numbers. Which is which?

Set 1	Set 2
0.9750	0.7464
0.0951	0.2349
0.3356	0.6017
0.8696	0.0213
0.4422	0.7935
0.9620	0.0336
0.1426	0.6476
0.4768	0.5630
0.9729	0.9116
0.1028	0.1748
0.3597	0.8706
0.8982	0.9058

Problem 41. The onset of ventricular fibrillation in the heart can be understood in part as a property of *cardiac restitution*. The action potential duration (APD) depends on the previous diastolic interval (DI): the time from the end of the last action potential until the start of the next one. The relationship between APD and DI is called the *restitution curve*. In cardiac muscle, a typical restitution curve has the form

$$APD_{i+1} = 300\left(1 - e^{-DI_i/100}\right)$$

where all times are given in ms. Suppose we apply to the heart a series of stimuli, with period (or "cycle length") CL.



Since $APD + DI = CL$, we have $DI_{i+1} = CL - APD_{i+1}$.

(a) Suppose we stimulate with a long cycle length ($CL =$ 400 ms). Using an initial value of $DI_1 = 200$, calculate APD_i and DI_i for ten iterations. What happens?

(b) Shorten the cycle length to $CL = 300$ ms. Using the same DI_1, calculate APD_i and DI_i for ten iterations. What happens now? (In the jargon of cardiac electrophysiology, this behavior is often called *alternans*).

(c) Shorten the cycle length further to $CL = 200$ ms. Using the same DI_1, calculate APD_i and DI_i for ten iterations. If DI_{i+1} is negative (corresponding to tissue so refractory that it fires no action potential), keep adding CL to it until it becomes positive before calculating the next APD. What happens now?

(d) Shorten the cycle length further to $CL = 100$ ms. Using the same DI_1, calculate APD_i and DI_i for twenty iterations. What happens now?

Your results in part (d) should be chaotic, resembling ventricular fibrillation. We must not overinterpret this simple model, however, because fibrillation consists of propagating wave fronts, whereas this simple model does not include spatial effects. For a more detailed account of a model similar to this one, see Hastings et al. (2000).

Problem 42. In Problem 41, the onset of alternans occurs when the slope of the restitution curve $APD_{i+1} = 300\left(1 - e^{-DI_i/100}\right)$ becomes greater than 1.

(a) Calculate the slope of the restitution curve $d\,(APD)\,/d\,(DI)$ analytically.

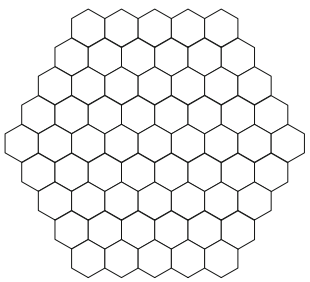
(b) Set the slope equal to 1 and solve for the resulting value of DI. Use the restitution curve to determine the corresponding values of APD and CL.

(c) Calculate APD_i and DI_i for twenty iterations for CL 10% above and 10 % below the value determined in part (b). What behaviors do you observe?

(d) Suppose you apply a drug to the heart that can change the restitution curve to $APD_{i+1} = 300\left(1 - be^{-DI_i/100}\right)$. Plot APD as a function of DI for $b = 0$, 0.5, and 1. What value of b ensures that the slope of the restitution curve is always less than 1? Garfinkel et al. (2000) have suggested that one way to prevent ventricular fibrillation is to use drugs that "flatten" the restitution curve.

Problem 43. Use the restitution curve from Problem 42 with $b = 1/3$ and $CL = 250$ to analyze the response of the system with initial diastolic intervals of 50, 60, 70, 80, and 90. You should find that the qualitative behavior depends on the initial condition. Which values of the initial diastolic interval give a $1:1$ response? Which give $2:1$? Determine the initial value of the DI to three significant figures for which the system makes a transition from one behavior to the other. When two qualitatively different behaviors can both occur, depending on the initial conditions, the system is *bistable*. To learn more about such behavior, see Yehia et al. (1999).

Problem 44. Elementary models of cellular excitable media (sometimes called *cellular automata*) provide valuable insight into the electrical behavior of the heart. Winfree (1987, pp. 106–107) describes one such model. A hexagonal array represents a sheet of cardiac tissue.



Each cell in the array can be in one of three states: excited (E), refractory (R), or quiescent (Q). A cell changes state by the following rules:

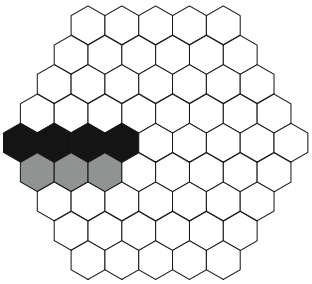
1. If in state E, then at the next time step it changes to state R,

2. If in state R, then at the next time step it changes to state Q,
3. If in state Q, then at the next time step it remains in state Q unless one of its six nearest neighbors is in state E, in which case it changes to state E.

(a) Start with the central cell in state E, and the rest of the cells in state Q. What happens in subsequent time steps? You should get an outwardly propagating wave front. Let the simulation run long enough to see what happens when the wave front hits the edge of the array. Does it "reflect" off the edge? Does the tissue ever go to the state of all Q?

(b) Start with the top five and bottom five cells in state E, and the rest in state Q. What happens when the two resulting wave fronts collide? Does the tissue ever go to the state of all Q?

(c) Start with the four black cells in state E, the three gray cells in state R, and the rest in state Q.



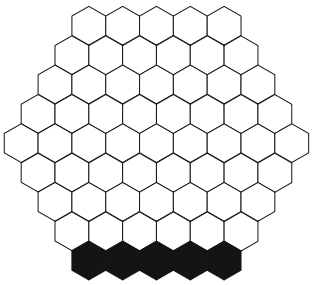
What happens? Does the tissue ever go to the state of all Q? This results in a spiral wave and may be responsible for some heart arrhythmias, such as ventricular tachycardia.

(d) In part (c), there is a special point called a *phase singularity* where cells in states E, R, and Q all meet at one point. Find the phase singularities in the results of part (c). How many are there? Do they move?

Problem 45. In the cellular excitable medium described in Problem 44, what happens if you apply an electrical stimulus? The stimulus is described by a fourth rule:

4. A stimulus changes the state to E, regardless of the previous state.

Assume the stimulus is applied only to the central cell. Start with the initial condition



which will initiate a wave front propagating upward. At a later time, apply the stimulus and see what happens. If the initial condition is $t = 1$, then try applying the stimulus at $t = 4$, 5, or 6. Are there any situations in which you produce phase singularities? If so, how many? Is the timing of the stimulus important?

Section 10.10

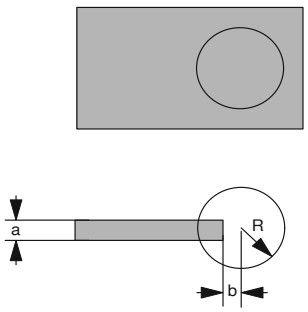
Problem 46. By substitution show that $w(t) = w_0 e^{-\gamma t} \cos \omega t$ can be a solution of the delay-differential equation, Eq. 10.44 if $\gamma = (1/\tau)\, e^{\gamma t_d} \cos \omega t_d$, $\omega = (1/\tau)\, e^{\gamma t_d} \sin \omega t_d$. Introduce the dimensionless variables $\alpha = t_d/\tau$, $\xi = \omega t_d$, and $\eta = \gamma t_d$ and show that the result is the simultaneous equations $\xi = \alpha e^{\eta} \sin \xi$, $\eta = \alpha e^{\eta} \cos \xi$. From these obtain the equivalent equations $\eta = \xi \cot \xi$ and $\xi^2 = \alpha^2 e^{2\eta} - \eta^2$. Show how these can be solved graphically if α is known.

Section 10.12

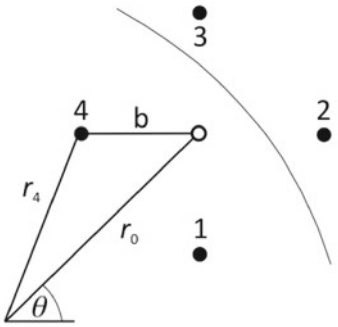
Problem 47. Find an equation relating L, the total amount of light energy per second reaching the retina, I, the intensity of the light (W m^{-2}), and R, the radius of the pupil. Calculate the gain $G = \partial L/\partial R$ and the logarithmic gain $g = (1/L)(\partial L/\partial R)$. Consider the two cases shown in the figure.

(a) There is uniform illumination of the pupil.

(b) The rectangle of illumination partially overlaps the pupil so that the area within the pupil is $a(R - b)$.



Problem 48. Suppose you measure the arrival time of an action potential wave front at four points (1–4) arranged in a diamond pattern each a distance b from a central point (open circle) located at (r_0, θ) relative to the origin. Use the steps below to calculate the wave front speed, direction, and curvature from these four measurements.



(a) Assume that the wave front is circular and propagates outward from the origin. Use the law of cosines to write r_1, r_2, r_3 and r_4 (the distance of each electrode from the origin) in terms of r_0, b and θ.
(b) Pull a factor r_0 outside the square root in each of your four expressions from part (a).
(c) Assume r_0 is much greater than b, and perform a Taylor expansion of each expression in terms of the small parameter $\xi = b/r_0$. Include terms that are constant, linear and quadratic in ξ.
(d) Write the arrival time at each of the four electrodes at $t_n = r_n/v$, where v is the wave speed.
(e) Let $\Delta t_{ij} = t_i - t_j$. Find expressions for Δt_{31} and Δt_{24} in terms of b, θ, and v. Solve these equations to determine v and θ in terms of Δt_{31}, Δt_{24}, and b.
(f) Find expressions for Δt_{14} and Δt_{23} in terms of b, θ, and v. Now (and this is the most difficult step), find an expression for the radius of curvature, r_0, in terms of b, Δt_{13}, Δt_{24}, Δt_{14}, and Δt_{23}.

Problem 49. Write a computer program to reproduce the numerical results in Fig. 10.35b and c. The calculation was originally performed by Glass and Mackey using the delay differential equation

$$\frac{dx}{dt} = \frac{\beta_0 x(t-\tau)}{1+x^n(t-\tau)} - \gamma x(t)$$

where x is the white blood cell count (equal to P/θ in the figure), $\beta_0 = 0.2$, $\gamma = 0.1$, and $n = 10$. The initial condition for x is 0.1. Figure 10.35b uses $\tau = 6$, and Fig. 10.35c uses $\tau = 20$. (See Sect. 6.14 for some guidance on how to solve differential equations numerically.)

References

Abraham R, Shaw CD (1992) Dynamics: the geometry of behavior, 2nd ed. Addison-Wesley, Reading

Bartlett AA, Braun TJ (1983) Death in a hot tub: the physics of heat stroke. Am J Phys 51(2):127–132

Begon M, Mortimer M, Thompson DJ (1996) Population ecology: a unified study of animals and plants, 3rd ed. Blackwell Science, Cambridge

Bub G, Shrier A, Glass L (2002) Spiral wave generation in heterogeneous excitable media. Phys Rev Lett 88:058101

Chick WL et al. (1977) Artificial pancreas using living beta cells: effects on glucose homeostasis in diabetic rats. Science 197:780–781

Clayton RH, Bernus O, Cherry EM, Dierckx H, Fenton FH, Mirabella L, Panfilov AV, Sachse FB, Seemann G, Zhang H (2011) Models of cardiac tissue electrophysiology: progress, challenges and open questions. Prog Biophys Mol Biol 104:22–48

El-Samad H, Goff JP, Khammash M (2002) Calcium homeostasis and parturient hypocalcemia: an integral feedback perspective. J Theor Biol 214:17–29

Epstein AE, Ideker RE (1995) Ventricular fibrillation. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 2nd ed. Saunders, Philadelphia, pp 927–933

Epstein IR, Kustin K, De Kepper P, Orban M (1983) Oscillating chemical reactions. Sci Am 248:112–123

Fox JJ, Bodenschatz E, Gilmour RF Jr (2002) Period-doubling instability and memory in cardiac tissue. Phys Rev Lett 89:138101

Garfinkel A, Spano ML, Ditto WL, Weiss JN (1992) Controlling cardiac chaos. Science 257:1230–1235

Garfinkel A, Kim Y-H, Voroshilovsky O, Qu Z, Kil JR, Lee M-H, Karagueuzian HS, Weiss JN, Chen P-S (2000) Preventing ventricular fibrillation by flattening cardiac restitution. Proc Natl Acad Sci U S A 97:6061–6066

Ginzburg L, Colyvan M (2004) Ecological orbits: how planets move and populations grow. Oxford University Press, Oxford

Glass L, Mackey MC (1988) From clocks to chaos. Princeton University Press, Princeton

Glass L, Nagai Y, Hall K, Talajic M, Nattel S (2002) Predicting the entrainment of reentrant cardiac waves using phase resetting curves. Phys Rev E Stat Nonlin Matter Phys 65:021908

Gray RA (2009) Rotors and spiral waves in the heart. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Springer, New York, pp 349–360

Guevara MR, Glass L, Shrier A (1981) Phase-locking, period-doubling bifurcations, and irregular dynamics in periodically stimulated cardiac cells. Science 214:1350–1353

Guyton AC (1995) Textbook of medical physiology, 9th ed. Saunders–Elsevier, Philadelphia

Guyton AC, Crowell JW, Moore JW (1956) Basic oscillating mechanism of Cheyne–Stokes breathing. Am J Physiol 187:395–398

Hastings HM, Fenton FH, Evans SJ, Hotomaroglu O, Geetha J, Gittelson K, Nilson J, Garfinkel A (2000) Alternans and the onset of ventricular fibrillation. Phys Rev E 62:4043–4048

Hayes ND (1950) Roots of the transcendental equation associated with a certain difference-differential equation. J Lond Math Soc 25:226–232

Hilborn RC (2000) Chaos and nonlinear dynamics, 2nd ed. Oxford University Press, New York

Jalife J, Moe GK (1976) Effect of electrotonic potential on pacemaker activity of canine Purkinje fibers in relation to parasystole. Circ Res 39:801–808

Kaplan D, Glass L (1995) Understanding nonlinear dynamics. Springer, New York

Keener JP, Sneyd J (2008a) Mathematical physiology I: cellular physiology, 2nd ed. Springer, New York

Keener JP, Sneyd J (2008b) Mathematical physiology II: systems physiology, 2nd ed. Springer, New York
Khammash M, El-Samad H (2004) Systems biology: from physiology to gene regulation. IEEE Control Syst Mag 24:62–76
Khoo MCK (2000) Physiological control systems: analysis, simulation and estimation. IEEE, New York
Le Duc PR Messner WC Wikswo JP (2011) How do control-based approaches enter into biology? Annu Rev Biomed Eng 13:369–396
Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
Luther S, Fenton FH, Kornreich BG, Squires A, Bittihn P, Hornung D, Zabel M, Flanders J, Gladuli A, Compoy L, Cherry EM, Luther G, Hasenfuss G, Krinsky VI, Pumir A, Gilmour RF, Bodenschatz E (2011) Low-energy control of electrical turbulence in the heart. Nature 475:235–239
Mackey MC, Glass L (1977) Oscillation and chaos in physiological control systems. Science 197:287–289
May RM (1976) Simple mathematical models with very complicated dynamics. Nature (London) 261:459–467
Mielczarek EV, Turner JS, Leiter D, Davis L (1983) Chemical clocks: experimental and theoretical models of nonlinear behavior. Am J Phys 51(1):32–42
Mines GR (1914) On circulating excitation on heart muscles and their possible relation to tachycardia and fibrillation. Trans R Soc Can 4:43–53
Murray JD (2007) Mathematical biology I: an introduction, 3rd ed. Springer, New York
Murray JD (2008) Mathematical biology II: spatial models and biomedical applications, 3rd ed. Springer, New York
Panfilov AV (2009) Theory of reentry. In: Zipes DP, Jalife J (eds) Cardiac electrophysiology: from cell to bedside, 5th ed. Saunders, Philadelphia, pp 329–338
Patton HD, Fuchs AF, Hille B, Scher AM, Steiner RF (eds) (1989) Textbook of physiology, 21st ed. Saunders, Philadelphia
Riggs DS (1970) Control theory and physiological feedback mechanisms. Williams and Wilkins, Baltimore
Scher AM, Young AC (1963) Servoanalysis of carotid sinus reflex effects on peripheral resistance. Circ Res 12:152–165
Stark L (1968) Neurological control systems: studies in bioengineering. Plenum, New York, pp 73–84
Stark LW (1984) The pupil as a paradigm for neurological control systems. IEEE Trans Biomed Eng 31:919–924
Stark L, Sherman PM (1957) A servoanalytic study of consensual pupil reflex to light. J Neurophysiol 20:17–26
Strogatz SH (1994) Nonlinear dynamics and chaos. Addison-Wesley, Reading
Strogatz SH (2003) Sync: the emerging science of spontaneous order. Hyperion, New York
Winfree AT (1987) When time breaks down. Princeton University Press, Princeton
Winfree AT (2001) The geometry of biological time, 2nd ed. Springer, New York
Witkowski FX, Kavanagh KM, Penkoske PA, Plonsey R (1993) In vivo estimation of cardiac transmembrane current. Circ Res 72(2):424–439
Witkowski FX, Plonsey R, Penkoske PA, Kavanagh KM (1994) Significance of inwardly directed transmembrane current in determination of local myocardial electrical activation during ventricular fibrillation. Circ Res 74(3):507–524
Witkowski FX, Kavanagh KM, Penkoske PA, Plonsey R, Spano ML, Ditto WL, Kaplan DT (1995) Evidence for determinism in ventricular fibrillation. Phys Rev Lett 75(6):1230–1233
Yehia AR, Jeandupeux D, Alonso F, Guevara MR (1999) Hysteresis and bistability in the direct transition from 1:1 to 2:1 rhythm in periodically driven single ventricular cells. Chaos 9:916–931

The Method of Least Squares and Signal Analysis 11

This chapter deals with three common problems in experimental science. The first is fitting a discrete set of experimental data with a mathematical function. The function usually has some parameters that must be adjusted to give a "best" fit. The second is to detect a periodic change in some variable, a signal, which may be masked by random changes, noise, superimposed on the signal. The third is to determine whether sets of apparently unsystematic data are from a random process or a process governed by deterministic chaotic behavior.

These techniques are used in many fields, including physiology and biophysics. The fitting techniques lead naturally to Fourier series, which are used extensively in image reconstruction and image analysis. Using least squares or Fourier series normally requires extensive computation. Commercial packages for making these calculations are readily available. The problems at the end of the chapter are often artificially designed for simple computation, rather than being "real." We hope that the chapter will help you develop some intuition for the techniques before you use the commercial packages.

This chapter is a self-contained discussion of signal analysis. It is a prerequisite to Chap. 12 on image reconstruction.

We will find that a periodic signal can be built up of sine waves of different frequencies, and that it is possible to speak of the *frequency spectrum* of a signal. The first five sections of the chapter show how to adjust the parameters in a polynomial or in a sum of sines and cosines to fit experimental data. Sections 11.5 and 11.6 discuss sine and cosine expansions for continuous periodic functions. Sections 11.7 and 11.8 introduce the cross-correlation and autocorrelation functions and their relation to the power spectrum. Sections 11.9 through 11.12 extend these techniques to pulses. Sections 11.13 and 11.14 introduce noise and the use of correlation functions to detect signals that are masked by noise.

Many linear feedback systems are most easily studied by how they respond to sinusoidal stimuli at various frequencies, and there are techniques using impulse or noise stimuli that provide the same information. Section 11.15 explains the frequency response of a linear system, and the next section describes the effect of a simple linear system on the power spectrum of Johnson noise. The next section introduces some of the concepts involved in testing data for chaotic behavior. Finally, Sect. 11.18 discusses stochastic resonance, where introducing noise into a nonlinear system can enhance a desired effect.

11.1 The Method of Least Squares and Polynomial Regression

In this section, we show how to approximate or "fit" a set of discrete data y_j with a polynomial function

$$y_j = y(x_j) = \sum_k a_k x_j^k.$$

Several criteria can be used to determine the "best" fit (Press et al. 1992, Sect. 15.7); the one described in this chapter is called the *method of least squares*. Instead of immediately deriving the general polynomial result, we first consider the simple (and rather useless) fit $y = x + b$ (the coefficient of x is unity), then the more useful linear fit $y = ax + b$.

11.1.1 The Simplest Example

Suppose that we wish to describe the data in Table 11.1 by a fitting function $y(x)$. A plot of the data suggests that a straight line will be a reasonable approximation to the data. For mathematical simplicity, we first try a line with unit slope but adjustable intercept:

$$y(x_j) = x_j + b. \tag{11.1}$$

Figure 11.1a plots y vs. x for different values of b. It is clear by inspection that the curves for $b = 1$ and $b = 2$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_11, © Springer International Publishing Switzerland 2015

Table 11.1 Sample data

x	y
1	2
4	6
5	7

are closer to the points than those for $b = 0$ or $b = 3$. For a quantitative measure of how good the fit is, we use the quantity

$$Q = \frac{1}{N}\sum_{j=1}^{N}\left[y_j - y(x_j)\right]^2, \tag{11.2}$$

which is called the *mean square error*. It is the square of the residuals (the differences between the measured values of y and the values of y calculated from the approximation to the data, $y_j - y(x_j)$) summed over all N data points and divided by N. It is reminiscent of the variance, with the mean replaced by the fitting function $y(x_j)$. The least-squares technique adjusts the parameters in the function $y(x_j)$ to make Q a minimum. Table 11.2 shows the steps in the calculation of Q for various values of b. Figure 11.1b shows how Q changes as b is changed.

It is tedious to calculate Q for many different values of b; instead we can treat this as a maximum–minimum problem in calculus. We write

$$\begin{aligned} Q &= \frac{1}{N}\sum_{j=1}^{N}(y_j - x_j - b)^2 \\ &= \frac{1}{N}\left[(y_1 - x_1 - b)^2 + (y_2 - x_2 - b)^2 + \cdots\right]. \end{aligned}$$

The derivative is

$$\begin{aligned} \frac{dQ}{db} &= -\frac{1}{N}\sum_{j=1}^{N}2(y_j - x_j - b) \\ &= \frac{1}{N}\left[-2(y_1 - x_1 - b) - 2(y_2 - x_2 - b) + \cdots\right]. \end{aligned}$$

Setting this equal to zero to find the extremum gives

$$\sum_{j=1}^{N} y_j = \sum_{j=1}^{N} x_j + \sum_{j=1}^{N} b$$

or, not bothering to show explicitly that the index ranges over all the data points,

$$\sum_j y_j = \sum_j x_j + Nb.$$

Using this result for the example above gives $15 = 10 + 3b$, or $b = 1.67$ for the smallest value of $Q = 0.22$.

11.1.2 A Linear Fit

The previous example was rather artificial, because for simplicity we did not allow the slope of the line to vary. The maximum–minimum procedure is easily extended to two or more parameters. If the fitting function is given by $y = ax + b$, then Q becomes

$$Q = \frac{1}{N}\sum_{j=1}^{N}(y_j - ax_j - b)^2.$$

At the minimum, both $\partial Q/\partial a = 0$ and $\partial Q/\partial b = 0$. The former gives

$$\frac{\partial Q}{\partial a} = \frac{2}{N}\sum_{j=1}^{N}(y_j - ax_j - b)(-x_j) = 0$$

or

$$\sum_j x_j y_j - a\sum_j x_j^2 - b\sum_j x_j = 0. \tag{11.3}$$

The latter gives

$$\frac{\partial Q}{\partial b} = \frac{2}{N}\sum_{j=1}^{N}(y_j - ax_j - b)(-1) = 0$$

or

$$\sum_{j=1}^{N} y_j - a\sum_{j=1}^{N} x_j - Nb = 0. \tag{11.4}$$

For the example in Table 11.1 $\sum x_j = 10$, $\sum y_j = 15$, $\sum x_j^2 = 42$, and $\sum x_j y_j = 61$. Therefore, Eqs. 11.3 and 11.4 become $42a + 10b = 61$ and $10a + 3b = 15$. These can be easily solved to give $a = 1.27$ and $b = 0.77$. The best straight-line fit to the data of Table 11.1 is $y = 0.77 + 1.27x$. The value of Q, calculated from Eq. 11.2, is 0.013. The best fit is plotted in Fig. 11.2. It is considerably better than the fit with the slope constrained to be one.

A general expression for the solution to Eqs. 11.3 and 11.4 is

$$a = \frac{N\left(\sum_{j=1}^{N} x_j y_j\right) - \left(\sum_{j=1}^{N} x_j\right)\left(\sum_{j=1}^{N} y_j\right)}{N\left(\sum_{j=1}^{N} x_j^2\right) - \left(\sum_{j=1}^{N} x_j\right)^2}, \tag{11.5a}$$

$$b = \frac{\sum_{j=1}^{N} y_j}{N} - \frac{a\left(\sum_{j=1}^{N} x_j\right)}{N} \equiv \overline{y} - a\overline{x}, \tag{11.5b}$$

where $\overline{x}$ and $\overline{y}$ are the means. In doing computations where the range of data is small compared to the mean, better numerical accuracy can be obtained from

$$a = S_{xy}/S_{xx}, \tag{11.5c}$$

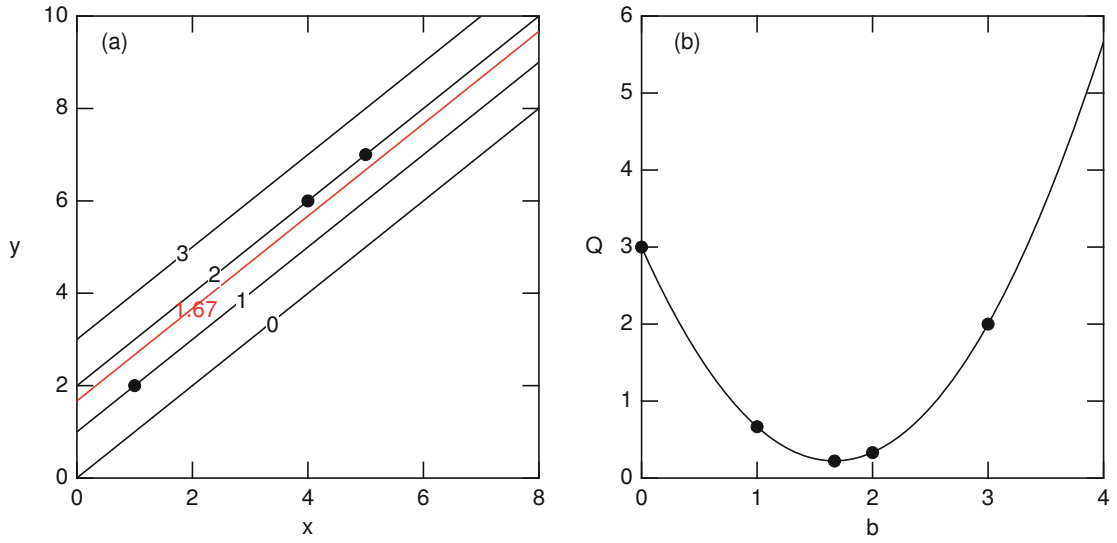


Fig. 11.1 Fits to the data of Table 11.1 by a curve of the form $y = x + b$. **a** Plots of y vs. x. **b** Plot of Q vs. b. Q is defined in Eq. 11.2

Table 11.2 Calculation of Q for the example of Eq. 11.1

Index			$b = 0$		$b = 1$		$b = 2$	
j	x_j	y_j	$y(x_j)$	$[y_j - y(x_j)]^2$	$y(x_j)$	$[y_j - y(x_j)]^2$	$y(x_j)$	$[y_j - y(x_j)]^2$
1	1	2	1	1	2	0	3	1
2	4	6	4	4	5	1	6	0
3	5	7	5	4	6	1	7	0
Sum				9		2		1
Q				3		0.67		0.33

using the sums

$$S_{xx} = \sum_{j=1}^{N} \left(x_j - \overline{x}\right)^2, \tag{11.5d}$$

and

$$S_{xy} = \sum_{j=1}^{N} (x_j - \overline{x})(y_j - \overline{y}). \tag{11.5e}$$

11.1.3 A Polynomial Fit

The method of least squares can be extended to a polynomial of arbitrary degree. The only requirement is that the number of adjustable parameters (which is one more than the degree of the polynomial) be less than the number of data points. If this requirement is not met, the equations cannot be solved uniquely; see Problem 8. If the polynomial is written as

$$y(x_j) = a_0 + a_1 x_j + a_2 x_j^2 + \cdots + a_n x_j^n = \sum_{k=0}^{n} a_k x_j^k, \tag{11.6}$$

then the expression for the mean square error is

$$Q = \frac{1}{N} \sum_{j=1}^{N} \left(y_j - \sum_{k=0}^{n} a_k x_j^k \right)^2. \tag{11.7}$$

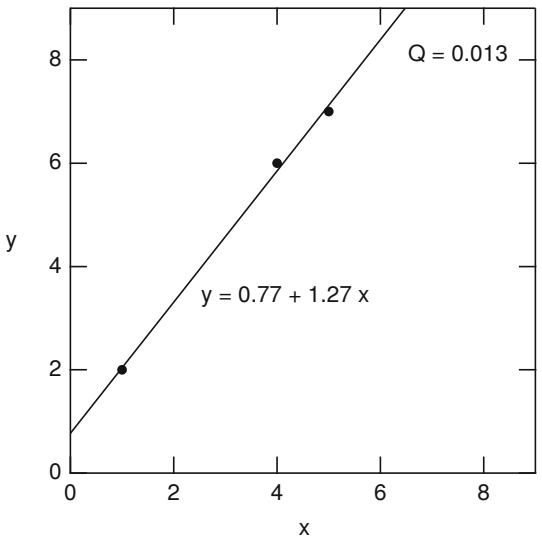


Fig. 11.2 The best fit to the data of Table 11.1 with the function $y = ax + b$. Both the slope and the intercept have been chosen to minimize Q

Index j ranges over the data points; index k ranges over the terms in the polynomial. This expression for Q can be differentiated with respect to one of the $n + 1$ parameters, say,

a_m:

$$\frac{\partial Q}{\partial a_m} = \frac{2}{N}\sum_{j=1}^{N}\left[\left(y_j - \sum_{k=0}^{n} a_k x_j^k\right)(-x_j^m)\right].$$

Setting this derivative equal to zero gives

$$\sum_{j=1}^{N} y_j x_j^m = \sum_{k=0}^{n}\sum_{j=1}^{N} a_k x_j^k x_j^m = \sum_{k=0}^{n} a_k \sum_{j=1}^{N} x_j^{k+m}.$$

This is one of the equations we need. Doing the same thing for all values of m, $m = 0, 1, 2, \ldots, n$, we get $n+1$ equations that must be solved simultaneously for the $n+1$ parameters $a_0, a_1, \ldots, a_n$.

For $m = 0$:

$$\sum_{j=1}^{N} y_j = Na_0 + a_1\sum_{j=1}^{N} x_j + a_2\sum_{j=1}^{N} x_j^2 + \cdots + a_n\sum_{j=1}^{N} x_j^n. \tag{11.8a}$$

For $m = 1$:

$$\sum_{j=1}^{N} x_j y_j = a_0\sum_{j=1}^{N} x_j + a_1\sum_{j=1}^{N} x_j^2 + a_2\sum_{j=1}^{N} x_j^3 + \cdots + a_n\sum_{j=1}^{N} x_j^{n+1}. \tag{11.8b}$$

For $m = n$:

$$\sum_{j=1}^{N} x_j^n y_j = a_0\sum_{j=1}^{N} x_j^n + a_1\sum_{j=1}^{N} x_j^{n+1} + a_2\sum_{j=1}^{N} x_j^{n+2} + \cdots + a_n\sum_{j=1}^{N} x_j^{2n}. \tag{11.8c}$$

Solving these equations is not as formidable a task as it seems. Given the data points (x_j, y_j), the sums are all evaluated. When these numbers are inserted in Eqs. 11.8, the result is a set of $n+1$ simultaneous equations in the $n+1$ unknown coefficients a_k. This technique is called *linear least-squares fitting of a polynomial* or *polynomial regression*. Routines for solving the simultaneous equations or for carrying out the whole procedure are readily available.

11.1.4 Variable Weighting

The least-squares technique described here gives each data point the same weight. If some points are measured more accurately than others, they should be given more weight. This can be done by assigning each data point its own weight, replacing Eq. 11.2 by

$$Q = \frac{1}{N}\sum_{j=1}^{N} w_j\left[y_j - y(x_j)\right]^2. \tag{11.9}$$

For example, repeated measurements of y_j for a particular x_j might give results that are Gaussian-distributed about a mean value with standard deviation σ_j and variance σ_j^2. (See Appendices G and I.) Then it would be appropriate to use the weight $w_j = 1/\sigma_j^2$. Setting all the weights equal to 1 as we have been doing is correct only if the variance is the same for each y_j. It is easy to show that the effect of this weighting is to add a factor of $1/\sigma_j^2$ to each term in the sums in Eqs. 11.8.

This analysis assumes that errors exist only in the y values. If there are errors in the x values as well, it is possible to make an approximate correction based on an effective error in the y values (Orear 1982) or to use an iterative but exact least-squares method (Lybanon 1984).

11.2 Nonlinear Least Squares

If we need to fit a single exponential to a set of data, we have two choices:

Method 1. Use semilog paper or make a linear fit to $v = \log y$.

Method 2. Use a statistical package that makes a fit directly to $y(x)$ using the method of nonlinear least squares.

Both methods can be used either with uniform weighting of each data point, Eq. 11.2, or with individual weightings, Eq. 11.9.

The linear least-squares technique, Method 1, can only be used to fit data with a single exponential $y = ae^{-bx}$, where a and b are to be determined. Take logarithms of each side of the equation:

$$\log y = \log a - bx\log e,$$

$$v = a' - b'x.$$

This is a linear equation, and constants a' and b' can be determined using Eqs. 11.5. With a sum of two or more exponentials, this method does not work. We will see below that even when it does work, Method 1 should not be used.

Method 2 can be used for any fitting function, for example $y = a^{-bx}$, $y = ae^{-bx} + c$, or even a sum of exponentials:

$$y = a_1 e^{-b_1 x} + a_2 e^{-b_2 x} + \cdots.$$

When we try to minimize Q by the technique of the previous section, we find that when the derivatives of this fitting function are set equal to zero, the equations in a_1, a_2, etc., are

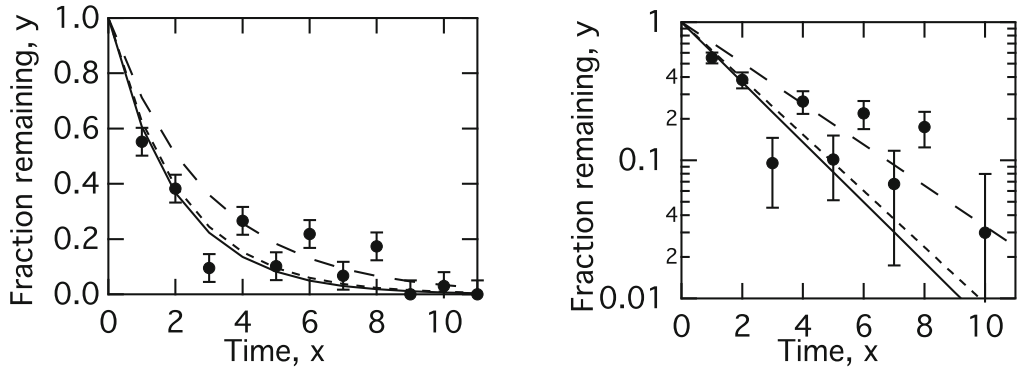


Fig. 11.3 A comparison of fitting techniques when each data point has the same weight. The solid line shows the original model, $y = e^{-0.5x}$. The data points have a Gaussian-distributed random error with standard deviation 0.09. The small-dashed line is a nonlinear least squares fit to y. The value of b is 0.47. The longer-dashed line is a linear-least squares fit to log-transformed data, $v = \log y$. The value of b is 0.34

linear if we assume that the values of the b_k are known. However, the equations for determining the b's are transcendental equations that are quite difficult to solve.

The problem can be solved using the technique of *nonlinear least squares*. In its simplest form, one makes an initial guess for each parameter[1] $b_{10}, b_{20}, \ldots, b_{k0}$ and says that the correct value of each b is given by $b_k = b_{k0} + h_k$. The calculated value of y is written as a Taylor's series expansion with all the derivatives evaluated for $b_{10}, b_{20}, \ldots$:

$$y(x_j; b_1, b_2, \ldots) = y(x_j; b_{10}, b_{20}, \ldots) + \frac{\partial y}{\partial b_1} h_1 + \frac{\partial y}{\partial b_2} h_2 + \cdots$$

Since y and its derivatives can be evaluated using the current guess for each b, the expression is linear in the h_k, and the linear least-squares technique can be used to determine the values of the h_k that minimize Q. After each h_k has been determined, the revised values $b_k = b_{k0} + h_k$ are used as the initial guesses, and the process is repeated until a minimum value of Q is found. The technique is not always stable; it can overshoot and give too large a value for h_k. There are many ways to improve the process to ensure more rapid convergence. The most common is called the Levenberg–Marquardt method (see Bevington and Robinson 2003 or Press et al. 1992).

Using the method of nonlinear least squares used to be quite difficult. Now, however, it is available in many statistical packages, such as R (see The R Project at http://www.r-project.org/).

Referring back to Fig. 2.6, we see that if each data point on a linear plot has the same weight (variance), then the weights of the log-transformed data should be very different. This fact has not always been appreciated. It is very easy to use Method 1 (take the logarithm of y and make a linear least-squares fit to the transformed data) giving the same weight to each data point. This can give substantial errors in the parameters.

For example, some ecologists study the decomposition of litter on the forest floor. They make a number of porous litter bags, fill each one with the same mass of litter material $m(0)$, put them on the forest floor, and retrieve bags at various later times. If several bags are retrieved at the same time, there is much more scatter in the mass from bag to bag than there is in the original mass measurement $m(0)$. The dependent variable is the fraction of mass remaining: $y(x_j) = m(x_j)/m(0)$. A model, such as simple exponential decay, is used to fit y. There is no data point for $x = 0$ but the fit $y = ae^{-bx}$ must have $a = 1$. Sometimes there is nothing left in a litter bag and $y(x_j) = 0$.

Incorrect analyses occur frequently in the literature. Adair et al. (2010) studied the way decomposition data have been analyzed in 498 papers. They also compared the results of using Method 1 and Method 2 on both real and artificial data.

At least 40 % of the papers used Method 1; only 15 % explicitly stated that they used Method 2. The other papers were not clear about the method used. The distinction between the methods is shown in Fig. 11.3. The original model was simple exponential decay with $b = 0.5$: $y = e^{-0.5x}$. Gaussian-distributed random noise with standard deviation $\sigma = 0.09$ was added to each data point. (These values are typical of decomposition experiments.) When this was fit using Method 1, linear least squares on the log-transformed data with all points weighted the same, the estimate of b was 0.34, a significant underestimate. With Method 2, a nonlinear least squares fit to the original data with equal weighting, the value of b was 0.47.

There is another problem with Method 1. If one of the values of $y(x_j) = 0$, it cannot be log-transformed. Some investigators have substituted arbitrary small values that are very far from the fit in the semilog plot, greatly distorting the estimated value of b. It is better to delete these data points in Method 1. Zero values present no problem with Method 2.

[1] The parameters a_k can either be included in the parameter list, or the values of a_k for each trial set b_k can be determined by linear least squares.

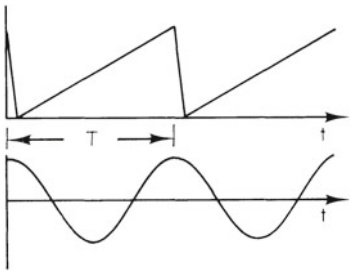


Fig. 11.4 Two different periodic functions

Nearly 60 % of the papers analyzed made the fit using $y = ae^{-bx}$ and adjusting a to improve the fit. This gave a smaller Q but the estimating function was not 1 at $x = 0$, and the value of b was quite different than if a is set equal to 1.

In the simulation studies Adair et al. found that Method 1 gave reliable parameter estimates only when the errors were lognormally distributed (i.e., Gaussian in the log-transformed data). Method 2, however, gave good parameter estimates in all cases. In the real data sets, Method 2 gave consistently larger estimates of b than Method 1. (This same effect was observed in the simulation studies.) The differences were most pronounced for rapid decay (large values of b).

The moral: use nonlinear least squares, Method 2!

The same problem occurs while using log–log plots for allometric scaling (Packard 2009).

11.3 The Presence of Many Frequencies in a Periodic Function

A function y that repeats itself after a time[2] T is said to be *periodic* with period T. The mathematical description of this periodicity is

$$y(t + T) = y(t). \tag{11.10}$$

Two examples of functions with period T are shown in Fig. 11.4. One of these functions is a sine wave, $y(t) = A\sin(\omega_0 t - \phi)$, where A is the amplitude, ω_0 is the *angular frequency*, and ϕ is the phase of the function. Changing the amplitude changes the height of the function. Changing the phase shifts the function along the time axis. The sine function repeats itself when the argument shifts by 2π radians. It repeats itself after time T, where $\omega_0 T = 2\pi$. Therefore the angular frequency is related to the period by

$$\omega_0 = \frac{2\pi}{T}. \tag{11.11}$$

[2] Although we speak of t and time, the technique can be applied to any independent variable if the dependent variable repeats as in Eq. 11.10. Zebra stripes are (almost) periodic functions of position.

(The units of ω_0 are radian s^{-1}, but radians are dimensionless.) It is completely equivalent to write the function in terms of the *frequency* as $y(t) = A\sin(2\pi f_0 t - \phi)$. The frequency f_0 is the number of cycles per second. Its units are s^{-1} or hertz (Hz) (hertz is not used for angular frequency):

$$f_0 = \frac{1}{T} = \frac{\omega_0}{2\pi}. \tag{11.12}$$

It is possible to write function y as a sum of a sine term and a cosine term instead of using phase ϕ:

$$\begin{aligned} y(t) &= A\sin(\omega_0 t - \phi) = A(\sin\omega_0 t\cos\phi - \cos\omega_0 t\sin\phi) \\ &= (A\cos\phi)\sin\omega_0 t - (A\sin\phi)\cos\omega_0 t \\ &= S\sin\omega_0 t - C\cos\omega_0 t. \end{aligned} \tag{11.13}$$

The upper function graphed in Fig. 11.4 also has period T.

Harmonics are integer multiples of the fundamental frequency. They have the time dependence $\cos(k\omega_0 t)$ or $\sin(k\omega_0 t)$, where $k = 2, 3, 4, \ldots$. These also have period T. (They also have shorter periods, but they still satisfy the definition Eq. 11.10 for a function of period T.)

We can generate periodic functions of different shapes by combining various harmonics. Different combinations of the fundamental, third harmonic, and fifth harmonic are shown in Fig. 11.5. In this figure, part a is a pure sine wave, parts b and c have some third harmonic added with a different phase in each case, and parts d and e show the addition of a fifth harmonic term to part b with different phases.

An *even function* is one for which $y(t) = y(-t)$. For an *odd function*, $y(t) = -y(-t)$. The cosine is even, and the sine is odd. A sum of sine terms gives an odd function. A sum of cosine terms gives an even function.

11.4 Fourier Series for Discrete Data

The ability to adjust the amplitude of sines and cosines to approximate a specific shape suggests that discrete periodic data can be fitted by a function of the form

$$\begin{aligned} y(t_j) &= a_0 + \sum_{k=1}^{n} a_k\cos(k\omega_0 t_j) + \sum_{k=1}^{n} b_k\sin(k\omega_0 t_j) \\ &= a_0 + \sum_{k=1}^{n} a_k\cos(k2\pi f_0 t_j) + \sum_{k=1}^{n} b_k\sin(k2\pi f_0 t_j). \end{aligned} \tag{11.14}$$

It is important to note that if we have a set of data to fit, in some cases we may not know the actual period of the data; we sample for some interval of length T. In that case the period $T = 2\pi/\omega_0$ is a characteristic of the fitting function that we calculate, not of the data being fitted.

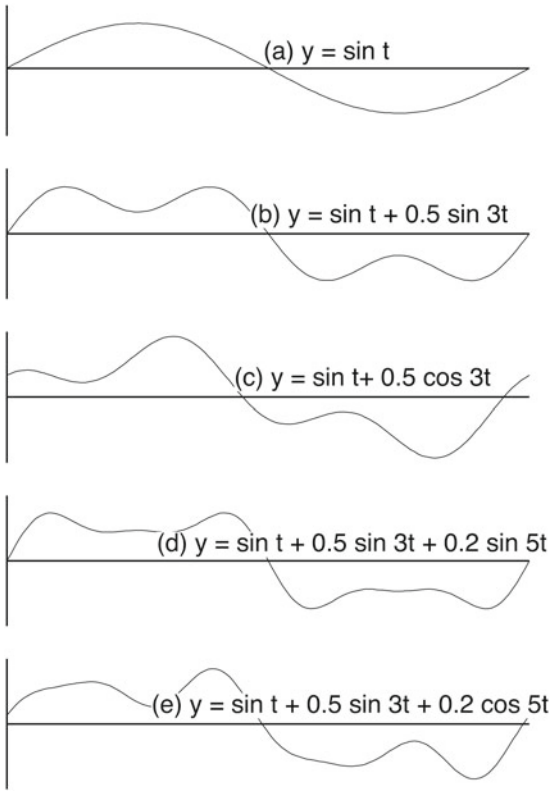


Fig. 11.5 Various periodic signals made by adding sine waves that are harmonically related. Each signal has an angular frequency $\omega_0 = 1$ and a period $T = 2\pi$

There are $2n + 1$ parameters $(a_0; a_1, \ldots, a_n; b_1, \ldots, b_n)$. Since there are N independent data points, there can be at most N independent coefficients. Therefore $2n + 1 \leq N$, or[3]

$$n \leq \frac{N-1}{2}. \tag{11.15}$$

This means that there must be at least two samples per period at the highest frequency present. This is known as the *Nyquist sampling criterion*.

11.4.1 Determining the Parameters

If the least-squares criterion is used to determine the parameters, Eq. 11.14 is a *Fourier-series* representation of the data. Using the least-squares criterion to determine the coefficients to fit N data points requires minimizing the mean square residual

$$Q = \frac{1}{N}\sum_{j=1}^{N}\left[y_j - a_0 - \sum_{k=1}^{n} a_k \cos(k\omega_0 t_j) - \sum_{k=1}^{n} b_k \sin(k\omega_0 t_j)\right]^2. \tag{11.16}$$

The derivatives that must be set to zero are

$$\frac{\partial Q}{\partial a_0} = -\frac{2}{N}\sum_{j=1}^{N}\left[\left(y_j - a_0 - \sum_{k=1}^{n} a_k \cos(k\omega_0 t_j) - \sum_{k=1}^{n} b_k \sin(k\omega_0 t_j)\right)(1)\right],$$

$$\frac{\partial Q}{\partial a_m} = -\frac{2}{N}\sum_{j=1}^{N}\left[\left(y_j - a_0 - \sum_{k=1}^{n} a_k \cos(k\omega_0 t_j) - \sum_{k=1}^{n} b_k \sin(k\omega_0 t_j)\right)\cos(m\omega_0 t_j)\right],$$

and

$$\frac{\partial Q}{\partial b_m} = -\frac{2}{N}\sum_{j=1}^{N}\left[\left(y_j - a_0 - \sum_{k=1}^{n} a_k \cos(k\omega_0 t_j) - \sum_{k=1}^{n} b_k \sin(k\omega_0 t_j)\right)\sin(m\omega_0 t_j)\right].$$

Setting each derivative equal to zero and interchanging the order of the summations give $2n + 1$ equations analogous to Eq. 11.8. The first is

$$\sum_{j=1}^{N} y_j = N a_0 + \sum_{k=1}^{n} a_k \sum_{j=1}^{N} \cos(k\omega_0 t_j) + \sum_{k=1}^{n} b_k \sum_{j=1}^{N} \sin(k\omega_0 t_j). \tag{11.17}$$

There are n equations of the form

$$\sum_{j=1}^{N} y_j \cos(m\omega_0 t_j) = a_0 \sum_{j=1}^{N} \cos(m\omega_0 t_j) + \sum_{k=1}^{n} a_k \sum_{j=1}^{N} \cos(k\omega_0 t_j)\cos(m\omega_0 t_j) + \sum_{k=1}^{n} b_k \sum_{j=1}^{N} \sin(k\omega_0 t_j)\cos(m\omega_0 t_j) \tag{11.18}$$

[3] For equally spaced data and N even, there are actually $n = N/2 + 1$ values of a_k and $n = N/2 - 1$ values of b_k. (We will find from Eq. 11.26c that b_k for $k = N/2$ is identically zero). Thus, there are N parameters and N coefficients. We will ignore this point in this chapter, since for large N it makes little difference.

for $m = 1, \ldots, n$, and n more of the form

$$\sum_{j=1}^{N} y_j \sin(m\omega_0 t_j) = a_0 \sum_{j=1}^{N} \sin(m\omega_0 t_j) + \sum_{k=1}^{n} a_k \sum_{j=1}^{N} \cos(k\omega_0 t_j) \sin(m\omega_0 t_j) + \sum_{k=1}^{n} b_k \sum_{j=1}^{N} \sin(k\omega_0 t_j) \sin(m\omega_0 t_j). \tag{11.19}$$

Since the t_j are known, each of the sums over the data points (index j) on the right hand side can be evaluated independent of the y_j.

11.4.2 Equally Spaced Data Points Simplify the Equations

If the data points are equally spaced, the equations become much simpler. There are N data points spread out over an interval T: $t_j = jT/N = 2\pi j/N\omega_0$, $j = 1, \ldots, N$. The arguments of the sines and cosines are of the form $(2\pi jk/N)$. One can show that

$$\sum_{j=1}^{N} \cos\left(\frac{2\pi jk}{N}\right) = \begin{cases} N, & k = 0 \text{ or } k = N, \\ 0 & \text{otherwise} \end{cases} \tag{11.20}$$

$$\sum_{j=1}^{N} \sin\left(\frac{2\pi jk}{N}\right) = 0, \quad \text{for all } k, \tag{11.21}$$

$$\sum_{j=1}^{N} \cos\left(\frac{2\pi jk}{N}\right) \cos\left(\frac{2\pi jm}{N}\right) = \begin{cases} N/2, & k = m \text{ or } k = N - m, \\ 0 & \text{otherwise}, \end{cases} \tag{11.22}$$

$$\sum_{j=1}^{N} \sin\left(\frac{2\pi jk}{N}\right) \sin\left(\frac{2\pi jm}{N}\right) = \begin{cases} N/2, & k = m, \\ -N/2, & k = N - m, \\ 0 & \text{otherwise}, \end{cases} \tag{11.23}$$

$$\sum_{j=1}^{N} \sin\left(\frac{2\pi jk}{N}\right) \cos\left(\frac{2\pi jm}{N}\right) = 0 \text{ for all } k. \tag{11.24}$$

Due to these properties, Eqs. 11.17–11.19 become a set of independent equations when the data are equally spaced:

$$a_0 = \frac{1}{N} \sum_{j=1}^{N} y_j, \tag{11.25a}$$

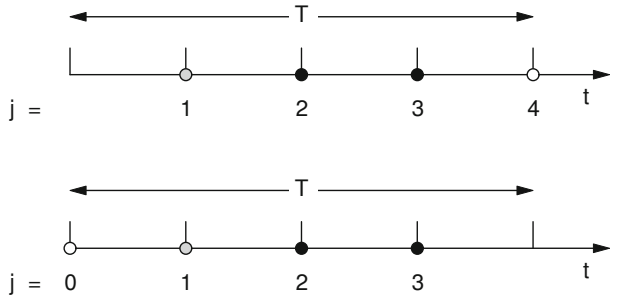


Fig. 11.6 A case where $N = 4$. The values of time are spaced by T/N and distributed uniformly. In the *top case* the values of j range from 1 to N. In the *lower case* they range from 0 to $N - 1$. The values of all the trigonometric functions are the same for $j = 0$ and for $j = N$

$$a_m = \frac{2}{N} \sum_{j=1}^{N} y_j \cos\left(\frac{2\pi jm}{N}\right), \tag{11.25b}$$

$$b_m = \frac{2}{N} \sum_{j=1}^{N} y_j \sin\left(\frac{2\pi jm}{N}\right). \tag{11.25c}$$

11.4.3 The Standard Form for the Discrete Fourier Transform

It is customary to change the notation to make the equations more symmetric. Figure 11.6 shows the four different times corresponding to $N = 4$ with $j = 1, 2, 3, 4$. Because of the periodicity of the sines and cosines, $j = N$ gives exactly the same value of a sine or cosine as does $j = 0$. Therefore, if we reassign the data point corresponding to $j = N$ to have the value $j = 0$ and sum from 0 to $N - 1$, the sums will be unchanged:

$$a_0 = \frac{1}{N} \sum_{j=0}^{N-1} y_j, \tag{11.26a}$$

$$a_m = \frac{2}{N} \sum_{j=0}^{N-1} y_j \cos\left(\frac{2\pi jm}{N}\right), \tag{11.26b}$$

$$b_m = \frac{2}{N} \sum_{j=0}^{N-1} y_j \sin\left(\frac{2\pi jm}{N}\right). \tag{11.26c}$$

For equally spaced data the function can be written as

$$y_j = y(t_j) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{2\pi jk}{N}\right) + \sum_{k=1}^{n} b_k \sin\left(\frac{2\pi jk}{N}\right). \tag{11.26d}$$

You can show (see the problems at the end of this chapter) that the symmetry and antisymmetry in Eqs. 11.22 and 11.23 for $k = N - m$ means that Eqs. 11.25 and 11.26 for $k > N/2$ repeat those for $k < N/2$. We can use this fact to make the equations more symmetric by changing the factor in front of

the summations in Eqs. 11.26b and 11.26c to be $1/N$ instead of $2/N$ and extending the summation in Eq. 11.26d all the way to $n = N - 1$. Since $\cos(0) = 1$ and $\sin(0) = 0$, we can include the term a_0 by including $k = 0$ in the sum. We then have the set of equations

$$y_j = y(t_j) = \sum_{k=0}^{N-1} a_k \cos\left(\frac{2\pi jk}{N}\right) + \sum_{k=0}^{N-1} b_k \sin\left(\frac{2\pi jk}{N}\right), \tag{11.27a}$$

$$a_k = \frac{1}{N}\sum_{j=0}^{N-1} y_j \cos\left(\frac{2\pi jk}{N}\right), \tag{11.27b}$$

$$b_k = \frac{1}{N}\sum_{j=0}^{N-1} y_j \sin\left(\frac{2\pi jk}{N}\right). \tag{11.27c}$$

This set of equations is the usual form for the *discrete Fourier transform*. We will continue to use our earlier form, Eqs. 11.26, in the rest of this chapter.

11.4.4 Complex Exponential Notation

The Fourier transform is usually written in terms of complex exponentials. We have avoided using complex exponentials. They are not necessary for anything done in this book. The sole advantage of complex exponentials is to simplify the notation. The actual calculations must be done with real numbers. Since you will undoubtedly see complex notation in other books, the notation is included here for completeness.

The numbers that we have been using are called real numbers. The number $i = \sqrt{-1}$ is called an *imaginary number*. A combination of a real and imaginary number is called a *complex number*. The remarkable property of imaginary numbers that make them useful in this context is that

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{11.28}$$

If we define the complex number $Y_k = a_k - ib_k$, we can write Eqs. 11.27 as

$$Y_k = \frac{1}{N}\sum_{j=0}^{N-1} y_j e^{-i2\pi jk/N} \tag{11.29a}$$

and

$$y_j = \sum_{k=0}^{N-1} Y_k e^{i2\pi jk/N}. \tag{11.29b}$$

Since our function y is assumed to be real, in the second equation we keep only the real part of the sum. To repeat: this gives only a more compact notation. It does not save in the actual calculations.

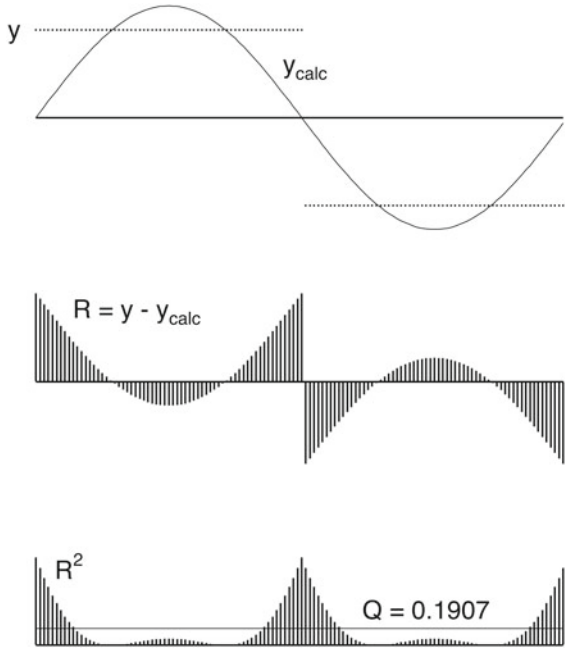


Fig. 11.7 A square wave $y(t_j)$ and the calculated function $y(t) = b_1 \sin(\omega_0 t)$ are shown, along with the residuals and the squares of the residuals for each point. The value of b_1 is $4/\pi$, which minimizes Q for that term

Table 11.3 Fourier coefficients obtained for a square wave fit

Term k	a_k	b_k
0	0.000	
1	0.000	1.273
2	0.000	0.000
3	0.000	0.424
4	0.000	0.000
5	0.000	0.253
6	0.000	0.000
7	0.000	0.181

11.4.5 Example: The Square Wave

Figures 11.7–11.10 show fits to a square wave with 128 data points. The function is $y_0 = 0$, $y_j = 1$, $j = 1, \ldots, 63$, $y_{64} = 0$, and $y_j = -1$, $j = 65, \ldots, 127$. This is an odd function of t. Therefore, the series should contain only sine terms; all a_k should be zero. The calculated coefficients are shown in Table 11.3. The even values of the b_k vanish. We will see why below.

Figure 11.7 shows the square wave as dots and $y(x)$ as a smooth curve when $b_1 = 1.273$ and all the other coefficients are zero. This provides the minimum Q obtainable with a single term. Figure 11.8 shows why Q is larger for any other value of b_1. Figure 11.9 shows the terms for $k = 1$ and $k = 3$. The value of Q is further reduced. Figure 11.10 shows why even terms do not reduce Q. In this case $b_2 = 0.5$ has been added to b_1. The fit is improved for the regions $0 < t < T/4$ and $3T/4 < t < T$, but between those regions the fit is made worse.

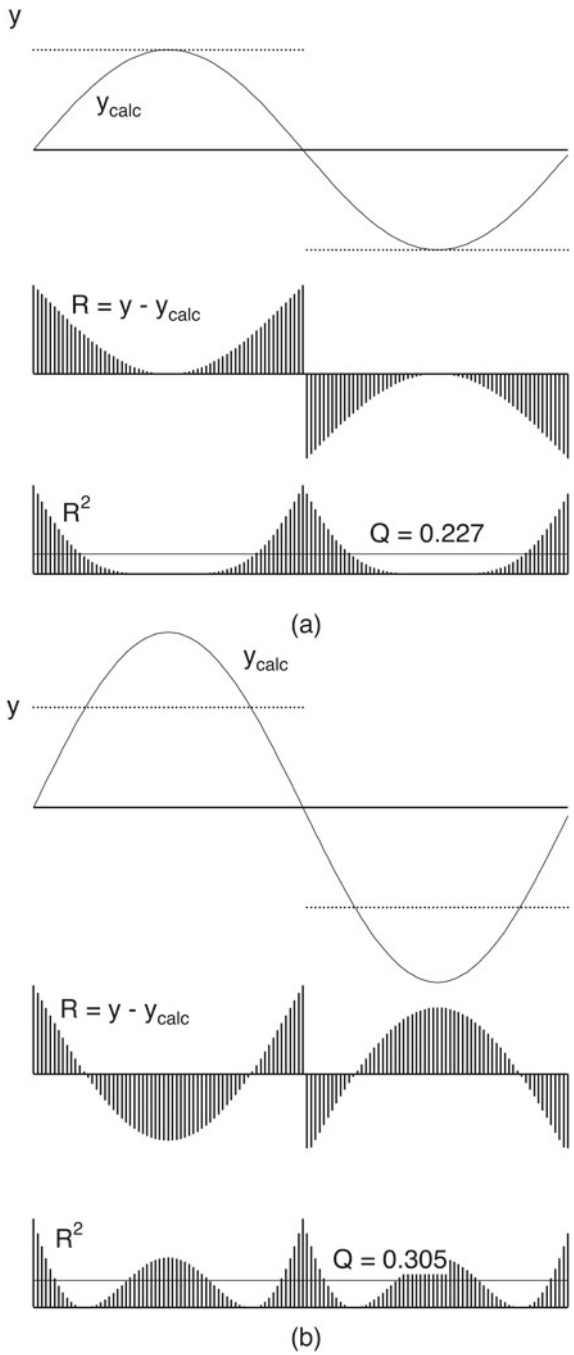


Fig. 11.8 A single term is used to approximate the square wave. **a** $b_1 = 1.00$, which is too small a value. **b** $b_1 = 1.75$, which is too large. In both cases Q is larger than the minimum value for a single term, shown in Fig. 11.7

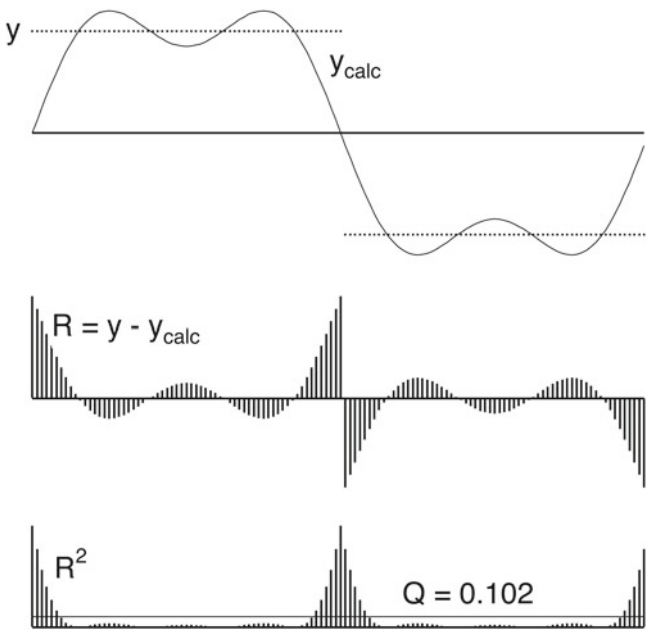


Fig. 11.9 Terms b_1 and b_3 have their optimum values. Q is smaller than in Fig. 11.7

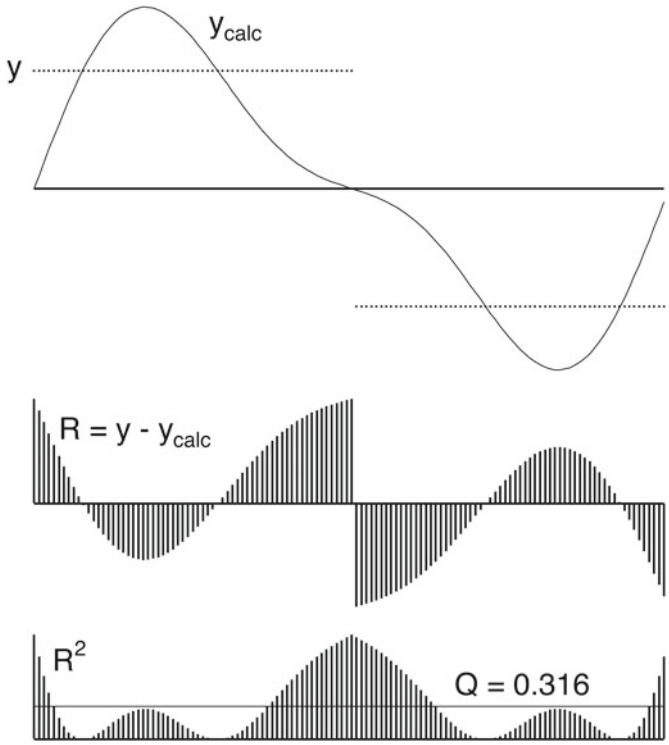


Fig. 11.10 This figure shows why even terms do not contribute. A term $b_2 = 0.5$ has been added to a term with the correct value of b_1. It improves the fit for $t < T/4$ and $t > 3T/4$ but makes it worse between $T/4$ and $3T/4$

11.4.6 Example: When the Sampling Time is not a Multiple of the Period of the Signal

The discussion just after Eq. 11.14 pointed out that in some cases we may not know the actual period and fundamental

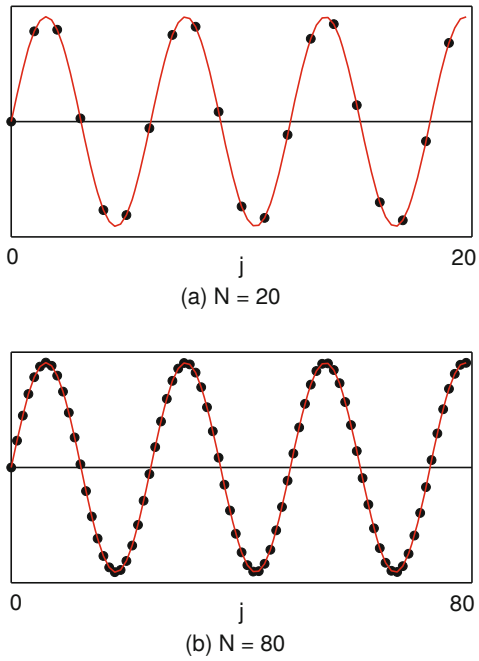


Fig. 11.11 Sine wave $y_j = \sin\left[3.3 \times 2\pi j/N\right]$ with 20 data points (**a**) and with 80 data points (**b**). The sampling time is not an integral number of periods

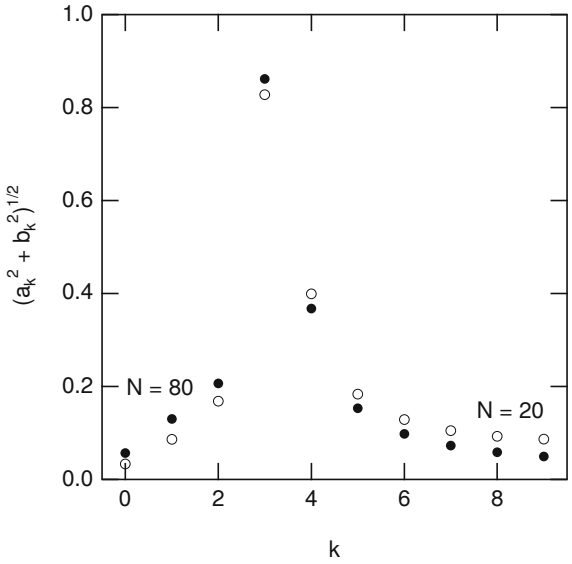


Fig. 11.12 The amplitude of the mixed sine and cosine coefficients $(a_k^2 + b_k^2)^{1/2}$ vs k for the function $y_j = \sin 3.3 \times 2\pi j/N$. The signal is sampled for 20 (open circles) or 80 (solid circles) data points. The amplitude spectrum is nearly independent of the number of samples

frequency ω_0 of the data. If we do know the actual period and the data points y_j are a sine or cosine with exact frequency ω_0 or a harmonic, and if no random errors are superimposed on the data, then only the coefficients corresponding to those frequencies will be nonzero. The reason is that if the function is exactly periodic, then by sampling for one period we have effectively sampled for an infinite time.

If the measuring duration T is not an integral multiple of the period of the signal—that is, the frequency of the signal y is not an exact harmonic of ω_0—then the Fourier series contains terms at several frequencies. This is shown in Figs. 11.11 and 11.12 for the data $y_j = \sin\left[3.3 \times 2\pi j/N\right]$. Figure 11.11 shows the y_j for $N = 20$ and $N = 80$ samples during the period of the measurement. For 20 samples, $n = 9$; for 80 samples, $n = 39$. Figure 11.12 shows $(a_k^2 + b_k^2)^{1/2}$ for both sample sets for $k = 0$ to 9, calculated using Eqs. 11.26. For 80 samples, the value of $(a_k^2 + b_k^2)^{1/2}$ is very small for $k > 9$ and is not plotted. The frequency spectrum is virtually independent of the number of samples. There is a zero-frequency component because there is a net positive area under the curve. The largest amplitude occurs for $k = 3$. If one imagines a smooth curve drawn through the histogram, its peak would be slightly above $k = 3$. We will see later that the width of this curve depends on the duration of the measurement, T.

Figure 11.13 shows the fit to the data of Fig. 11.11a. Since the data points had no errors, the fitting function with 20 parameters passes through each of the 20 data points. However,

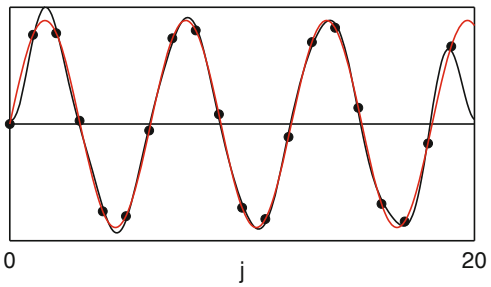


Fig. 11.13 The solid line shows the calculated fit for the 20 data points in Fig. 11.11a. The red line is the same as the red lines in Fig. 11.11

it does not match y of Fig. 11.11a at other points. Note particularly the difference between the function near $j = 1$ and near $j = 19$.

11.4.7 Example: Spontaneous Births

Figure 11.14 shows the number of spontaneous births per hour vs local time of day for 600,000 live births in various parts of the world. The basic period is 24 h; there is also a component with $k = 3$ ($T = 8$ h). These data were reported by Kaiser and Halberg (1962). More recent data show peaks at different times (Anderka et al. 2000). Changes might be due to a difference in the duration of labor.

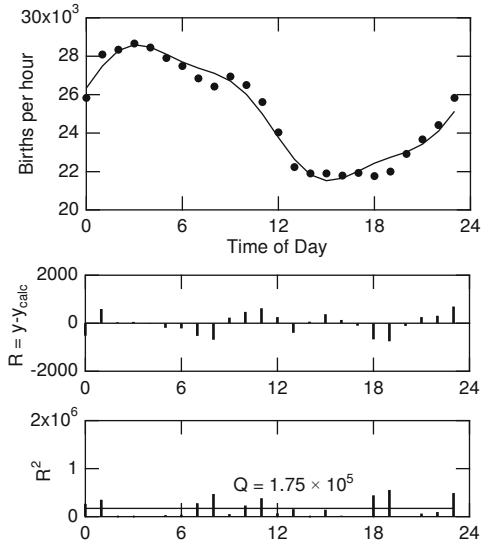


Fig. 11.14 Data on the number of spontaneous births per hour, fit with terms having periods of 24 and 8 h

11.4.8 Example: Photosynthesis in Plants

Tobacco plant leaves were exposed to white light similar to sunlight, with the amplitude varying sinusoidally with a frequency ω_0 corresponding to a period of 60 or 80 s (Nedbal and Březina 2002). Fluorescence measurements showed an oscillation with predominant frequencies of ω_0, $2\omega_0$, and $3\omega_0$. This is shown in Fig. 11.15. The authors present a feedback model, very similar to those in Sects. 10.10.6 and 11.15. A nonlinearity in the model is responsible for generating the second and third harmonics.

11.4.9 Pitfalls of Discrete Sampling: Aliasing

We saw in the preceding section that N samples in time T allow the determination of unique Fourier coefficients only for the terms from $k = 0$ to $n = (N-1)/2$. This means that for a sampling interval T/N, the maximum angular frequency is $(N-1)\omega_0/2$. The period of the highest frequency that can be determined is $T_{\min} = 2T/(N-1)$. This is approximately twice the spacing of the data points. One must sample at least twice per period to determine the coefficient at a particular frequency.

If a component is present whose frequency is more than half the sampling frequency, it will appear in the analysis at a *lower* frequency. This is the familiar stroboscopic effect in which the wheels of the stagecoach appear to rotate backward because the samples (movie frames) are not made rapidly enough. In signal analysis, this is called *aliasing*. It can be seen in Fig. 11.16, which shows a sine wave sampled at regularly spaced intervals that are longer than half a period.

This phenomenon is inescapable if frequencies greater than $(N-1)\omega_0/2$ are present. They must be removed by analog or digital techniques *before* the sampling is done. For a more detailed discussion, see Blackman and Tukey (1958)

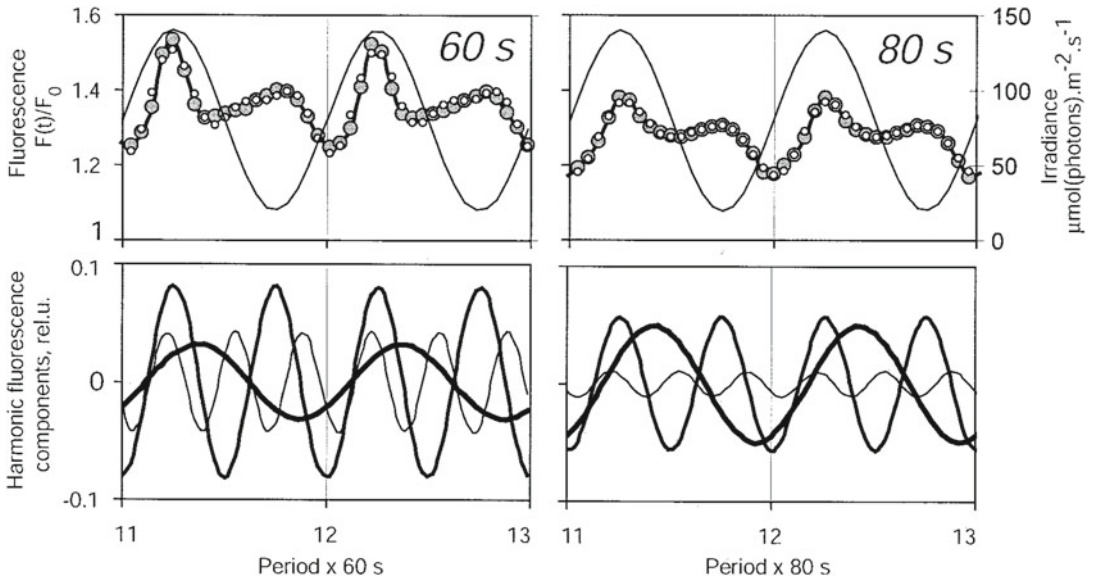


Fig. 11.15 Tobacco leaves were exposed to sinusoidally varying light with a period of 60 or 80 s (thin line, upper panels). The leaves were also interrogated with a measuring flash of orange light which stimulated fluorescence. The large circles show the resulting fluorescence. The lower panels show the frequencies in the fluorescence signal. (From Nedbal and Březina 2002. Used by permission)

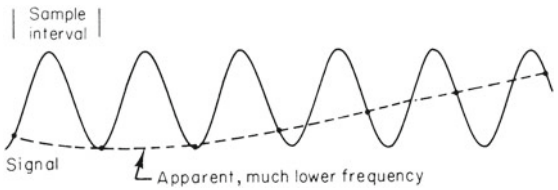


Fig. 11.16 An example of aliasing. The data are sampled less often than twice per period and appear to be at a much lower frequency

or Press et al. (1992). An example of aliasing is found in a later section, in Fig. 11.42. Maughan et al. (1973) pointed out how researchers have been "stung" by this problem in hematology.

11.4.10 Fast Fourier Transform

The calculation of the Fourier coefficients using our equations involves N evaluations of the sine or cosine, N multiplications, and N additions for each coefficient. There are N coefficients, so that there must be N^2 evaluations of the sines and cosines, which uses a lot of computer time. Cooley and Tukey (1965) showed that it is possible to group the data in such a way that the number of multiplications is about $(N/2)\log_2 N$ instead of N^2 and the sines and cosines need to be evaluated only once, a technique known as the *fast Fourier transform* (FFT). For example, for $1024 = 2^{10}$ data points, $N^2 = 1,048,576$, while $(N/2)\log_2 N = (512)(10) = 5120$. This speeds up the calculation by a factor of 204. The techniques for the FFT are discussed by many authors (see Press et al. 1992 or Visscher 1996). Bracewell (1990) has written an interesting review of all the popular numerical transforms. He points out that the grouping used in the FFT dates back to Gauss in the early nineteenth century.

11.5 Fourier Series for a Periodic Function

It is possible to define the Fourier series for a continuous periodic function $y(t)$ as well as for discrete data points y_j. In fact, the function need only be *piecewise continuous*, that is, with a finite number of discontinuities. The calculated function is given by the analog of Eq. 11.14:

$$y_{\text{calc}}(t) = a_0 + \sum_{k=1}^{n} a_k \cos(k\omega_0 t) + \sum_{k=1}^{n} b_k \sin(k\omega_0 t). \quad (11.30)$$

The quantity to be minimized is still the mean square error, in this case

$$Q = \frac{1}{T}\int_0^T \left[y(t) - y_{\text{calc}}(t)\right]^2 dt. \quad (11.31)$$

When Q is a minimum, $\partial Q/\partial a_m$ and $\partial Q/\partial b_m$ must be zero for each coefficient. For example,

$$\frac{\partial Q}{\partial a_m} = \frac{1}{T}\frac{\partial}{\partial a_m}\int_0^T \times \left(y(t) - a_0 - \sum_{k=1}^{n}[a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t)]\right)^2 dt$$
$$= -\frac{2}{T}\int_0^T \left[\left(y(t) - a_0 - \sum_{k=1}^{n}[a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t)]\right) \times \cos(m\omega_0 t)\right] dt = 0.$$

This integral must be zero for each value of m from 1 to n. If the order of integration and summation is interchanged, the result is

$$\int_0^T y(t)\cos(m\omega_0 t)\,dt - a_0\int_0^T \cos(m\omega_0 t)\,dt - \sum_{k=1}^{n} a_k \int_0^T \cos(k\omega_0 t)\cos(m\omega_0 t)\,dt - \sum_{k=1}^{n} b_k \int_0^T \sin(k\omega_0 t)\cos(m\omega_0 t)\,dt = 0. \quad (11.32)$$

The integral of $\cos(m\omega_0 t)$ over a period vanishes. The last two integrals are of the form given in Appendix E, Eqs. E.4 and E.5:

$$\int_0^T \cos(k\omega_0 t)\cos(m\omega_0 t)\,dt = \begin{cases} 0, & k \neq m \\ T/2, & k = m, \end{cases}$$
$$\int_0^T \sin(k\omega_0 t)\cos(m\omega_0 t)\,dt = 0. \quad (11.33)$$

These results are the *orthogonality relations* for the trigonometric functions. Inserting these values, we find that only one term in the first summation over k remains, and we have

$$\int_0^T y(t)\cos(m\omega_0 t)\,dt - a_m\frac{T}{2} = 0,$$

or

$$a_m = \frac{2}{T}\int_0^T y(t)\cos(m\omega_0 t)\,dt. \quad (11.34a)$$

Minimizing with respect to b_m gives

$$b_m = \frac{2}{T}\int_0^T y(t)\sin(m\omega_0 t)\,dt, \quad (11.34b)$$

and minimizing with respect to a_0 gives

$$a_0 = \frac{1}{T}\int_0^T y(t)dt. \quad (11.34c)$$

Table 11.4 Value of the kth coefficient and the value of Q when terms through the kth are included from Eq. 11.36

k	b_k	Q
1	1.2732	0.19
3	0.4244	0.10
5	0.2546	0.07
7	0.1819	0.05
9	0.1415	0.04

These equations are completely general. Because of the orthogonality of the integrals, the coefficients are independent, just as they were in the discrete case for equally spaced data. This is not surprising, since the continuous case corresponds to an infinite set of uniformly spaced data.

Note the similarity of these equations to the discrete results, Eqs. 11.25. In each case a_0 is the average of the function over the period. The other coefficients are twice the average of the signal multiplied by the sine or cosine whose coefficient is being calculated.

The integrals can be taken over any period. Sometimes it is convenient to make the interval $-T/2$ to $T/2$. As we would expect, the integrals involving sines vanish when y is an even function, and those involving cosines vanish when y is an odd function. For a continuous function, the upper limit of the sum in Eq. 11.30 can be extended from n to ∞:

$$y_{\text{calc}}(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(k\omega_0 t) + \sum_{k=1}^{\infty} b_k \sin(k\omega_0 t). \quad (11.35)$$

Lighthill (1958) proves that any piecewise continuous function converges to its Fourier series if $n = \infty$ (the *Fourier theorem*).

For the square wave $y(t) = 1$, $0 < t < T/2$; $y(t) = -1$, $T/2 < t < T$, we find

$$a_k = 0,$$
$$b_k = \begin{cases} 0, & k \text{ even}, \\ 4/\pi k, & k \text{ odd}. \end{cases} \quad (11.36)$$

Table 11.4 shows the first few coefficients for the Fourier series representing the square wave, obtained from Eq. 11.36. They are the same as those for the discrete data in Table 11.3. Figure 11.17 shows the fits for $n = 3$ and $n = 39$. As the number of terms in the fit is increased, the value of Q decreases. However, spikes of constant height (about 18 % of the amplitude of the square wave or 9 % of the discontinuity in y) remain. These are seen in Fig. 11.17. These spikes appear whenever there is a discontinuity in y and are called the *Gibbs phenomenon*.

Figure 11.18 shows the blood flow in the pulmonary artery of a dog as a function of time. It has been fitted by a mean and four terms of the form $M_k \sin(k\omega_0 t - \phi_k)$. The technique is useful because the elastic properties of the arterial wall can be described in terms of sinusoidal pressure variations at various frequencies (Milnor 1972).

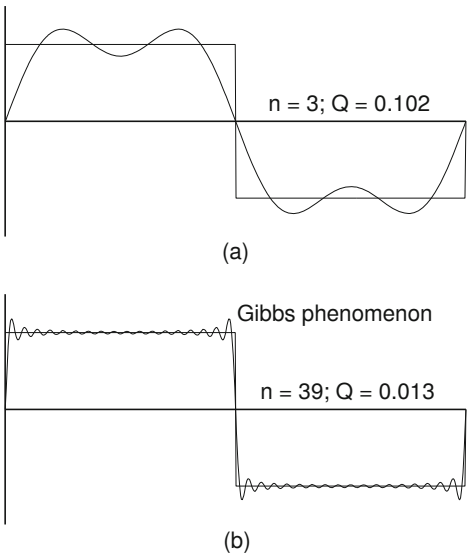


Fig. 11.17 Fit to the square wave. **a** Fit with the terms for $k = 1$ and $k = 3$. The value of Q is 0.102. **b** Fit with terms through $k = 39$. Q is very small, but the Gibbs phenomenon—*spikes* near the discontinuity—is apparent

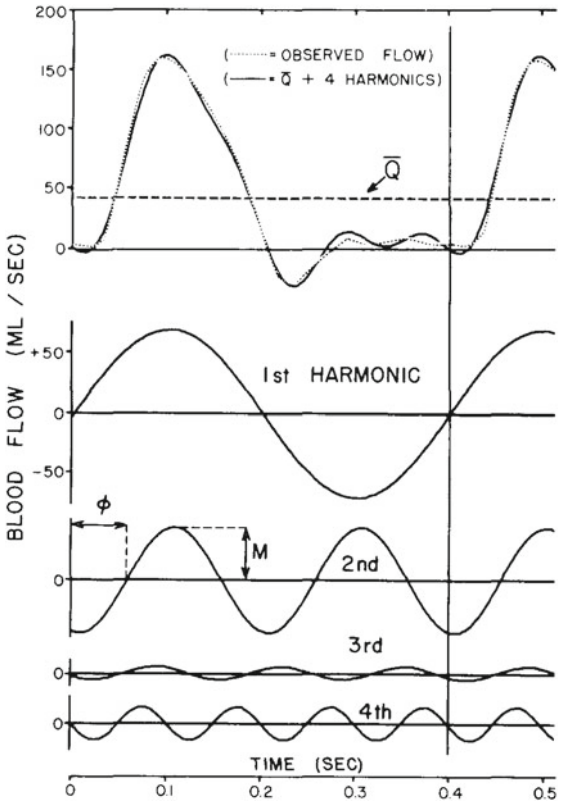


Fig. 11.18 Analysis of the pulmonary arterial blood flow in a dog, in terms of a Fourier series. (From Milnor 1972. Copyright ©Massachusetts Medical Society. All rights reserved. Drawing courtesy of Prof. Milnor)

11.6 The Power Spectrum

Since the power dissipated in a resistor is v^2/R or i^2R, the square of any function (or signal) is often called the power. A periodic signal $y(t)$ can be written as

$$y(t) = a_0 + \sum_{k=1}^{\infty} [a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t)]. \quad (11.37)$$

The coefficients are given by Eqs. 11.34.

The average "power" in the signal is defined to be[4]

$$\langle y^2 \rangle = \lim_{T' \to \infty} \frac{1}{2T'} \int_{-T'}^{T'} y^2(t)\, dt. \quad (11.38)$$

For a periodic signal, the same result can be obtained by integrating over one period:

$$\langle y^2 \rangle = \frac{1}{T} \int_0^T y^2(t)\, dt. \quad (11.39)$$

To calculate this using Eq. 11.37 for $y(t)$, we have to write the sum twice and multiply both sums together:

$$\frac{1}{T}\int_0^T y^2(t)\, dt = \frac{1}{T}\int_0^T \times \left(a_0 + \sum_{k=1}^{n} [a_k \cos(k\omega_0 t) + b_k \sin(k\omega_0 t)] \right) \times \left(a_0 + \sum_{j=1}^{n} [a_j \cos(j\omega_0 t) + b_j \sin(j\omega_0 t)] \right) dt.$$

When these terms are multiplied together and written out, we have

$$\langle y^2 \rangle = \frac{1}{T}\int_0^T dt \left(\overset{\text{(i)}}{a_0^2} + 2a_0 \sum_{k=1}^{\infty} \left[\overset{\text{(ii)}}{a_k \cos(k\omega_0 t)} + \overset{\text{(iii)}}{b_k \sin(k\omega_0 t)} \right] + \sum_{k=1}^{\infty} \left[\overset{\text{(iv)}}{a_k^2 \cos^2(k\omega_0 t)} + \overset{\text{(v)}}{b_k^2 \sin^2(k\omega_0 t)} \right] + \sum_{k=1}^{\infty} \sum_{j \neq k} \overset{\text{(vi)}}{a_k a_j \cos(k\omega_0 t) \cos(j\omega_0 t)} + \sum_{k=1}^{\infty} \sum_{j \neq k} \overset{\text{(vii)}}{b_k b_j \sin(k\omega_0 t) \sin(j\omega_0 t)} + 2 \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \overset{\text{(viii)}}{a_k b_j \cos(k\omega_0 t) \sin(j\omega_0 t)} \right).$$

[4] The time average of a variable will be denoted by ⟨⟩ brackets.

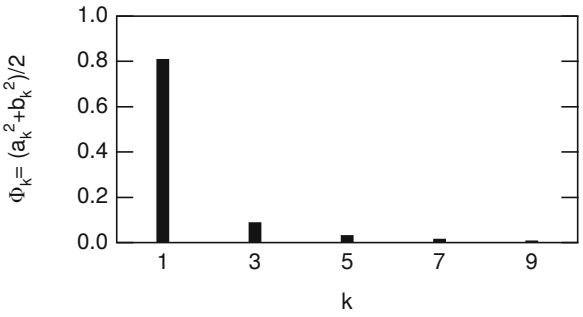


Fig. 11.19 The power spectrum Φ_k for the square wave of Fig. 11.7 or Fig. 11.18, calculated using the values of b_k from Table 11.4

Each term has been labeled (i) through (viii). Assume that the function y is sufficiently well behaved so that the order of integration and summation can be interchanged. Term (i) gives a_0^2. Terms (ii) and (iii) are integrals of the cosine or sine over an integral number of cycles and vanish. Terms (iv) and (v) give $a_k^2/2$ and $b_k^2/2$. Terms (vi), (vii), and (viii) all vanish because of Eq. 11.33. We finally have for the average power

$$\langle y^2(t) \rangle = a_0^2 + \frac{1}{2}\sum_{k=1}^{\infty}(a_k^2 + b_k^2) = \sum_{k=0}^{\infty} \Phi_k. \quad (11.40)$$

The coefficients are defined by Eqs. 11.34. We could have made a similar argument for the discrete Fourier series of Eqs. 11.25 or 11.26 and obtained the same result. In both cases the average power is a sum of terms Φ_k that represent the average power at each frequency $k\omega_0$. The term $\Phi_0 = a_0^2$ is the average of the square of the zero-frequency or dc (direct-current) term and $\Phi_k = (a_k^2 + b_k^2)/2$ is the average of the squares of the terms $a_k \cos(k\omega_0 t)$ and $b_k \sin(k\omega_0 t)$. Figure 11.19 shows the power spectrum of the square wave that was used in the example.

11.7 Correlation Functions

The correlation function is useful to test whether two functions of time are correlated, that is, whether a change in one is accompanied by a change in the other. Let the two variables to be tested be $y_1(t)$ and $y_2(t)$. The change in y_2 may occur earlier or later than the change in y_1; therefore the correlation must be examined as one of the variables is shifted in time. Examples of pairs of variables that may be correlated are wheat price and rainfall, urinary output and fluid intake, and the voltage changes at two different points along the nerve axon. The variables may or may not be periodic. Exhibiting a correlation does not establish a cause-and-effect relationship. (The height of a growing tree may correlate for several years with an increase in the stock market.)

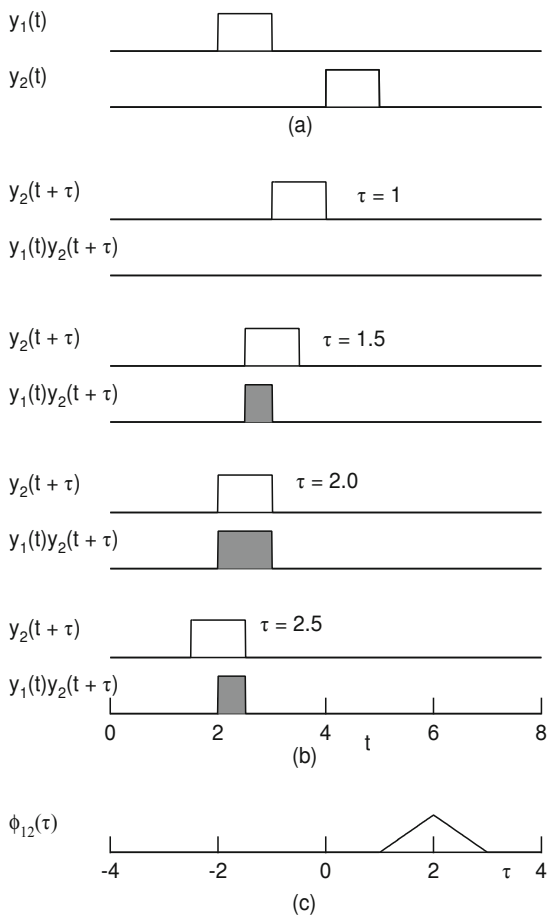


Fig. 11.20 An example of the cross-correlation function. **a** The two signals to be correlated. **b** Plots of $y_2(t+\tau)$ and the product $y_1(t)y_2(t+\tau)$ for different values of τ. **c** Plot of $\phi_{12}(\tau)$. The peak occurs when signal y_2 has been advanced 2 s

11.7.1 Cross-Correlation of a Pulse

To calculate the cross-correlation function of y_1 and y_2, advance $y_2(t)$ by an amount τ, multiply y_1 by the shifted y_2, and integrate the product. Figure 11.20 shows the process for two square pulses one-second long. The second pulse occurs 2 s later than the first. As the second pulse is advanced the pulses begin to overlap. When the second pulse has been advanced by 2 s the overlap is greatest; as it is advanced more, the overlap falls to zero. The cross-correlation function depends on τ and is plotted in Fig. 11.20c. The mathematical statement of this procedure for a pulse is

$$\phi_{12}(\tau) = \int_{-\infty}^{\infty} y_1(t)y_2(t+\tau)\,dt. \tag{11.41}$$

The integrand makes a positive contribution to the integral if $y_1(t)$ and $y_2(t+\tau)$ are both positive at the same time or both negative at the same time. It makes a negative contribution if one function is positive while the other is negative.

11.7.2 Cross-Correlation of a Nonpulse Signal

If the signals are not pulses, then the cross-correlation integral is defined as

$$\phi_{12}(\tau) = \langle y_1(t)y_2(t+\tau)\rangle\,. \tag{11.42}$$

As before, the average is the integral over a long time divided by the time interval:

$$\phi_{12}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y_1(t)y_2(t+\tau)\,dt. \tag{11.43}$$

If the signals have period T, the average can be taken by integrating over a single period:

$$\phi_{12}(\tau) = \frac{1}{T}\int_{t'}^{t'+T} y_1(t)y_2(t+\tau)\,dt. \tag{11.44}$$

Note the difference in units between ϕ_{12} as defined for pulses in Eq. 11.41 where the units of ϕ are the units of y^2 times time, and ϕ_{12} defined in Eqs. 11.42–11.44 where the units are those of y^2.

The cross-correlation depends only on the relative shift of the two signals. It does not matter whether y_2 is advanced by an amount τ or y_1 is delayed by the same amount:

$$\phi_{21}(-\tau) = \phi_{12}(\tau). \tag{11.45}$$

11.7.3 Cross-Correlation Example

As an example of the cross-correlation, consider a square wave that has value ± 1 and a sine wave with the same period (Fig. 11.21). When the square wave and sine wave are in phase, the product is always positive and the cross-correlation has its maximum value. As the square wave is shifted the product is sometimes positive and sometimes negative. When they are 1/4 period out of phase, the average of the integrand is zero, as shown in Fig. 11.21b. Still more shift results in the correlation function becoming negative, then positive again, with a shift of one full period giving the same result as no shift.

11.7.4 Autocorrelation

The *autocorrelation* function is the correlation of the signal with itself:

$$\phi_{11}(\tau) = \int y_1(t)y_1(t+\tau)\,dt \qquad \text{(pulse)}, \tag{11.46}$$

$$\phi_{11}(\tau) = \langle y_1(t)y_1(t+\tau)\rangle \qquad \text{(nonpulse)}. \tag{11.47}$$

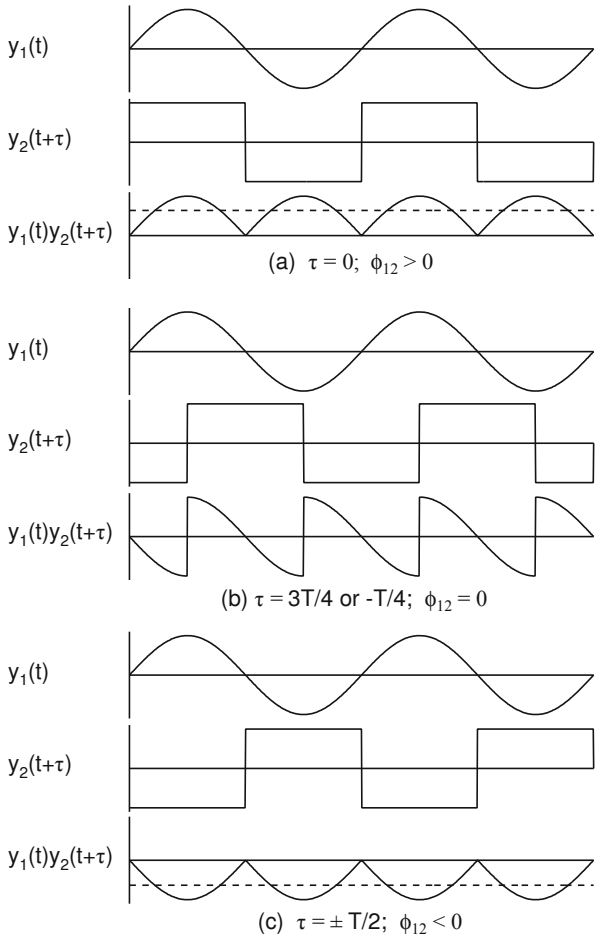


Fig. 11.21 Cross-correlation of a square wave and a sine wave of the same period

Since the signal is correlated with itself, advancing one copy of the signal is the same as delaying the other. The autocorrelation is an even function of τ:

$$\phi_{11}(\tau) = \phi_{11}(-\tau). \tag{11.48}$$

11.7.5 Autocorrelation Examples

The autocorrelation function for a sine wave can be calculated analytically. If the amplitude of the sine wave is A and the frequency is $\omega = 2\pi/T$,

$$\phi_{11}(\tau) = \frac{A^2}{T}\int_0^T \sin(\omega t)\sin(\omega t + \omega\tau)\,dt$$

$$= \frac{A^2}{T}\int_0^T \sin(\omega t) \times [\sin(\omega t)\cos(\omega\tau) + \cos(\omega t)\sin(\omega\tau)]\,dt$$

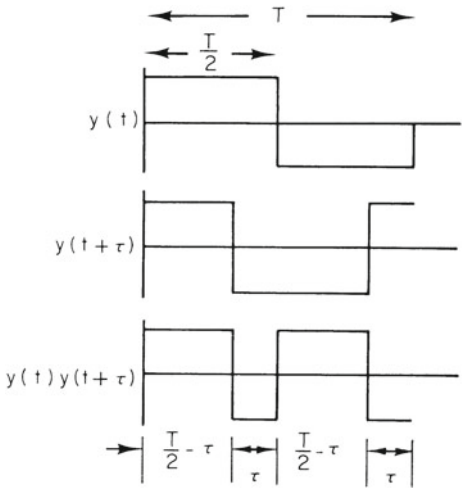


Fig. 11.22 Plots of $y(t)$, $y(t+\tau)$, and their product for a square wave

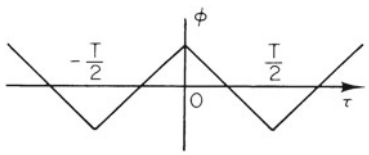


Fig. 11.23 Plot of $\phi_{11}(\tau)$ for the square wave

$$= A^2\cos(\omega\tau)\left[\frac{1}{T}\int_0^T \sin^2(\omega t)\,dt\right] + A^2\sin(\omega\tau)\left[\frac{1}{T}\int_0^T \sin(\omega t)\cos(\omega t)\,dt\right].$$

It is shown in Appendix E that the first term in square brackets is $\frac{1}{2}$ and the second is 0. Therefore the autocorrelation function of the sine wave is

$$\phi_{11}(\tau) = \frac{A^2}{2}\cos(\omega\tau). \tag{11.49}$$

As a final example, consider the autocorrelation of a square wave of unit amplitude. One period is drawn in Fig. 11.22 showing the wave, the advanced wave, and the product. The average is the net area divided by T. The area above the axis is $(2)(T/2-\tau)(1)$ since there are two rectangles of height 1 and width $T/2-\tau$. From this must be subtracted the area of the two rectangles of height 1 and width τ that are below the axis. The net area is $T-4\tau$. The autocorrelation function is

$$\phi_{11}(\tau) = 1 - 4\tau/T, \quad 0 < \tau < T/2. \tag{11.50}$$

The plot of the integrand in Fig. 11.22 is only valid for $0 < \tau < T/2$. We can use the fact that the autocorrelation is an even function to draw ϕ for $-T/2 < \tau < 0$. We then have ϕ for the whole period. It is plotted in Fig. 11.23.

11.8 The Autocorrelation Function and the Power Spectrum

We saw that the power spectrum of a periodic signal is related to the coefficients in its Fourier series (Eq. 11.40):

$$\left\langle y^2(t)\right\rangle = a_0^2 + \frac{1}{2}\sum_{k=1}^{\infty}(a_k^2 + b_k^2),$$

with the term for each value of k representing the amount of power carried in the signal component at that frequency. The Fourier series for the autocorrelation function carries the same information. To see this, calculate the autocorrelation function of

$$y_1(t) = a_0 + \sum_{k=1}^{\infty}\left[a_k\cos(k\omega_0 t) + b_k\sin(k\omega_0 t)\right].$$

We can write

$$\begin{aligned}\phi_{11}(\tau) &= \langle y_1(t)y_1(t+\tau)\rangle \\ &= \left\langle\left(a_0 + \sum_{k=1}^{\infty}\left[a_k\cos(k\omega_0 t) + b_k\sin(k\omega_0 t)\right]\right)\right. \\ &\left.\times\left(a_0 + \sum_{j=1}^{\infty}\left\{a_j\cos\left[j\omega_0(t+\tau)\right] + b_j\sin\left[j\omega_0(t+\tau)\right]\right\}\right)\right\rangle.\end{aligned}$$

The next step is to multiply out all the terms as we did when deriving Eq. 11.40. We then use the trigonometric identities[5]

$$\begin{aligned}\cos(x+y) &= \cos x\cos y - \sin x\sin y,\\ \sin(x+y) &= \cos x\sin y + \sin x\cos y.\end{aligned}$$

For many of the terms, either the averages are zero or pairs of terms cancel. We finally obtain

$$\phi_{11}(\tau) = a_0^2 + \sum_{k=1}^{\infty}\frac{1}{2}(a_k^2 + b_k^2)\cos(k\omega_0\tau). \qquad (11.51)$$

This has only cosine terms, since the autocorrelation function is even.

For zero shift,

$$\phi_{11}(0) = a_0^2 + \sum_{k=1}^{\infty}\frac{1}{2}(a_k^2 + b_k^2).$$

Comparison with Eq. 11.40 shows that this is the power in the signal. We can get this result directly from Eq. 11.38. The integral is the same as the definition of the autocorrelation function when $\tau = 0$.

[5] One virtue of the complex notation is that these addition formulae become the standard rule for multiplying exponentials: $e^{i(x+y)} = e^{ix}e^{iy}$.

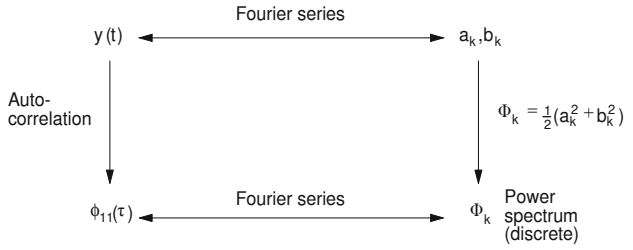


Fig. 11.24 The power spectrum of a periodic signal can be obtained either from the squares of the Fourier coefficients of the signal or from the Fourier coefficients of the autocorrelation function

The Fourier series for the autocorrelation function is particularly easy to obtain. We need only pick out the coefficients in Eq. 11.51. Write the Fourier expansion of the autocorrelation function as

$$\phi_{11}(\tau) = \alpha_0 + \sum_{k=1}^{\infty}\alpha_k\cos(k\omega_0\tau). \qquad (11.52)$$

Comparing terms in Eqs. 11.51 and 11.52 shows that $\alpha_0 = a_0^2$ and $\alpha_k = (a_k^2 + b_k^2)/2$. We can also compare these with the definition of Φ_k in Eq. 11.40 and say that

$$\begin{aligned}\Phi_0 = &\quad \text{average dc (zero-frequency) power} &&= \alpha_0 = a_0^2,\\ \Phi_k = &\quad \text{average power at frequency } k\omega_0 &&= \alpha_k = \tfrac{1}{2}(a_k^2 + b_k^2).\end{aligned} \qquad (11.53)$$

The autocorrelation function contains no phase information about the signal. The sine and cosine terms at a given frequency are completely mixed together.

There are two ways to find the power Φ_k at frequency $k\omega_0$. Both are shown in Fig. 11.24. The function $y(t)$ and its Fourier coefficients are completely equivalent, and one can go from one to the other. Squaring the coefficients and adding them gives the power spectrum. This is a one-way process; once they have been squared and added, there is no way to separate them again. The autocorrelation function also involves squaring and adding and is a one-way process. The autocorrelation function and the power spectrum are related by a Fourier series and can be calculated from each other.

11.9 Nonperiodic Signals and Fourier Integrals

Sometimes we have to deal with a signal that is a pulse that occurs just once. Several pulses are shown in Fig. 11.25; they come in an infinite variety of shapes. Noise is another signal that never repeats itself and is therefore not periodic. The *Fourier integral* or *Fourier transform* is an extension of the

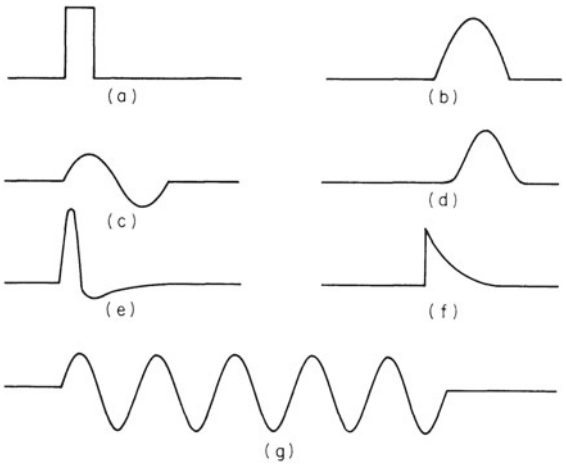


Fig. 11.25 Various pulses. The common feature is that they occur once. **a** Square pulse. **b** Half cycle of a sine wave. **c** One cycle of a sine wave. **d** Gaussian. **e** Nerve pulse. **f** Exponentially decaying pulse. **g** Gated sine wave

Fourier series that allows us to deal with nonperiodic signals (Bracewell 2000).

11.9.1 Introduce Negative Frequencies and Make the Coefficients Half as Large

The Fourier series expansion of a periodic function $y(t)$ was seen in Eq. 11.35 with the coefficients given by Eqs. 11.34. Since $y(t)$ has period T, the integrals in Eqs. 11.34 can be over any interval that is one period long. Let us therefore make the limits of integration $-T/2$ to $T/2$ and also remember that $1/T = \omega_0/2\pi$. With these substitutions, Eqs. 11.34 become

$$a_0 = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\,dt,$$

$$a_k = \frac{\omega_0}{\pi}\int_{-T/2}^{T/2} y(t)\cos(k\omega_0 t)\,dt, \qquad (11.54)$$

$$b_k = \frac{\omega_0}{\pi}\int_{-T/2}^{T/2} y(t)\sin(k\omega_0 t)\,dt.$$

Now allow k to have negative as well as positive values. If the coefficients for negative k are also defined by Eqs. 11.54, they have the properties [since $\cos(k\omega_0 t) = \cos(-k\omega_0 t)$ and $\sin(k\omega_0 t) = -\sin(-k\omega_0 t)$],

$$a_{-k} = a_k, \qquad b_{-k} = -b_k.$$

Therefore, the terms $a_k\cos(k\omega_0 t)$ and $b_k\sin(k\omega_0 t)$ in Eq. 11.30 are the same function of t whether k is positive or negative. By introducing negative values of k we can make the coefficients in front of the integrals for a_k and b_k in Eqs. 11.54 become $\omega_0/2\pi$. This is the same trick used to obtain the discrete equations, Eqs. 11.27. With negative values of k allowed, we have

$$y(t) = a_0' + \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}\left[a_k'\cos(k\omega_0 t) + b_k'\sin(k\omega_0 t)\right],$$

$$a_0' = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\,dt,$$

$$a_k' = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\cos(k\omega_0 t)\,dt,$$

$$b_k' = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\sin(k\omega_0 t)\,dt.$$

Since $\cos(0) = 1$ and $\sin(0) = 0$, we can incorporate the definition of a_0' into the definition of a_k' and introduce b_0' which is always zero. The sum can then include $k = 0$:

$$y(t) = \sum_{k=-\infty}^{\infty}\left[a_k'\cos(k\omega_0 t) + b_k'\sin(k\omega_0 t)\right],$$

$$a_k' = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\cos(k\omega_0 t)\,dt, \qquad (11.55)$$

$$b_k' = \frac{\omega_0}{2\pi}\int_{-T/2}^{T/2} y(t)\sin(k\omega_0 t)\,dt.$$

A final change of variables defines $C_k = 2\pi a_k'/\omega_0$ and $S_k = 2\pi b_k'/\omega_0$. With these changes the Fourier series and its coefficients are

$$y(t) = \frac{\omega_0}{2\pi}\sum_{k=-\infty}^{\infty}\left[C_k\cos(k\omega_0 t) + S_k\sin(k\omega_0 t)\right],$$

$$C_k = \int_{-T/2}^{T/2} y(t)\cos(k\omega_0 t)\,dt, \qquad (11.56)$$

$$S_k = \int_{-T/2}^{T/2} y(t)\sin(k\omega_0 t)\,dt.$$

To recapitulate, there is nothing fundamentally new in Eqs. 11.56. Negative values of k were introduced so that the sum goes over each value of $|k|$ twice (except for $k = 0$). This allowed the coefficients to be made half as large.

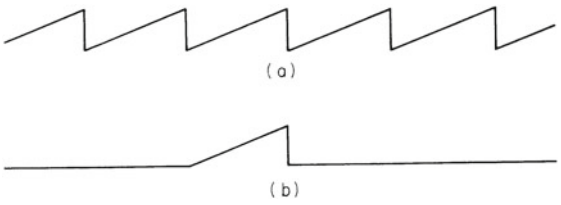


Fig. 11.26 **a** A periodic signal. **b** A nonperiodic signal

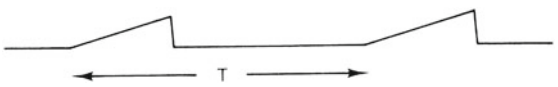


Fig. 11.27 An approximation to the nonperiodic signal shown in Fig. 11.26b

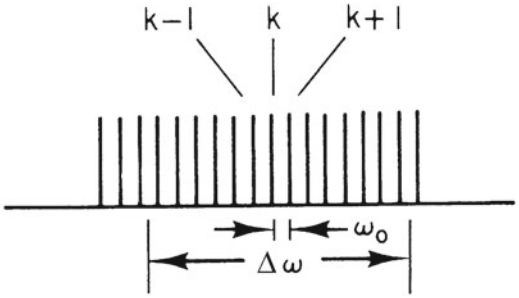


Fig. 11.28 A histogram of C_k vs. k

11.9.2 Make the Period Infinite

These equations can be used to calculate the Fourier series for a periodic signal such as that shown in Fig. 11.26a. Suppose that instead we want to find the coefficients for the nonperiodic signal shown in Fig. 11.26b. This signal can be approximated by another periodic signal shown in Fig. 11.27. The approximation to Fig. 11.26b becomes better and better as T is made longer. As T becomes infinite, the fundamental angular frequency ω_0 approaches 0. Define $\omega = k\omega_0$. The frequencies ω are discrete with spacing ω_0. Consider a small frequency interval encompassing many values of k, as shown in Fig. 11.28. Since ω_0 is approaching zero, there can be many values of ω and k between ω and $\omega + \Delta\omega$. The frequencies will be nearly the same, so the values of C_k will be nearly the same. All of the terms in the sum in Eq. 11.56 can be replaced by an average value of C_k or S_k multiplied by the number of values of k in the interval, which is $\Delta\omega/\omega_0$. Finally, we set C_k=$C(\omega)$ and $\Delta\omega = d\omega$. The sum becomes an integral with $\Delta\omega = d\omega$:

$$y(t) = \frac{\omega_0}{2\pi}\int_{-\infty}^{\infty} [C(\omega)\cos\omega t + S(\omega)\sin\omega t]\,\frac{d\omega}{\omega_0},$$

or finally, since $d\omega = 2\pi df$,

$$\begin{aligned} y(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} [C(\omega)\cos\omega t + S(\omega)\sin\omega t]\; d\omega \\ &= \int_{-\infty}^{\infty} \left[C(f)\cos(2\pi f t) + S(f)\sin(2\pi f t)\right] df, \\ C(\omega) &= \int_{-\infty}^{\infty} y(t)\cos\omega t\, dt, \\ S(\omega) &= \int_{-\infty}^{\infty} y(t)\sin\omega t\, dt. \end{aligned} \tag{11.57}$$

These equations constitute a *Fourier integral pair* or *Fourier transform pair*. They are completely symmetric in the variables f and t and symmetric apart from the factor 2π in the variables ω and t. We obtain $C(\omega)$ or $S(\omega)$ by multiplying the function $y(t)$ by the appropriate trigonometric function and integrating over time. We obtain $y(t)$ by multiplying C and S by the appropriate trigonometric function and integrating over frequency.

11.9.3 Complex Notation

Using complex notation, we define

$$Y(\omega) = C(\omega) - iS(\omega) \tag{11.58}$$

and write

$$\begin{aligned} y(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} Y(\omega)e^{i\omega t}\, d\omega = \int_{-\infty}^{\infty} Y(\omega)e^{i\omega t}\, df, \\ Y(\omega) &= \int_{-\infty}^{\infty} y(t)e^{-i\omega t}\, dt. \end{aligned} \tag{11.59}$$

11.9.4 Example: The Exponential Pulse

As an example, consider the function

$$y(t) = \begin{cases} 0, & t \le 0 \\ Ae^{-at}, & t > 0. \end{cases} \tag{11.60}$$

The functions C and S are evaluated using Eqs. 11.57. Since $y(t)$ is zero for negative times, the integrals extend from zero to infinity. They are found in all standard integral tables:

$$\begin{aligned} C(\omega) &= A\int_0^{\infty} e^{-at}\cos\omega t\, dt = \frac{A/a}{1+(\omega/a)^2}, \\ S(\omega) &= A\int_0^{\infty} e^{-at}\sin\omega t\, dt = \frac{(A/a)(\omega/a)}{1+(\omega/a)^2}. \end{aligned} \tag{11.61}$$

These are plotted in Fig. 11.29. Function C is even, while S is odd. The functions are plotted on log–log graph paper in

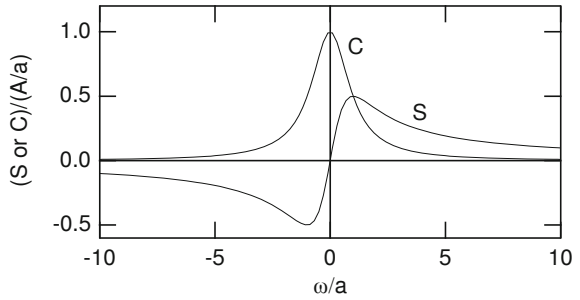


Fig. 11.29 The sine and cosine coefficients in the Fourier transform of an exponentially decaying pulse

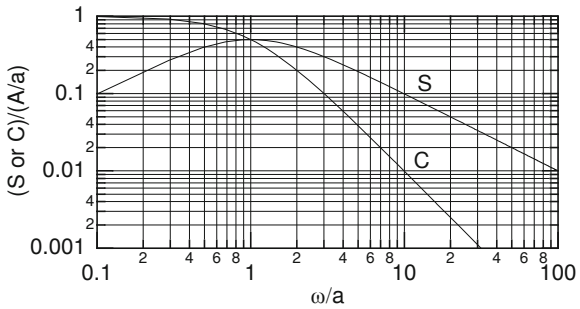


Fig. 11.30 Log–log plot of the coefficients in Fig. 11.29

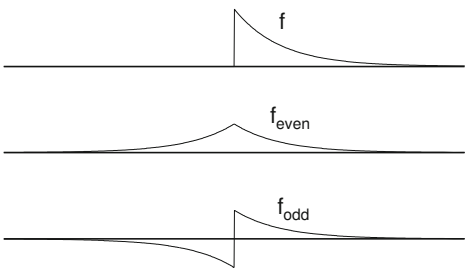


Fig. 11.31 Function $f(t)$ and its even and odd parts

Fig. 11.30. Remember that only positive values of ω/a can be shown on a logarithmic scale, so the origin and negative frequencies cannot be shown. It is apparent from the slopes of the curves that C falls off as $(\omega/a)^{-2}$ while S falls more slowly, as $(\omega/a)^{-1}$. One way of explaining this difference is to note that the function $y(t)$ can be written as a sum of even and odd parts as shown in Fig. 11.31. The odd function, which is given by the sine terms in the integral, has a discontinuity, while the even function does not. A more detailed study of Fourier expansions shows that a function with a discontinuity has coefficients that decrease as $1/\omega$ or $1/k$, while the coefficients of a function without a discontinuity decrease more rapidly. (Recall that the coefficients of the square wave were $4/\pi k$.)

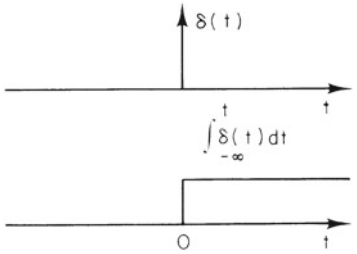


Fig. 11.32 The δ function and its integral

11.10 The Delta Function

It will be useful in the next sections to introduce a pulse that is very narrow, very tall, and has unit area under its curve. Physicists call this function the *delta function* $\delta(t)$. Engineers call it the *impulse function* $u_0(t)$.

The δ function is defined by the equations

$$\delta(t) = 0, \quad t \neq 0$$
$$\int_{-\epsilon}^{\epsilon} \delta(t)\,dt = \int_{-\infty}^{\infty} \delta(t)\,dt = 1. \tag{11.62}$$

The δ function can be thought of as a rectangle of width a and height $1/a$ in the limit $a \to 0$, or as a Gaussian function (Appendix I) as $\sigma \to 0$. Many other functions have the same limiting properties. The δ function is not like the usual function in mathematics because of its infinite discontinuity at the origin. It is one of a class of *generalized functions* whose properties have been rigorously developed by mathematicians[6] since they were first used by the physicist P. A. M. Dirac.

Since integrating across the origin picks up this spike of unit area, the integral of the δ function is a step of unit height at the origin. The δ function and its integral are shown in Fig. 11.32. The δ function can be positioned at $t = a$ by writing $\delta(t - a)$ because the argument vanishes at $t = a$.

Multiplying any function by the δ function and integrating picks out the value of the function when the argument of the δ function is zero:

$$\int_{-\infty}^{\infty} y(t)\delta(t)\,dt = y(0)\int_{-\infty}^{\infty} \delta(t)\,dt = y(0),$$
$$\int_{-\infty}^{\infty} y(t)\delta(t-a)\,dt = y(a)\int_{-\infty}^{\infty} \delta(t-a)\,dt = y(a). \tag{11.63}$$

The second integral on each line is based on the fact that $y(t)$ has a constant value when the δ function is nonzero so it can be taken outside the integral.

[6] A rigorous but relatively elementary mathematical treatment is given by Lighthill (1958).

The δ function has the following properties that are proved in Problem 34:

$$\begin{aligned} \delta(t) &= \delta(-t), \\ t\,\delta(t) &= 0 \\ \delta(at) &= \frac{1}{a}\delta(t). \end{aligned} \tag{11.64}$$

11.11 The Energy Spectrum of a Pulse and Parseval's Theorem

For a signal with period T, the *average power* is $\frac{1}{T}\int_0^T y^2(t)dt$. We can also define the average power for a signal lasting a very long time as

$$\lim_{T'\to\infty}\frac{1}{2T'}\int_{-T'}^{T'} y^2(t)dt = a_0^2 + \frac{1}{2}\sum_k (a_k^2 + b_k^2).$$

In the limit of infinite duration, both the integral and the denominator are infinite, but the ratio is finite.

For a pulse the integral is finite and the average power vanishes. In that case, we use the integral without dividing by T and call it the *energy* in the pulse.

Since $y(t)$ is given by a Fourier integral, the energy in the pulse can be written as

$$\int_{-\infty}^{\infty} y^2(t)\,dt = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega'$$
$$[C(\omega)\cos\omega t + S(\omega)\sin\omega t]\times\left[C(\omega')\cos\omega' t + S(\omega')\sin\omega' t\right].$$

If the terms are multiplied out, this becomes

$$\begin{aligned} \int_{-\infty}^{\infty} y^2(t)\,dt = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \\ \big[C(\omega)C(\omega')\cos\omega t\cos\omega' t \\ + C(\omega)S(\omega')\cos\omega t\sin\omega' t \\ + S(\omega)C(\omega')\sin\omega t\cos\omega' t \\ + S(\omega)S(\omega')\sin\omega t\sin\omega' t\big]. \end{aligned} \tag{11.65}$$

To simplify this expression, we interchange the order of integration, carrying out the time integration first. We assume that the function $y(t)$ is sufficiently well behaved so that this can be done.

Changing the order gives three integrals to consider:

$$\int_{-\infty}^{\infty} \cos\omega t\ \cos\omega' t\,dt,$$
$$\int_{-\infty}^{\infty} \cos\omega t\ \sin\omega' t\,dt,$$
$$\int_{-\infty}^{\infty} \sin\omega t\ \sin\omega' t\,dt.$$

These are analogous to the trigonometric integrals of Appendix E, except that they extend over all time instead of just one period. Therefore, we might expect that an integral such as $\int_{-\infty}^{\infty}\cos\omega t\,\sin\omega' t\,dt$ would vanish for all possible values of ω and ω'. We might expect that the integrals $\int_{-\infty}^{\infty}\cos\omega t\,\cos\omega' t\,dt$ and $\int_{-\infty}^{\infty}\sin\omega t\,\sin\omega' t\,dt$ would vanish when $\omega\neq\omega'$ and be infinite when $\omega=\omega'$. Such is indeed the case. This is reminiscent of the δ function, but it does not tell us the exact relationship of these integrals to it.

To find the exact values of the integrals we use the following trick. Let $y(t)$ be the function for which $C(\omega) = \delta(\omega-\omega')$. Then, using Eqs. 11.57 and 11.63 we get

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\delta(\omega-\omega')\cos\omega t\,d\omega = \frac{1}{2\pi}\cos\omega' t.$$

The inverse equation for $C(\omega)$ is

$$C(\omega) = \int_{-\infty}^{\infty} y(t)\cos\omega t\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\cos\omega' t\cos\omega t\,dt.$$

But $C(\omega)=\delta(\omega-\omega')$. Therefore

$$\int_{-\infty}^{\infty}\cos\omega t\cos\omega' t\,dt = 2\pi\,\delta(\omega-\omega'). \tag{11.66a}$$

A similar argument shows that

$$\int_{-\infty}^{\infty}\sin(\omega t)\sin(\omega' t)\,dt = 2\pi\,\delta(\omega-\omega'). \tag{11.66b}$$

The fact that both the sine and cosine integrals are the same should not be surprising, since a sine curve is jut a cosine curve shifted along the axis and we are integrating from $-\infty$ to ∞.

11.11.1 Parseval's Theorem

The integrals in Eqs. 11.66 can be used to evaluate Eq. 11.65. The result is

$$\int_{-\infty}^{\infty} y^2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega\int_{-\infty}^{\infty}d\omega'[C(\omega)C(\omega')\delta(\omega-\omega') + S(\omega)S(\omega')\delta(\omega-w')]$$

$$\int_{-\infty}^{\infty} y^2(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega\left[C^2(\omega)+S^2(\omega)\right]. \tag{11.67}$$

This result is known as *Parseval's theorem*. If we define the function

$$\Phi'(\omega) = C^2(\omega) + S^2(\omega), \tag{11.68a}$$

then Eq. 11.67 takes the form

$$\int_{-\infty}^{\infty} y^2(t)\,dt = \int_{-\infty}^{\infty}\Phi'(\omega)\frac{d\omega}{2\pi} = \int_{-\infty}^{\infty}\Phi'(f)df. \tag{11.68b}$$

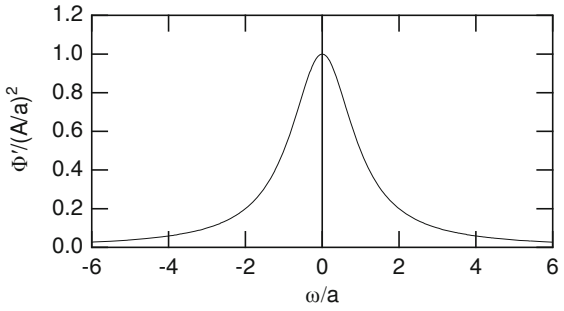


Fig. 11.33 The energy spectrum $\Phi'(\omega)$ for an exponential pulse

The prime is to remind us that this is energy and not power. The left-hand side is the total energy in the signal, and $y^2(t)\,dt$ is the amount of energy between t and $t + dt$. This suggests that we call $\Phi'(\omega)\,d\omega/2\pi = \Phi'(f)\,df$ the amount of energy in the angular frequency interval between ω and $\omega + d\omega$ or the frequency interval between f and $f + df$.

11.11.2 Example: The Exponential Pulse

The energy spectrum of the exponential pulse that was used earlier as an example is

$$\Phi'(\omega) = C^2(\omega) + S^2(\omega) = \left(\frac{A}{a}\right)^2 \frac{1}{1 + (\omega/a)^2}. \quad (11.69)$$

It is plotted in Fig. 11.33.

11.12 The Autocorrelation of a Pulse and its Relation to the Energy Spectrum

The correlation functions for pulses are defined as integrals instead of averages:

$$\begin{aligned} \phi_{12}(\tau) &= \int_{-\infty}^{\infty} y_1(t) y_2(t+\tau)\,dt, \\ \phi_{11}(\tau) &= \int_{-\infty}^{\infty} y_1(t) y_1(t+\tau)\,dt. \end{aligned} \quad (11.70)$$

Consider the autocorrelation function of the exponential pulse, Eq. 11.60. Figure 11.34 shows the functions involved in calculating the autocorrelation for a typical positive value of τ. Since the autocorrelation function is even, negative values of τ need not be considered. The product of the function and the shifted function is $(Ae^{-at})(Ae^{-a(t+\tau)}) = A^2 e^{-a\tau} e^{-2at}$. It can be seen from Fig. 11.34 that the limits of integration are from zero to infinity. Thus,

$$\phi_{11}(\tau) = A^2 e^{-a\tau} \int_0^{\infty} e^{-2at}\,dt = \frac{A^2 e^{-a\tau}}{2a}, \quad \tau > 0.$$

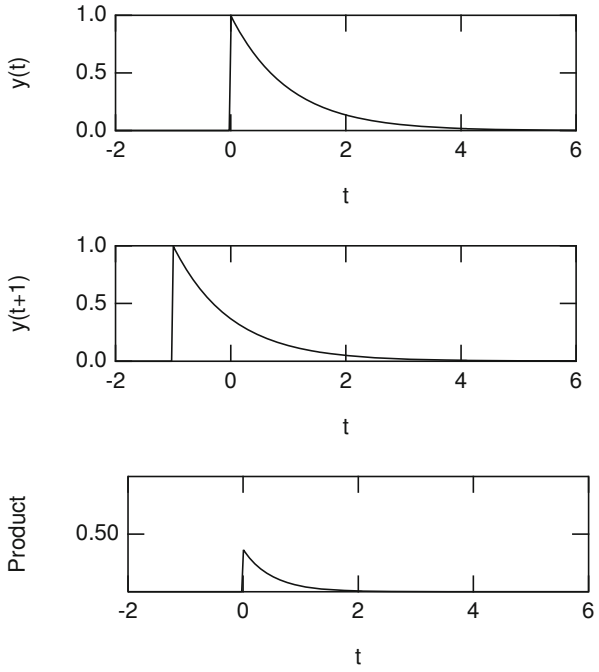


Fig. 11.34 Calculation of the autocorrelation of the exponential pulse. The figure shows $y(t)$, $y(t + \tau)$, and their product, for $\tau = 1$

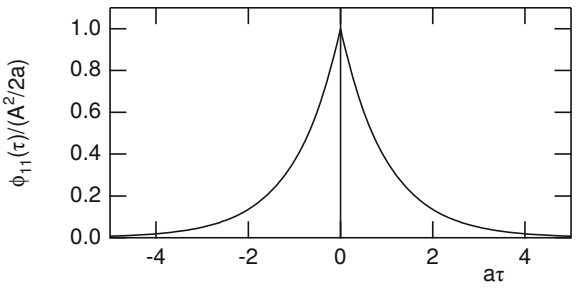


Fig. 11.35 The autocorrelation function for an exponentially decaying pulse

Because ϕ_{11} is even, the full autocorrelation function is

$$\phi_{11}(\tau) = \frac{A^2}{2a} e^{-a|\tau|}. \quad (11.71)$$

This is plotted in Fig. 11.35.

The autocorrelation function has a Fourier transform Φ'. Only the cosine term appears, since ϕ_{11} is even:

$$\begin{aligned} \Phi'(\omega) &= \frac{A^2}{2a} \int_{-\infty}^{\infty} e^{-a|t|} \cos \omega t\,dt \\ &= \frac{A^2}{a} \int_0^{\infty} e^{-at} \cos \omega t\,dt. \end{aligned}$$

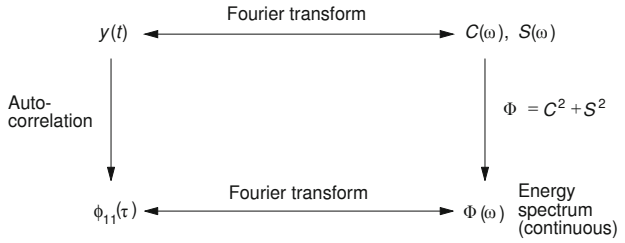


Fig. 11.36 Two ways to obtain the energy spectrum of a pulse signal

We have seen this integral before, in conjunction with Eq. 11.61. The result is

$$\Phi'(\omega) = \left(\frac{A}{a}\right)^2 \frac{1}{1+(\omega/a)^2}.$$

Comparing this with Eq. 11.68, we again see that

$$\Phi'(\omega) = C^2(\omega) + S^2(\omega). \tag{11.72}$$

This relationship between the autocorrelation and Φ' can be proved in general by representing each function in the definition of the autocorrelation function by its Fourier transform, using the trigonometric addition formulas, carrying out the time integration first, and using the δ-function definitions. The result is

$$\begin{aligned}\phi_{11}(\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[C^2(\omega)+S^2(\omega)\right]\cos\omega\tau\, d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi'(\omega)\cos\omega\tau\, d\omega. \end{aligned}\tag{11.73}$$

As with the periodic signal, there are two ways to go from the signal to the energy spectrum. The Fourier transform is taken of either the original function or the autocorrelation function. Squaring and adding is done either in the time domain to $y(t)$ to obtain the autocorrelation function, or in the frequency domain by squaring and adding the coefficients. The Fourier transforms can be taken in either direction. Squaring and adding is one-directional and makes it impossible to go from the energy spectrum back to the original function. These processes are illustrated in Fig. 11.36.

11.13 Noise

The function $y(t)$ we wish to study is often the result of a measurement of some system: the electrocardiogram, the electroencephalogram (EEG), blood flow, etc., and is called a *signal*. Most signals are accompanied by *noise*. Random noise fluctuates in such a way that we cannot predict what its value will be at some future time. Instead we must talk about the probability that the noise has a certain value. A key problem is to learn as much as one can about a signal that is contaminated by noise. The techniques discussed in this chapter are often useful.

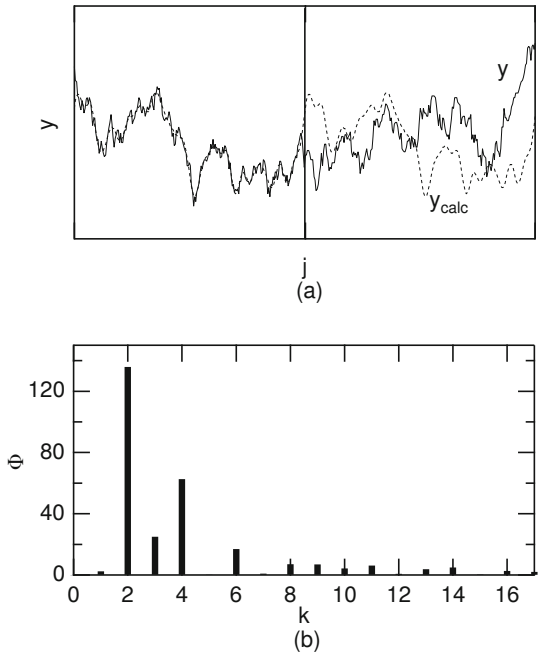


Fig. 11.37 **a** The solid line shows function y_j that was calculated from a one-dimensional random walk with a Gaussian-distributed step length. The dashed line shows the function calculated from the Fourier coefficients of y_j based on the first half of the time interval. It does not fit the second half of the function. This is characteristic of random functions. **b** The power spectrum calculated from the first half of y_j. The zero-frequency component has been suppressed because it depends on the starting value of y

A very important property of noise can be seen from the data shown in Fig. 11.37a. The data consist of 460 discrete values that appear to have several similar peaks. A discrete Fourier transform of the first 230 values gives fairly large values for the first few coefficients a_k and b_k. Yet these values of a_k and b_k fail to describe subsequent values of y_j. The reason is that the y_j are actually random. In this case they are the net displacement after j steps in a random walk in which each step length is Gaussian distributed with standard deviation $\sigma = 5$. The Fourier transform of a random function does not exist. We can apply the recipe to the data and calculate the coefficients. But if we apply the same recipe to some other set of data points from the random function we get different values of the coefficients, although the sum of their squares, $(a_k^2 + b_k^2)^{1/2}$, would be nearly the same. The sum of the squares of the coefficients is plotted in Fig. 11.37b. It is the phases that change randomly, while the amount of energy at a particular frequency remains constant or fluctuates slightly about some average value.

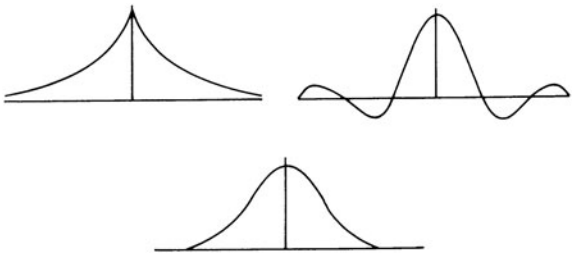


Fig. 11.38 Some possible autocorrelation functions of noise

Noise is not periodic, but neither is it a pulse. It has finite power, but it will have infinite energy if the noise goes on "forever." To describe noise we must use averages, calculated over a time interval that is "long enough" so that the average does not change. Suppose that we are measuring the electrical potential between a pair of electrodes on the scalp. Assume that there is no obvious periodicity, and we think it is noise. If we measure the potential for only a few milliseconds, we will get one average value. If we measure for the same length of time a few minutes later, we may get a different average. But if we average for 2 or 3 min, then a repetition gives almost the same average.

In general, random signals may vary with time in such a way that this average changes. (If we repeat the measurements on the scalp in a few hours, the averages may be different.) We will *assume* that properties such as the mean and standard deviation and power spectrum do not change with time, so that if we average over a "long enough" interval and repeat the average at a later time, we get the same result. Processes that generate data with these properties are called *stationary*. We limit our discussion to stationary random processes.

The correlation functions are not particularly useful for well-defined periodic signals, but they are very useful to describe noise or a signal that is contaminated by noise. (In fact, they allow us to detect a periodic signal that is completely hidden by noise. The technique is described in the next section.)

Space limitations require us to state some properties of the autocorrelation function of noise without proof, though the results are plausible. Many discussions of noise are available. An excellent one with a biological focus is by DeFelice (1981).

The autocorrelation function is given by Eq. 11.47:

$$\phi_{11}(\tau) = \langle y_1(t)y_1(t+\tau)\rangle = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y_1(t)y_1(t+\tau)\,dt.$$

The properties of the autocorrelation function depend on the details of the noise. Some possible shapes for the autocorrelation function are shown in Fig. 11.38.

The following properties of the autocorrelation function can be proved:

1. The autocorrelation function is an even function of τ. This follows from the definition.
2. The autocorrelation function for $\tau = 0$, $\phi_{11}(0)$, measures the average power in the signal. This also follows from the definition.
3. For a random signal with no constant or periodic components, the autocorrelation function goes to zero as $\tau \to \infty$. This is plausible, since for large shifts, if the signal is completely random, there is no correlation.
4. The autocorrelation function has its peak value at $\tau = 0$. This is also plausible, since for any shift of a random signal there will be some loss of correlation.

11.14 Correlation Functions and Noisy Signals

The autocorrelation function is useful for detecting a periodic signal in the presence of noise. We assume that the system that measures these is linear: the response to two simultaneous signals is the sum of the responses to each individually. Section 11.18 will consider what happens when the response is nonlinear.

11.14.1 Detecting Signals in Noise

Suppose that the periodic signal is $s(t)$, the random noise is $n(t)$, and the average of both is zero. The combination of signal and noise is

$$y(t) = s(t) + n(t). \tag{11.74}$$

The autocorrelation of the combination is

$$\begin{aligned}\phi_{yy}(\tau) &= \langle [s(t)+n(t)]\,[s(t+\tau)+n(t+\tau)]\rangle \\ &= \langle s(t)s(t+\tau)\rangle + \langle s(t)n(t+\tau)\rangle \\ &\quad + \langle n(t)s(t+\tau)\rangle + \langle n(t)n(t+\tau)\rangle\,.\end{aligned}$$

Each term in the average can be identified as a correlation function:

$$\phi_{yy}(\tau) = \phi_{ss}(\tau) + \phi_{sn}(\tau) + \phi_{ns}(\tau) + \phi_{nn}(\tau).$$

Since the noise is random, the cross-correlations ϕ_{ns} and ϕ_{sn} should be zero if the averages were taken over a sufficiently long time. Therefore,

$$\phi_{yy}(\tau) = \phi_{ss}(\tau) + \phi_{nn}(\tau). \tag{11.75}$$

The autocorrelation of a periodic signal is periodic in τ, while the autocorrelation of the noise approaches zero if τ is long enough.

If we suspect that a periodic signal is masked by noise, we can calculate the autocorrelation function. If the autocorrelation function shows periodicity that persists for long shift times τ, a periodic signal is present. The period of the correlation function is the same as that of the signal. Acquisition of the data and calculation of the correlation function are done with digital techniques. Press et al. (1992) have an excellent discussion of the techniques and pitfalls.

11.14.2 Signal Averaging

If the period of a signal is known to be T, perhaps from the autocorrelation function or more likely because one is looking for the response evoked by a periodic stimulus, it is possible to take consecutive segments of the combined signal plus noise of length T, place them one on top of another, and average them. One can also do this for the response evoked by a stimulus. The signal will be the same in each segment, while the noise will be uncorrelated. After N sampling periods, *signal averaging* reduces the noise by $1/\sqrt{N}$.

Examples of this are the *visual* or *auditory evoked response*. The signal in the electroencephalogram or magnetoencephalogram is measured in response to a flash of light or an audible click. (In other experiments the subject may perform a repetitive task.) The stimulus is repeated over and over while the signal plus noise is recorded and averaged. The average reproduces the shape of the signal. Figure 11.39 shows an example of signal averaging for an evoked response in the EEG for increasing values of N.

The signal-averaging procedure can also be described in terms of a cross-correlation with a series of δ functions at the stimulus times. Suppose a local signal $l(t)$ is produced in synchrony with the stimulus. The cross-correlation of $l(t)$ with $y(t)$ is

$$\phi_{yl}(\tau) = \langle [s(t) + n(t)]\, l(t+\tau) \rangle = \phi_{sl} + \phi_{nl}.$$

Whatever the local signal is, its cross-correlation with the noise approaches zero for long averaging times, so

$$\phi_{yl}(\tau) = \phi_{sl}(\tau). \tag{11.76}$$

If the local signal is a series of narrow spikes approximated by δ functions, then

$$l(t) = \delta(t) + \delta(t-T) + \delta(t-2T) + \cdots .$$

Since both $s(t)$ and $l(t)$ are periodic with the same period, the average can be taken over a single period. The integral then contains one δ function:

$$\phi_{yl}(-\tau) = \phi_{sl}(-\tau) = \frac{1}{T}\int_0^T s(t)\delta(t-\tau)\,dt = \frac{s(\tau)}{T}.$$

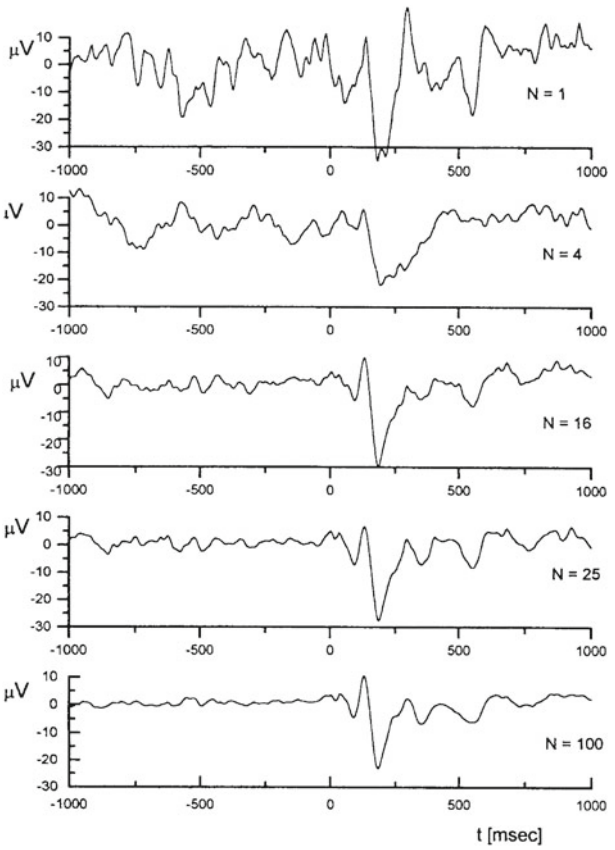


Fig. 11.39 An example of signal averaging. An evoked response is recorded along with the EEG from a scalp electrode. As the number of repetitions N is increased, the EEG background noise decreases and the evoked response stands out. (Copyright © 2006 from Mainardi et al. 2006, pp. 2-1–2-25. Reproduced by permission of Taylor & Francis LLC. Permission conveyed through Copyright Clearance Center, Inc.)

11.14.3 Power Spectral Density

We have already seen that the Fourier transform of a random signal does not exist. Because the phases of a random signal are continually changing, we were unable to predict the future behavior of a time series in Fig. 11.37. If the signal is stationary, averages, including the average power, do not change with time and have meaning. The autocorrelation function of a random signal does exist, and so does the Fourier transform of the autocorrelation. If $\Phi(\omega)$ is the Fourier transform of the autocorrelation function of a random signal, then

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} y^2(t)\,dt = \int_{-\infty}^{\infty} \Phi(\omega)\frac{d\omega}{2\pi}, \tag{11.77}$$

and we can think of Φ as giving the power between frequencies f and $f + df$. This is called the *Wiener theorem for random signals*. The quantity Φ is often called the *power spectral density* or PSD. Figure 11.40 summarizes how the

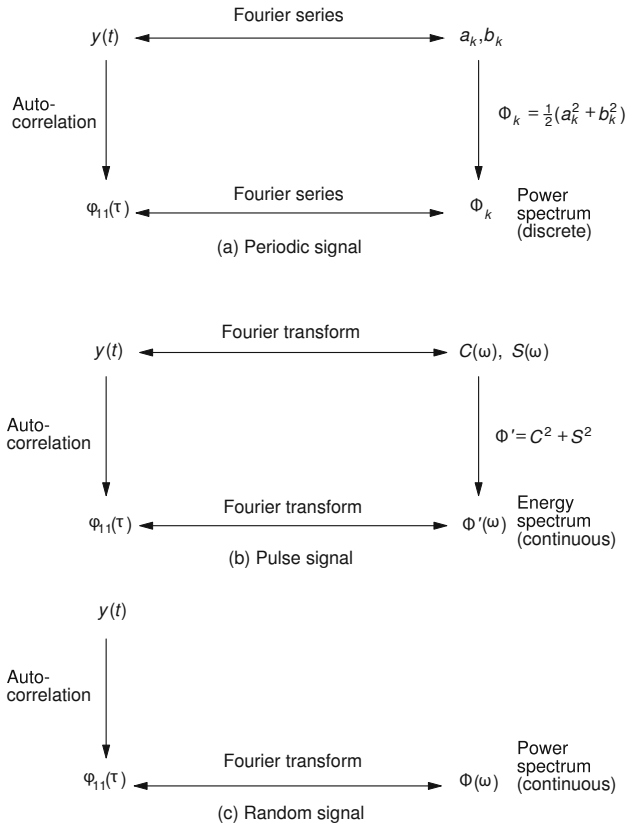


Fig. 11.40 The relationships between the power spectrum or energy spectrum and a periodic signal (**a**), a pulse (**b**), and a random signal (**c**). The Fourier transform and series are bidirectional; the other processes are not

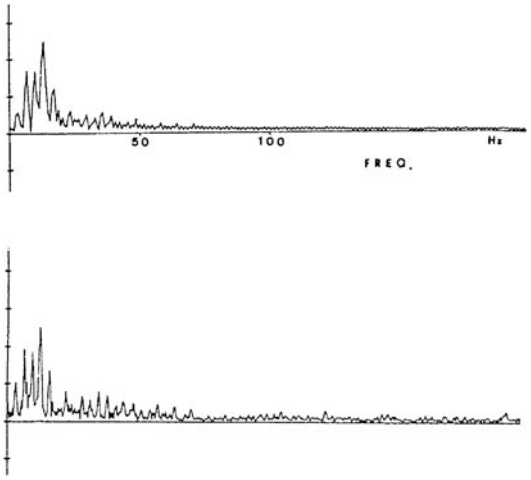


Fig. 11.41 The power spectrum from a surface electromyogram calculated two different ways. The upper panel shows the Blackman–Tukey method, which is a fast Fourier transform (FFT) of a digital estimate of the autocorrelation function. The lower panel is the sum of the squares of the coefficients in a direct fast Fourier transform of the discrete data. (Copyright © 2006 from Cohen 2006. Reproduced by permission of Taylor & Francis Group, LLC. Permission conveyed through Copyright Clearance Center, Inc.)

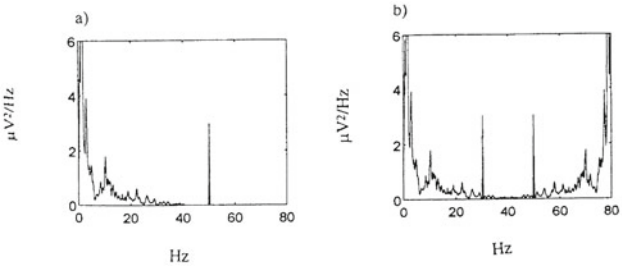


Fig. 11.42 The power spectrum of an electroencephalogram signal showing the problem with aliasing, and also the presence of negative frequencies appearing as positive frequencies above the Nyquist frequency. (Copyright © 2006 from Mainardi et al. 2006. Reproduced by permission of Taylor & Francis Group, LLC. Permission conveyed through Copyright Clearance Center, Inc.)

power or energy spectrum can be obtained for a periodic signal, a pulse, and a random signal.

In the digital realm there are several ways to calculate the power spectral density.[7] The Blackman–Tukey method makes a digital estimate of the correlation function and takes its discrete Fourier transform, as described in Fig. 11.40c. The *periodogram* uses the discrete Fourier transform directly. Though the Fourier transform of a random signal does not exist because of the randomly changing phases, the sum of the squares of the coefficients is stable. In fact, we plotted Φ_k calculated from the discrete Fourier transform in Fig. 11.37b. Figure 11.41 shows both ways of calculating $\Phi(f)$ for a surface electromyogram—the signal from a muscle measured on the surface of the skin. Slight differences can be seen, but they are not significant.

Figure 11.42 shows the power spectrum of an EEG signal and also the effect of aliasing. The original signal has no frequency components above 40 Hz. Sampling was done at 80 Hz. A 50-Hz power frequency signal was added, and the Fourier transform shows a spurious response at 30 Hz. The second panel also shows the mirror-image power spectrum from 40 to 80 Hz that should be thought of as occurring at negative frequencies (the factor of 2 again).

11.14.4 Units

It is worth pausing to review the units of the various functions we have introduced. They become confusing because we have three different cases: a periodic signal that is infinite in extent, a pulse signal that is of finite duration, and a

[7] See Press et al. (1992), Cohen (2006), or Mainardi et al. (2006).

Table 11.5 Units used in the various functions in this chapter, assuming that y is measured in (unit)

Type of function Signal	Expansion coefficients	Correlation functions	Power or energy
Discrete periodic			Power (unit^2)
y (unit)	a_k, b_k (unit)	ϕ (unit^2)	Φ_k (unit^2)
Pulse			Energy
y (unit)	C, S (unit s)	ϕ (unit^2 s)	$\Phi'(\omega)$ (unit^2 s^2)
			$\Phi'(\omega)\, d\omega$ (unit^2 s)
Random			Power (unit^2)
y (unit)		ϕ (unit^2)	$\Phi(\omega)$ (unit^2 s)
			$\Phi(\omega)\, d\omega$ (unit^2)

random-noise signal that is also infinite in extent but not periodic. For both signals that are infinite in extent we must use the "power," and for the pulse we must use "energy."

Often in signal analysis the units of "power" and "energy" may not be watts or joules. If the signal is a voltage, then the power dissipated in resistance R is v^2/R in watts. Our "power" defined from the equations above would be just v^2.

Suppose that the signal y is measured in "units." Then the "power" is in (units)^2 and the "energy" for a pulse is in (units)^2 s. The correlation functions for the infinite signals are in (units)^2 while those for pulses are in (units)^2 s. Table 11.5 summarizes the situation.

11.15 Frequency Response of a Linear System

Chapter 10 discussed feedback in a linear system in terms of the solution of a differential equation that described the response of the system as a function of time. The simplest system treated there was described by Eq. 10.20:

$$\tau_1 \frac{dx}{dt} + x = ap(t) + G_1 y(t). \tag{11.78}$$

Function $p(t)$ is the *input signal*. This equation was combined with Eq. 10.21 to obtain

$$\tau_1 \frac{dx}{dt} + (1 - G_1 G_2)x = ap(t). \tag{11.79}$$

It is often useful to characterize the behavior of a system by its response to sine waves of different frequencies instead of by its time response. The most familiar example is the audio amplifier: the output signal $x(t)$ is some function of an input signal $p(t)$ that is seldom a pure sinusoid. An equation analogous to Eq. 11.79 relates x and p. The amplifier is usually descried as having "a frequency response of −0.5 dB at 10 Hz and 30 kHz." It is easy to feed a sinusoidal signal of different frequencies into the amplifier and measure the amplitude ratio of the output sine wave to the input sine wave.[8]

To describe the amplifier completely, it is also necessary to measure the phase delay or the time delay at each frequency. The combination of amplitude and phase response is called the *transfer function* of the amplifier.

In principle, once the properties of a linear system are known, either in terms of a differential equation or the transfer function, its response to any input can be calculated. In the time domain, one solves the differential equation with input $p(t)$ on the right-hand side. In the frequency domain, one computes the Fourier transform of the input, makes the appropriate changes in amplitude and phase at every frequency according to the transfer function, and takes the inverse Fourier transform of the result. The inverse transform gives the output response as a function of time. Sometimes the differential equation may be impossible to solve analytically or the inverse Fourier transform cannot be obtained, and numerical solutions are all that can be obtained.

The frequency-response technique may be particularly useful if the system has several stages (a microphone, an amplifier, one or more loudspeakers); one can multiply the amplitudes and add the phases of each stage.

If the differential equation is known, the frequency response can be calculated. Conversely, if the frequency and phase responses are known, the differential equation can be deduced. We give an example of the former approach in this section. The latter technique requires more mathematics than we have developed.

11.15.1 Example of Calculating the Frequency Response

As an example of the frequency response method of describing the system, consider Eq. 11.79. With $G_2 = 0$, the results apply to the case without feedback, Eq. 11.78. Let $p(t) = \cos \omega t$ and $a = 1$. We want a solution of the form

$$x(t) = G(\omega) \cos(\omega t - \theta), \tag{11.80}$$

where $G(\omega)$ is the overall gain or amplitude ratio, and $\theta(\omega)$ the phase shift, at frequency ω. We can show by substitution

[8] The technique works only for a linear system. If the system is not linear, the output will not be sinusoidal.

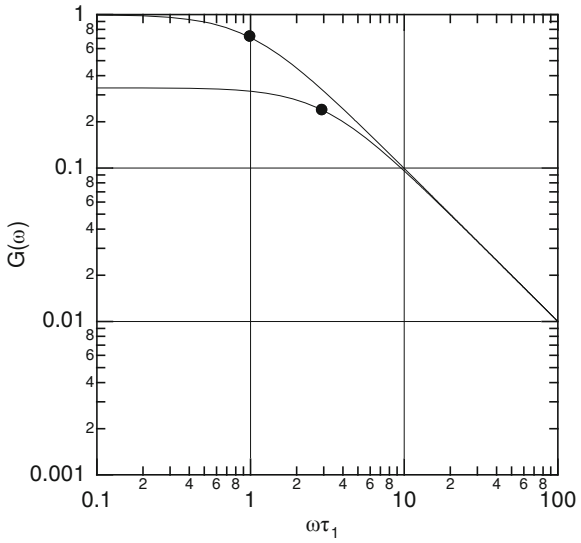


Fig. 11.43 Plot of $G(\omega)$ for a system described by Eq. 11.81. Two cases are shown: without feedback ($1 - G_1G_2 = 1$) and with feedback ($1 - G_1G_2 = 3$). The dots mark the half-power frequencies (see text)

that Eq. 11.80 is a solution of Eq. 11.79 if

$$G(\omega) = \frac{1}{1 - G_1G_2}\left(\frac{1}{1 + \omega^2\tau_1^2/(1 - G_1G_2)^2}\right)^{1/2},$$
$$\tan\theta = \frac{\omega\tau_1}{1 - G_1G_2}. \tag{11.81}$$

The behavior of the gain is plotted in Fig. 11.43, both without feedback ($1 - G_1G_2 = 1$) and with feedback ($1 - G_1G_2 = 3$). At low frequencies the gain is constant. It falls at high frequencies ($\omega\tau_1 \gg 1$) as ω^{-1}. When $\omega = 1/\tau_1$ (without feedback) or $\omega = 3/\tau_1$ (with feedback), the gain is $1/\sqrt{2}$ times its value at zero frequency. This frequency is called the *half-power frequency* because the power is proportional to the square of the signal and its value at the half-power frequency is 1/2 times its value at zero frequency.

Negative feedback reduces the gain and also raises the half-power frequency from $1/\tau_1$ to $(1 - G_1G_2)/\tau_1$. The time constant is reduced by the feedback from τ_1 to $\tau_1/(1 - G_1G_2)$. Recall Eq. 10.23.

11.15.2 The Decibel

The gain is often expressed in decibels[9] (dB):

$$\text{gain(dB)} = 20\log_{10} G(\omega). \tag{11.82}$$

A gain ratio of unity is equivalent to 0 dB; a gain of 1000 is $20\log_{10}(1000) = 60$ dB. One advantage to expressing gain in decibels is that the gains in dB for several stages add. If the first process has a gain of 2 (6 dB) and the second has a gain of 100 (40 dB), the overall gain is 200 (46 dB). For the amplifier whose gain has fallen by 0.5 dB at 10 Hz and 30 kHz, the ratio $G(\omega)/G_{\max}$ is given by solving

$$-0.5 = 20\log_{10}(G/G_{\max}),$$
$$G/G_{\max} = 10^{-0.025} = 0.944.$$

The gain has fallen to 94.4 % of its maximum value at 10 Hz and 30 kHz. If the maximum gain were 1000 (60 dB), then the gain would have fallen to 944 (59.5 dB) at 10 and 30 kHz.

The fall in gain is called the *roll-off*, in this case the high-frequency roll-off. At high frequencies the gain is proportional to $1/\omega$, so it drops by a factor of 2 (6 dB) when the frequency doubles (1 octave). Therefore, the gain has a high-frequency roll-off of 6 dB per octave. A roll-off of 6 dB per octave is characteristic of systems with a single time constant, as in Eq. 11.79.

11.15.3 Example: Impulse Response

As an example, we show that the response of the system to a δ function calculated in the time domain is consistent with the frequency response. Let the input be $p(t) = \delta(t)$. The Fourier transform of the input is

$$C_{\text{in}}(\omega) = \int_{-\infty}^{\infty} \delta(t)\cos\omega t\, dt = 1,$$
$$S_{\text{in}}(\omega) = \int_{-\infty}^{\infty} \delta(t)\sin\omega t\, dt = 0.$$

The δ function contains constant power at all frequencies. The sine coefficients are zero because a δ function at $t = 0$ is an even function. The gain and phase delay are applied to $C(\omega)$ to get the Fourier transform of the output signal. Although we started with a purely even function (only cosine terms) the phase shift means that the output contains both sine and cosine terms. To calculate the output, we write Eq. 11.80 as

$$x(t) = \int [G(\omega)\cos\theta\cos\omega t + G(\omega)\sin\theta\sin\omega t]\, d\omega,$$

from which

$$C_{\text{out}}(\omega) = G(\omega)\cos\theta,$$
$$S_{\text{out}}(\omega) = G(\omega)\sin\theta.$$

[9] The bel is the logarithm to the base 10 of the *power* ratio. The decibel is one tenth as large as the bel. Since the power ratio is the square of the voltage ratio or gain, the factor in Eq. 11.82 is 20.

From Eq. 11.81 we get (letting $G_2 = 0$ and doing a fair amount of algebra)

$$C_{\text{out}}(\omega) = \frac{1}{1+\omega^2\tau_1^2}, \qquad S_{\text{out}}(\omega) = \frac{\omega\tau_1}{1+\omega^2\tau_1^2}. \tag{11.83}$$

It is easier to solve the differential equation, take the Fourier transform of the solution, and compare it to Eq. 11.83 than it is to find the inverse transform with the mathematical tools at our disposal. For $G_2 = 0$ the equation to be solved is

$$\tau_1 \frac{dx}{dt} + x = \delta(t).$$

For all positive t a steady-state solution is $x(t) = 0$. The solution of the homogeneous equation is $x(t) = Ae^{-t/\tau_1}$. The value A is obtained by integrating the equation from $-\epsilon$ to ϵ as $\epsilon \to 0$:

$$\tau_1 \int_{-\epsilon}^{\epsilon} \frac{dx}{dt}\, dt + \int_{-\epsilon}^{\epsilon} x\, dt = \int_{-\epsilon}^{\epsilon} \delta(t)\, dt.$$

The first term is $x(\epsilon) - x(-\epsilon) \to x(0) - 0$. The second term vanishes in the limit, since x is finite and the width goes to zero. From the definition of the δ function the right-hand side of the equation is 1. Therefore

$$x = \begin{cases} 0, & t < 0 \\ (1/\tau_1)e^{-t/\tau_1}, & t > 0. \end{cases} \tag{11.84}$$

The Fourier coefficients of this function were calculated in Eqs. 11.61. They are

$$C(\omega) = \frac{1}{1+\omega^2\tau_1^2}, \qquad S(\omega) = \frac{\omega\tau_1}{1+\omega^2\tau_1^2}.$$

These agree with Eqs. 11.83. We have demonstrated that the response of this particular linear system to a δ function is the Fourier transform of the transfer function of the system.

11.16 The Frequency Spectrum of Noise

In Sect. 9.8, we introduced Johnson noise and shot noise. Both are inescapable. Johnson noise arises from the Brownian motion of charge carriers in a conductor; shot noise arises from fluctuations due to the discrete nature of the charge carriers.

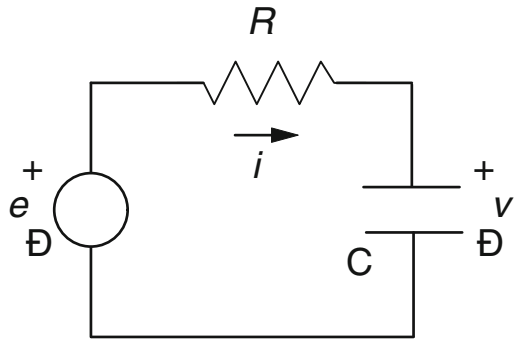


Fig. 11.44 The circuit for analyzing the noise produced by a resistance R connected to capacitance C. The circuit assumes that the noise is generated in a voltage source $e(t)$ in series with the resistance. The voltage across the capacitance is v

11.16.1 Johnson Noise

When we introduced Johnson noise we said nothing about its frequency spectrum. We used the equipartition theorem to argue that since the energy on a capacitor depends on the square of the voltage, there would be fluctuations in a capacitor whose average voltage is zero given by (in the notation of this chapter)

$$\frac{1}{2}C\left\langle v^2\right\rangle = \frac{1}{2}k_B T. \tag{11.85}$$

(In this section we will use T both for time and, when immediately following the Boltzmann constant, for temperature. We also have, briefly, C for capacitance as well as for the Fourier cosine coefficient. We will avoid the use of C for capacitance as much as possible.)

If the capacitor is completely isolated the charge on its plates, and hence the voltage between them, cannot fluctuate. The equipartition theorem applies to the capacitor only when it is in thermal equilibrium with its surroundings. This thermal contact can be provided by a resistor R between the plates of the capacitor. It is actually the Brownian movement of the charge carriers in this resistor that cause the Johnson noise. In analyzing the noise in electric circuits, it is customary to imagine that the noise arises in an *ideal voltage source*: a "battery" that maintains the voltage across its terminals, fluctuating randomly with time, regardless of how much current flows through it. It is placed in series with the resistor. This is not a real source. It is a fictitious source that gives the correct results in circuit analysis. We call the voltage across this noise source $e(t)$ and we want to learn about its properties.

Imagine that we place the noise source and its associated resistor across the plates of a capacitor, as shown in Fig. 11.44. We want to relate the voltage across the capacitor, v, to the voltage across the noise source, e. We know that $e(t) = v(t) + Ri(t)$, and that $i = Cdv/dt$. (See the

discussion surrounding Eq. 6.36 and 6.37.) Therefore

$$e(t) = v(t) + RC\frac{dv}{dt} = \tau_1\frac{dv}{dt} + v. \tag{11.86}$$

(By introducing $\tau_1 = RC$ we eliminate the need to use C for capacitance until the very end of the argument. We use the subscript on τ_1 to distinguish it from the argument of the correlation function.)

Even though the voltage is random, let us assume we can write it as a Fourier integral. Our final results depend only on the power spectrum and not on the phases. We write

$$v(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}[C(\omega)\cos\omega t + S(\omega)\sin\omega t]\,d\omega. \tag{11.87}$$

Differentiating this gives an expression for dv/dt:

$$\frac{dv}{dt} = \frac{1}{2\pi}\int_{-\infty}^{\infty}[-\omega C(\omega)\sin\omega t + \omega S(\omega)\cos\omega t]\,d\omega. \tag{11.88}$$

Combining these with Eq. 11.86 gives us the Fourier transform of $e(t)$:

$$\begin{aligned} e(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\{[C(\omega) + \omega\tau_1 S(\omega)]\cos\omega t \\ &\quad + [S(\omega) - \omega\tau_1 C(\omega)]\sin\omega t\}\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}[\alpha(\omega)\cos\omega t + \beta(\omega)\sin\omega t)]\,d\omega. \end{aligned}$$

We now need to calculate $\langle v^2(t)\rangle$ and $\langle e^2(t)\rangle$. The calculation is exactly the same as what we did to derive Parseval's theorem in Eqs. 11.67–11.68, except that we are dealing with random signals instead of pulses and we have to introduce

$$\lim_{T\to\infty}\frac{1}{2T}$$

on each side of the equation. When we do this, we find

$$\begin{aligned} \langle v^2(t)\rangle &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[C^2(\omega) + S^2(\omega)\right]d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_v(\omega)\,d\omega, \\ \langle e^2(t)\rangle &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\alpha^2(\omega) + \beta^2(\omega)\right]d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_e(\omega)\,d\omega. \end{aligned} \tag{11.89}$$

If we expand Φ_e, we find that

$$\begin{aligned} \Phi_e(\omega) &= \alpha^2(\omega) + \beta^2(\omega) \\ &= \left[C^2(\omega) + S^2(\omega)\right](1 + \omega^2\tau_1^2) \\ &= \Phi_v(\omega)(1 + \omega^2\tau_1^2). \end{aligned} \tag{11.90}$$

Johnson noise was discovered experimentally by J. B. Johnson in 1926. The next year Nyquist explained its origin using thermodynamic arguments and showed that until one reaches frequencies high enough so that quantum-mechanical effects are important, Φ_e is a constant independent of frequency (Nyquist 1928). We will not reproduce his argument; rather we will assume that Φ_e is a constant and find the value of Φ_e for which the mean square voltage across the capacitor satisfies the equipartition theorem, Eq. 11.85.

The expression for Φ_v becomes

$$\Phi_v(\omega) = \frac{\Phi_e}{1 + \omega^2\tau_1^2}, \tag{11.91}$$

and from the first of Eqs. 11.89,

$$\begin{aligned} \langle v^2(t)\rangle &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_v(\omega)\,d\omega = \frac{\Phi_e}{2\pi}\int_{-\infty}^{\infty}\frac{d\omega}{1 + \omega^2\tau_1^2} \\ &= \frac{\Phi_e}{2\pi\tau_1}\int_{-\infty}^{\infty}\frac{dx}{1 + x^2} \\ &= \frac{\Phi_e}{2\pi\tau_1}\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right] = \frac{\Phi_e}{2\tau_1}. \end{aligned} \tag{11.92}$$

Putting this expression in the equipartition statement, Eq. 11.85, and remembering that $\tau_1 = RC$, we obtain

$$\begin{aligned} \frac{C\langle v^2(t)\rangle}{2} &= \frac{1}{2}C\frac{\Phi_e}{2RC} = \frac{k_B T}{2}, \\ \Phi_e &= 2Rk_B T. \end{aligned} \tag{11.93}$$

The units of Φ_e are $\mathrm{V^2\ s}$ or $\mathrm{V^2\ Hz^{-1}}$. This is for frequencies that extend from $-\infty$ to ∞. If we were dealing with only positive frequencies, we would have

$$\Phi_e = 4Rk_B T \quad \text{(using positive frequencies only)}. \tag{11.94}$$

Either way, this says that the power spectrum for the fictitious source $e(t)$ is constant so there is equal power at all frequencies (up to the limits imposed by quantum mechanical effects). For this reason, Johnson noise is called *white noise,* in analogy with white light that contains all frequencies. The voltage fluctuations across the capacitor have the power spectrum

$$\Phi_v(\omega) = \begin{cases} \dfrac{2Rk_B T}{1 + \omega^2\tau_1^2}, & -\infty < \omega < \infty \\ \dfrac{4Rk_B T}{1 + \omega^2\tau_1^2}, & 0 < \omega < \infty. \end{cases} \tag{11.95}$$

Figure 11.45 shows the Johnson-noise power spectra and rms voltage spectra plotted vs frequency. These are based on $T = 300$ K, $R = 10^6\ \Omega$, $C = 10^{-9}$ F, and $\tau_1 = RC = 10^{-3}$ s. The labels on the ordinates are worth discussion. On the left we have Φ/R, which from Eq. 11.95 is in joules,

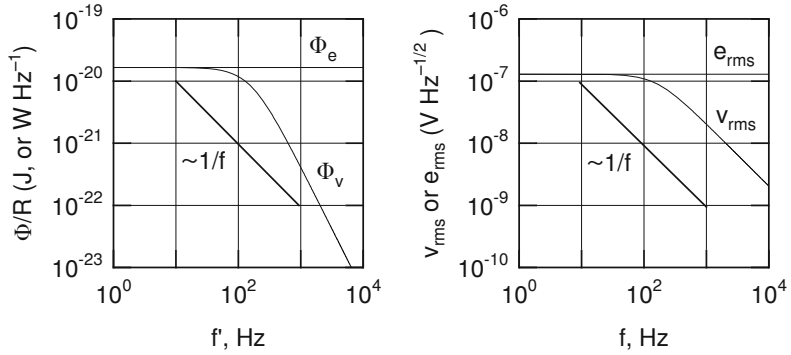


Fig. 11.45 The power spectrum of the noise source e and the voltage across the capacitor v. The left panel plots Φ/R vs f. The right panel plots $v_{\rm rms}$ in each frequency interval. The parameters are described in the text

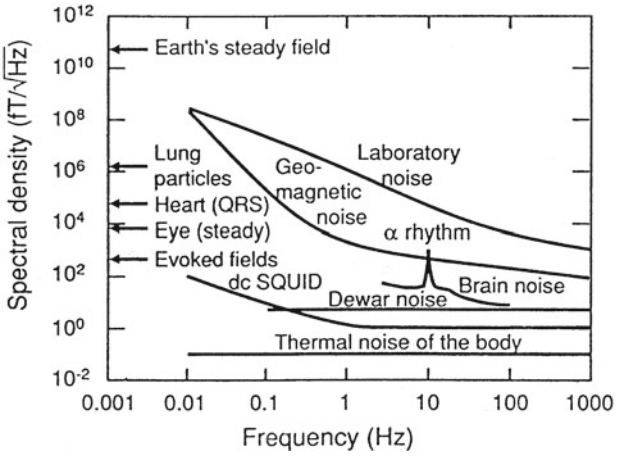


Fig. 11.46 Spectral density of various sources of the magnetic field, expressed in terms of the magnetic field in femtotesla (1 fT = 10^{-15} T). (Reprinted with permission from Hämäläinen et al. 1993. Copyright 1993 by the American Physical Society)

which is W s or W Hz^{-1}. The units for the graph on the right that are consistent with this are W$^{1/2}$ s$^{1/2}$ = W$^{1/2}$ Hz$^{-1/2}$ = V $\Omega^{-1/2}$ Hz$^{-1/2}$. The resistance has been included to make the units V Hz$^{-1/2}$. The $1/f^2$ falloff at high frequencies is due to the frequency response of the RC circuit and is not characteristic of the noise.

Figure 11.46 shows an example: the spectral density of the magnetic field from an article on the magnetoencephalogram. The units are femtotesla Hz$^{-1/2}$ (1 femtotesla = 1 fT = 10^{-15} T).

We can determine the autocorrelation functions $\Phi_{ee}(\tau)$ and $\Phi_{vv}(\tau)$. Equation 11.73 gave the Fourier transform of the autocorrelation function for a pulse. For a random signal the autocorrelation is very similar but involves the power instead of the energy:

$$\phi_{ee}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_e(\omega)\cos\omega t\, d\omega,$$
$$\phi_{vv}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_v(\omega)\cos\omega t\, d\omega. \tag{11.96}$$

For the voltage source the autocorrelation function is

$$\phi_{ee}(\tau) = \frac{2Rk_BT}{2\pi}\int_{-\infty}^{\infty}\cos\omega t\, d\omega. \tag{11.97}$$

To evaluate this, consider Eq. 11.66a, which shows the Fourier transform of the δ function. The integral there is over time. Interchange the time and angular frequency variables to write

$$\int_{-\infty}^{\infty}\cos\omega\tau\ \cos\omega\tau'\, d\omega = 2\pi\delta(\tau-\tau'). \tag{11.98}$$

Let $\tau' = 0$:

$$\int_{-\infty}^{\infty}\cos\omega\tau\, d\omega = 2\pi\delta(\tau). \tag{11.99}$$

The final expression for the autocorrelation function of the noise source is

$$\Phi_{ee}(\tau) = 2Rk_BT\,\delta(\tau). \tag{11.100}$$

To find $\phi_{vv}(\tau)$, consider the discussion surrounding Eqs. 11.70 and 11.71. There we discussed the Fourier transform pair (letting $a = 1/\tau_1$)

$$\frac{A^2}{1+\omega^2\tau_1^2} \xleftrightarrow{\text{Fourier transform}} \frac{A^2}{2\tau_1}e^{-|\tau|/\tau_1}, \tag{11.101}$$

from which we obtain the autocorrelation function for the voltage across the capacitor:

$$\Phi_{vv}(\tau) = \frac{Rk_BT}{\tau_1}e^{-|\tau|/\tau_1}. \tag{11.102}$$

Let us compare these two results. The autocorrelation of the noise source is a δ function. Any shift at all destroys the correlation. The noise equivalent voltage source and resistor, isolated from anything else, respond instantaneously to random noise changes, the correlation function is infinitely narrow, and all frequencies are present. When the source and resistor are connected to a capacitor, the voltage across the capacitor cannot change instantaneously. There is a high-frequency roll-off, and the voltage at one time is correlated with the voltage at surrounding times. As the time constant of the circuit becomes smaller, $\Phi_{vv}(\tau)$ becomes narrower and taller, approaching the δ function.

The power spectrum across the capacitor has the same form as the square of the magnitude of the gain (transfer function) of Eq. 11.81. This is the transfer function for an RC circuit, as can be seen by comparing Eq. 11.79 with Eq. 11.86. This is a special case of a general result, that linear systems can be analyzed by measuring how they respond to white noise.

11.16.2 Shot Noise

Chapter 9 also mentioned shot noise, which occurs because the charge carriers have a finite charge, so the number of them passing a given point in a circuit in a given time fluctuates about an average value. One can show that shot noise is also white noise.

11.16.3 $1/f$ Noise

Johnson noise and shot noise are fundamental and independent of the details of the construction of the resistance. The former depends on the Brownian motion of the charge carriers, and the latter depends on the number of charge carriers required to transport a given amount of charge. They are irreducible lower limits on the noise (for a given resistance and temperature). If one measures the noise in a real resistor in a circuit, one finds additional or "excess" noise that can be reduced by changing the materials or construction of the resistor. This excess noise often has a $1/f$ frequency dependence. For white noise the power in every frequency interval is proportional to the width of the interval, so there is 10 times as much power in the frequency decade from 10 to 100 Hz as in the decade from 1 to 10 Hz. For $1/f$ noise, on the other hand, there is equal power in each frequency decade. This kind of noise is sometimes called "pink noise" in allusion to the fact that pink light has more power in the red (lower frequency) part of the spectrum than the rest.

Noise with a $1/f$ spectrum had been discovered in many places: resistors, transistors, and the fluctuations in the flow of sand in an hourglass, in traffic flow, in the heartbeat, and even in human cognition. It is thought that there might be some universal principle underlying $1/f$ noise, possibly related to chaos, but this is still an area of active investigation.

11.17 Testing Data for Chaotic Behavior

A major problem in data analysis is to find the meaningful signal due to the physical or biological process in the presence of noise. We have introduced some of the analysis techniques in this chapter. A problem that has become important in recent years is to determine whether a variable that is apparently random is due to truly random behavior in the underlying process or whether the process is displaying chaotic behavior. The techniques for determining this are still under development and are beyond the scope of this book. An excellent introduction is found in Chap. 6 of Kaplan and Glass (1995). We close by mentioning two of the tools used in this analysis: embedding and surrogate data.

One of the problems in analyzing data from complex systems is that we may not be able to measure all of the variables. For example, we may have the electrocardiogram or even an intracellular potential recording but have no information about the details of the ionic currents of several species through the membrane that change the potential. We may measure the level of thyroid hormones T3 and T4 but have no information about the other hormones in the thyroid–hypothalamus–pituitary feedback system. Fortunately, we do not need to measure all the variables. There is a data-reduction technique that can be applied to a few of the variables that shows the dynamics of the full system.

11.17.1 Embedding

To see how embedding works, consider a system with two degrees of freedom described by a set of nonlinear differential equations with the form of Eqs. 10.35. In order to make the subscript on x available to index measurements of the variable at different times, we write the variables as x and y instead of x_1 and x_2:

$$\frac{dx}{dt} = f_1(x, y), \qquad \frac{dy}{dt} = f_2(x, y).$$

A phase-space plot would be in the xy plane. Suppose we only measure variable x, and that we obtain a sequence of measurements $x_j = x(t_j)$. The time derivative is approximately

$$\frac{x_{j+h} - x_j}{t_{j+h} - t_j} \approx \frac{dx}{dt} = f_1(x, y).$$

A series of measurements at different times gives us information about how function f_1 depends on x. A remarkable

result that we state without proof is that it also gives information about the entire system. (See Kaplan and Glass 1995 for a more detailed discussion and references to the literature.) Figure 11.47 shows this in a specific case. It is a calculation using the *van der Pol oscillator*. This nonlinear oscillator has been used to model many systems since it was first proposed in the 1920s. It can be written as the pair of first-order equations

$$\frac{dx}{dt} = \frac{1}{a}\left(y - \frac{x^3}{3} + x\right), \qquad \frac{dy}{dt} = -ax,$$

where a is a very small positive number. The top panel of Fig. 11.47 shows values of x_j vs j (labeled as D_t vs t). The middle panel shows a phase-plane plot of y vs x. The bottom panel plots x_{j+10} vs x_j. Shading is used to identify some of the early data points in all three panels. The trajectory in the bottom panel has all the same characteristics as the phase-plane plot.

This is an example of a general technique called *time-lag embedding*. The set of differential equations with two degrees of freedom has been converted into a nonlinear map in one degree of freedom.

For a system with three degrees of freedom, we could make a three-dimensional plot by creating sets of three numbers from the n measured values, which we can think of and plot as the three components of a vector

$$\mathbf{x}_j = (x_j, x_{j-h}, x_{j-2h}), \qquad j = 2h, 2h+1, \ldots, n-1.$$

In general, we can construct a p-dimensional set of vectors

$$\mathbf{x}_j = (x_j, x_{j-h}, \ldots x_{j-ph}), \qquad j = ph, \ldots, n-1.$$

We call p the *embedding dimension* and h the *embedding lag*. There are a number of further calculations that can be done to the embedded vector to help decide on the behavior of the underlying system. These are described in Kaplan and Glass (1995).

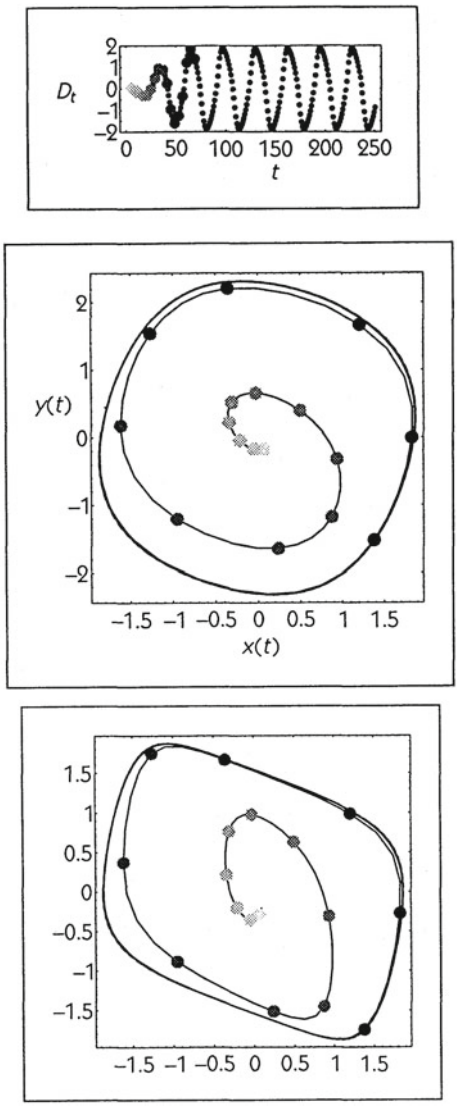


Fig. 11.47 Plots of the solution to the van der Pol equation with a certain set of initial conditions. The top panel shows values of x_j vs j (labeled as D_t vs t). The middle panel is a phase-plane plot of y vs x. The bottom panel plots x_{j+10} vs x_j. Shading is used to identify some of the early data points in all three panels. The trajectory in the bottom panel has the same characteristics as the phase-plane plot. (Used by permission from Kaplan and Glass 1995)

11.17.2 Surrogate Data

In general, a fully conclusive answer to the question of whether the data are due to a random process or a chaotic process cannot be obtained, though strong indications can be. The most rigorous way to test for the presence of chaotic behavior is to make the hypothesis, called the *null hypothesis*, that the data are explained by a linear process plus random noise. One then develops a test statistic (several standard tests are used) and compares the value of the test statistic for the real data to its value for sets of data that are consistent with the null hypothesis. These sets are called *surrogate data*. We examined one linear system with noise: the random walk of Fig. 11.37. The next value in the sequence was the previous value plus random noise. We saw that the power spectrum was defined, but the phases changed randomly. We can think of any linear system driven by random noise as having a defined transfer function $G(\omega)$ with random phases. Therefore, we can generate sets of surrogate data by taking the transform of the original data in the form of an amplitude and phase, related to C and S by Eq. 11.13. We then randomize the phases and calculate the inverse Fourier transform of the randomized coefficients to generate the surrogate data

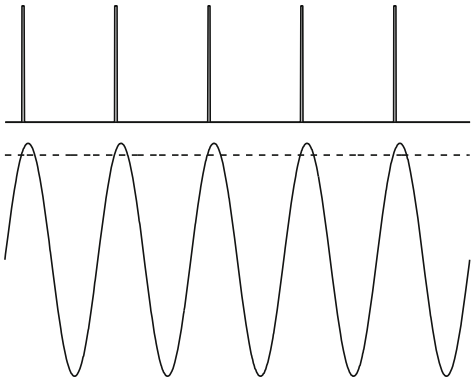


Fig. 11.48 A sine wave of unit amplitude drives a threshold detector. A spike is generated every time the signal rises through 0.9

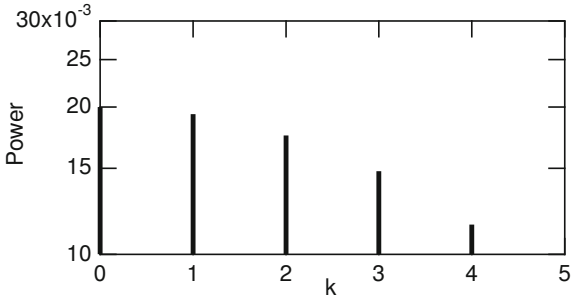


Fig. 11.49 Power spectrum for a train of rectangular pulses of width $2d$ when $d/T = 1/20$

sequence. The surrogate data have the same power spectrum and autocorrelation function as the original data. We then apply the various test statistics. If we were to do this to the data from Fig. 11.37, we would find the tests the same for the original data and the sets of surrogate data, because the original data set is consistent with the null hypothesis.

11.18 Stochastic Resonance

A nonlinear phenomenon called *stochastic resonance* has been recognized in recent years. In stochastic resonance, random fluctuations increase the sensitivity to detect weak signals or allow some other desirable process to take place, such as ion transport. Stochastic resonance takes many forms. It was first invoked in 1981 to explain why the earth has periodic ice ages.[10] It has been proposed as a mechanism in biological processes, but the models are rather complicated.[11] We discuss two simple physical examples.

11.18.1 Threshold Detection

In a linear system, any amount of noise decreases the signal-to-noise ratio. In a nonlinear system, weak noise can enhance signal detection. The simplest nonlinear system that shows this is a threshold detector: an output signal is generated when the input (signal + noise) exceeds a fixed threshold.

Suppose that a sine-wave signal is sent to a threshold detector. Every time the signal rises above the threshold, a pulse is generated, as shown in Fig. 11.48. The output signal is a series of pulses spaced by T, the period of the sine wave. Problem 29 shows that for a series of pulses of width $2d$ separated by time T, the power at frequency $k\omega_0 = 2\pi k/T$ is $\Phi_k = (2/\pi^2 k^2)\sin^2(2\pi kd/T)$. This power spectrum is plotted in Fig. 11.49.

If the amplitude of the sine wave in Fig. 11.48 is less than 0.9, the threshold will never be exceeded. However, if sufficient noise is added to a sine wave that is below threshold, the signal and noise combined will occasionally exceed the threshold. This will happen more frequently when the sine-wave signal is positive than when it is negative so output pulses will occur more frequently during peaks of the signal.

Experiments were done with an electronic circuit that behaves as we have described. The results are shown in Figs. 11.50 and 11.51. Fig. 11.50 shows the weak sinusoidal signal with and without the noise added to it, along with the resulting pulses and the power spectrum. Fig. 11.51 shows the power in the pulse train at the signal frequency and the signal-to-noise ratio, as a function of noise level. The amplitude of the sine wave is 0.1 V. As the noise level increases, both the signal and the SNR increase, reach a maximum, and decrease. The signal-to-noise ratio peaks when the rms noise level is about 0.25 V; the power at the signal frequency peaks at about 0.3 V. As the noise increases above these values the SNR and signal decrease. The lines are theoretical fits; both the theory and the data are described by Gingl et al. (1995).

11.18.2 Feynman's Ratchet

Perpetual motion machines violate either the first or second law of thermodynamics (or both). In his *Lectures on Physics,* Richard Feynman (1963) analyzed a microscopic cog wheel (ratchet) and pawl as shown in Fig. 11.52. Feynman's analysis is elegant, full of insight, and well worth reading. The analysis here follows that in Astumian and Moss (1998). An amount of energy ΔU is required to compress the spring enough to lift the pawl over the tooth. This energy can come

[10] References can be found in the articles by Wiesenfeld and Jaramillo (1998) and by Astumian and Moss (1998).

[11] See Astumian (1997); Astumian and Moss (1998); Wiesenfeld and Jaramillo (1998); Gammaitoni et al. (1998); Adair et al. (1998); Glass (2001).

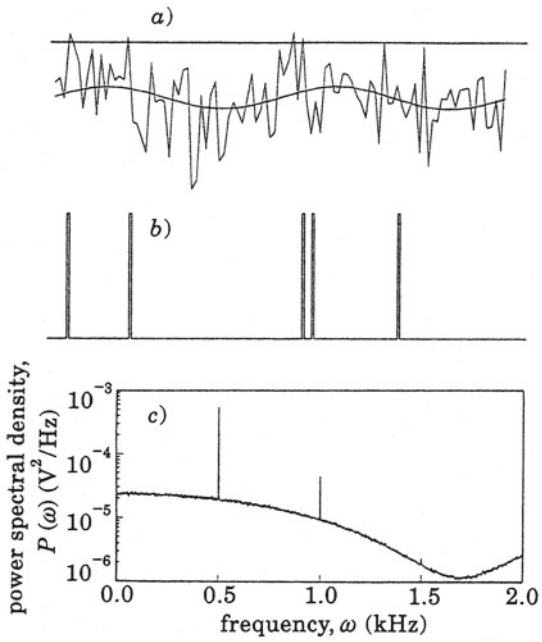


Fig. 11.50 Stochastic resonance. **a** The two curves show the sinusoidal signal and the combination of Gaussian noise plus signal. The latter occasionally exceeds the threshold value shown by the straight line. **b** The pulses generated when the combination of signal plus noise rises above threshold. **c** The averaged power spectrum of the pulse train. (From Gingl et al. 1995. Used by permission)

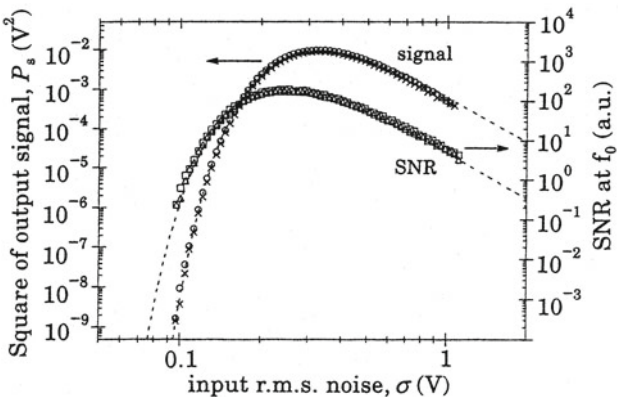


Fig. 11.51 The results of an electronics experiment and a theoretical calculation of threshold detection. One curve shows the square of the output sinusoidal signal, P_s. The other shows the signal-to-noise ratio. (From Gingl et al. 1995. Used by permission)

either from an imbalance of the molecular bombardment of the paddle wheel at temperature T_1, or from molecular bombardment of the pawl spring, which is at temperature T_2. Clockwise rotation will result if the pawl rides up the ramped side of the ratchet and will occur with a probability proportional to $e^{-\Delta U/k_B T_1}$; counterclockwise rotation requires energy transfer to the pawl spring, with a probability proportional to $e^{-\Delta U/k_B T_2}$. With $T_1 = T + \Delta T$ and $T_2 = T - \Delta T$,

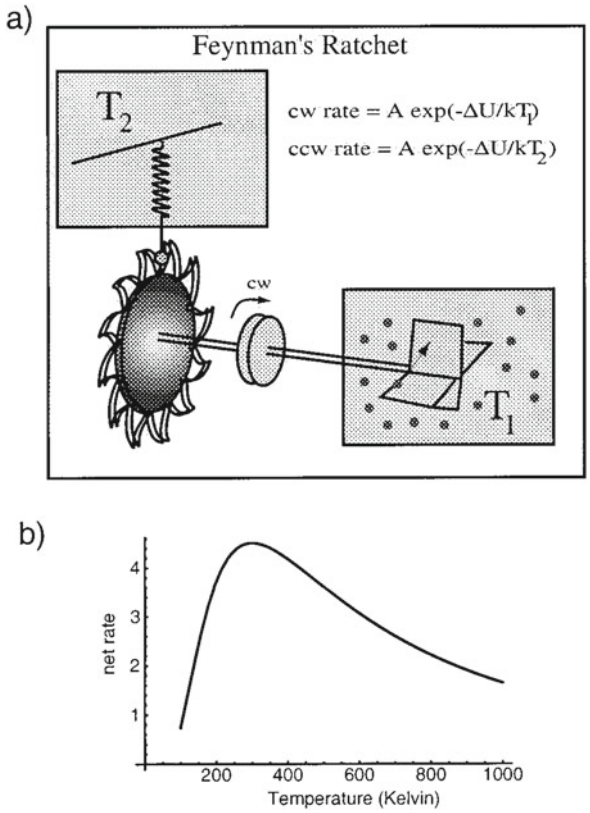


Fig. 11.52 Feynman's Ratchet. **a** A cog wheel is attached to a paddle wheel in a reservoir at temperature T_1. A pawl is attached to a spring located in a reservoir at temperature T_2. **b** The net rate of clockwise motion vs. $T = (T_1 + T_2)/2$. The details are discussed in the text. (Reproduced by permission from Astumian and Moss 1998. Copyright 1998 American Institute of Physics.)

one can show (see Problem 44), that the net rate is

$$\text{net rate} \propto \frac{2\Delta U \Delta T}{k_B T^2} e^{-\Delta U/k_B T}. \tag{11.103}$$

Fig. 11.52b plots the net rate for the parameters $\Delta U = 0.05$ eV and $\Delta T = 10$ K. While thermal gradients are not found in the body, Astumian and Moss show that particles in similar asymmetrically shaped potentials can be driven by having the barrier height vary randomly with time.

Symbols Used in Chap. 11

Symbol	Use	Units	First used page
a	Coefficient in polynomial fit		303
a	Slope		303
a	Coefficient of even (cosine) term		308
a	Parameter in exponential		306
a	Arbitrary constant		323

b	Intercept		303
b	Parameter in exponential		306
b	Coefficient of odd (sine) term		308
e	Noise voltage source	V	333
f, f_0	Frequency	Hz	308
f	Function		335
h	Small quantity		306
h	Shift index		335
i	$\sqrt{-1}$		311
i	Current	A	317
j	Index, usually denoting a data point		303
k	Index denoting terms in a sum		303
k_B	Boltzmann constant	J K^{-1}	332
l, m	Particular values of index k		309
l	Local signal		328
n	Maximum value of index k		308
n	Noise		327
p	Parameter or input signal		330
p	Dimension of a vector		336
s	Signal		327
t	Time	s	308
v	$\log y$		306
v	Voltage	V	317
x	Independent variable		303
x	Vector of data points		336
y	Dependent variable		303
A	Amplitude		308
C, C_k	Amplitude of cosine term		308
C	Capacitance	F	332
G	Gain		330
N	Number of data points		304
Q	Goodness of fit or mean square residual		304
R	Residual		311
R	Resistance	Ω	317
S, S_k	Amplitude of sine term		308
S_{xx}, S_{xy}	Sums of residuals and their products		304
T	Period	s	308
T	Temperature	K	332
U	Energy	J	337
Y, Y_k	Complex Fourier transform or series of y		311
α	Fourier coefficient in autocorrelation function		320
α, β	Fourier coefficients	V Hz$^{-1/2}$	333
δ	Delta function		323
ϕ, θ	Phase		308
ϕ	Correlation function		318
τ	Shift time	s	318
τ_1	Time constant	s	330
ω, ω_0	Angular frequency	s^{-1}	308
Φ_k	Power at frequency $k\omega_0$		317
$\Phi(\omega)$	Power in frequency interval		328
$\Phi'(\omega)$	Energy in frequency interval		324
$\langle\rangle$	Time average		317

Problems

Section 11.1

Problem 1. Find the least squares straight line fit to the following data:

x	y
0	2
1	5
2	8
3	11

Problem 2. Suppose that you wish to pick one number to characterize a set of data $x_1, x_2, \ldots, x_N$. Prove that the mean $\overline{x}$, defined by

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j,$$

minimizes the mean square error

$$Q = \frac{1}{N}\sum_{j=1}^{N}(x_j - \overline{x})^2.$$

Problem 3. Derive Eqs. 11.5.

Problem 4. Suppose that the experimental values $y(x_j)$ are exactly equal to the calculated values plus random noise for each data point: $y(x_j) = y_{\text{calc}}(x_j) + n_j$. What is Q?

Problem 5. You wish to fit a set of data (x_j, y_j) with an expression of the form $y = Bx^2$. Differentiate the expression for Q to find an equation for B.

Problem 6. Assume a dipole **p** is located at the origin and is directed in the xy plane. The z component of the magnetic field, B_z, produced by this dipole is measured at nine points on the surface $z = 50$ mm. The data are

i	x_i (mm)	y_i (mm)	B_{zi} (fT)
1	−50	−50	−154
2	0	−50	−170
3	50	−50	−31
4	−50	0	−113
5	0	0	0
6	50	0	113
7	−50	50	31
8	0	50	170
9	50	50	154

The magnetic field of a dipole is given by Eq. 8.17, which in this case is

$$B_z = \frac{\mu_0}{4\pi}\left[\frac{p_x y_i}{\left(x_i^2 + y_i^2 + z_i^2\right)^{3/2}} - \frac{p_y x_i}{\left(x_i^2 + y_i^2 + z_i^2\right)^{3/2}}\right].$$

Use the method of least squares to fit the data to the equation, and determine p_x and p_y.

Problem 7. Consider the data

x	y
100	4004
101	4017
102	4039
103	4063

(a) Fit these data with a straight line $y = ax + b$ using Eqs. 11.5a and 11.5b to find a.
(b) Use Eq. 11.5c to determine a. Your result should be the same as in part (a).
(c) Repeat parts (a) and (b) while rounding all the intermediate numbers to four significant figures. Do Eqs. 11.5a and 11.5b give the same result as Eq. 11.5c? If not, which is more accurate? To explore more about how numerical errors can creep into computations, see Acton (1990).

Problem 8. This problem is designed to show you what happens when the number of parameters exceeds the number of data points. Suppose that you have two data points:

x	y
0	1
1	4

Find the best fits for one parameter (the mean) and two parameters ($y = ax + b$). Then try to fit the data with three parameters (a quadratic). What happens when you try to solve the equations?

Problem 9. The strength-duration curve for electrical stimulation of a nerve is described by Eq. 7.45: $i = i_R(1+t_C/t)$, where i is the stimulus current, i_R is the rheobase, and t_C is the chronaxie. During an experiment you measure the following data:

t (ms)	i (mA)
0.5	2.004
1.0	1.248
1.5	0.997
2.0	0.879
2.5	0.802
3.0	0.749

Determine the rheobase and chronaxie by fitting these data with Eq. 7.45. Hint: let $a = i_R$ and $b = i_R t_C$, so that the equation is linear in a and b : $i = a+b/t$. Use the linear least squares method to determine a and b. Plot i vs t, showing both the theoretical expression and the measured data points.

Section 11.2

Problem 10.

(a) Obtain equations for the linear least-squares fit of $y = Bx^m$ to data by making a change of variables.
(b) Apply the results of (a) to the case of Problem 5. Why does it give slightly different results?
(c) Carry out a numerical comparison of Problems 5 and (b) with the data points

x	y
1	3
2	12
3	27

Repeat with

x	y
1	2.9
2	12.1
3	27.1

Problem 11. Consider the data given in Problem 2.40 relating molecular weight M and molecular radius R. Assume the radius is determined from the molecular weight by a power law: $R = BM^n$. Fit the data to this expression to determine B and n. Hint: take logarithms of both sides of the equation.

Problem 12. In Prob. 6 the dipole strength and orientation were determined by fitting the equation for the magnetic field of a dipole to the data, using the linear least squares method. In that problem the location of the dipole was known. Now, suppose the location of the dipole (x_0, y_0, z_0) is not known. Derive an equation for $B_z(p_x, p_y, x_0, y_0, z_0)$ in this more general case. Determine which parameters can be found using linear least squares, and which must be determined using nonlinear least squares.

Section 11.4

Problem 13. Write a computer program to verify Eqs. 11.20–11.24.

Problem 14. Consider Eqs. 11.17–11.19 when $n = N$ and show that all equations for $m > N/2$ reproduce the equations for $m < N/2$.

Problem 15. The secretion of the hormone cortisol by the adrenal gland is subject to a 24-h (circadian) rhythm (Guyton 1991). Suppose the concentration of cortisol in the blood, K (in μg per 100 ml) is measured as a function of time, t (in hours, with 0 being midnight and 12 being noon), resulting in the following data:

t	K
0	10.3
4	16.1
8	18.3
12	13.7
16	7.9
20	6.0

Fit these data to the function $K = a + b\cos(2\pi t/24) + c\sin(2\pi t/24)$ using the method of least squares, and determine a, b, and c.

Problem 16. Verify that Eqs. 11.29 follow from Eqs. 11.27.

Problem 17. This problem provides some insight into the fast Fourier transform. Start with the expression for an N-point Fourier transform in complex notation, Y_k in Eq. 11.29a. Show that Y_k can be written as the sum of two $N/2$-point Fourier transforms: $Y_k = \frac{1}{2}\left[Y_k^e + W^k Y_k^o\right]$, where $W = \exp(-i2\pi/N)$, superscript e stands for even values of j, and o stands for odd values.

Problem 18. The following data from Kaiser and Halberg (1962) show the number of spontaneous births vs time of day. Note that the point for 2300–2400 is much higher than for 0000–0100. This is probably due to a bias: if a woman has been in labor for a long time and the baby is born a few minutes after midnight, the birth may be recorded in the previous day. Fit these data with a 24-h period and again including an 8-h period as well. Make a correction for the midnight bias.

Time	Births	Time	Births
0000 – 0100	23,847	1200 – 1300	24,038
0100 – 0200	28,088	1300 – 1400	22,234
0200 – 0300	28,338	1400 – 1500	21,900
0300 – 0400	28,664	1500 – 1600	21,903
0400 – 0500	28,452	1600 – 1700	21,789
0500 – 0600	27,912	1700 – 1800	21,927
0600 – 0700	27,489	1800 – 1900	21,761
0700 – 0800	26,852	1900 – 2000	21,995
0800 – 0900	26,421	2000 – 2100	22,913
0900 – 1000	26,947	2100 – 2200	23,671
1000 – 1100	26,498	2200 – 2300	24,149
1100 – 1200	25,615	2300 – 2400	27,819

Problem 19. Calculate the discrete Fourier transform of the data $y_i = 0.00, 0.25, 0.50, 0.75, 0.50, 0.25$ using Eq. 11.26.

Section 11.5

Problem 20. Let $y(t)$ be a periodic function with period T:

$$y(t) = t, \quad 0 < t < T.$$

(a) Plot $y(t)$ over the range $-2T < t < 2T$.
(b) Use Eqs. 11.30 and 11.34 to calculate the Fourier series for $y(t)$.
(c) Plot the Fourier series using only the term $k = 0$, then using $k = 0$ and $k = 1$, and finally $k = 0, k = 1$ and $k = 2$. Compare these plots to the plot in part (a).

Problem 21. Let $y(t)$ be a periodic function with period T:

$$y(t) = sin(\pi t/T), \quad 0 < t < T.$$

(a) Plot $y(t)$ over the range $-2T < t < 2T$.
(b) Use Eqs. 11.30 and 11.34 to calculate the Fourier series for $y(t)$.
(c) Plot the Fourier series using only the term $k = 0$, then using $k = 0$ and $k = 1$, and finally $k = 0, k = 1$ and $k = 2$. Compare these plots to the plot in part (a).

Problem 22. Use Eqs. 11.34 to derive Eq. 11.36.

Section 11.6

Problem 23. Calculate the power spectrum for the function given in Problem 20.

Section 11.7

Problem 24. Suppose that $y(x,t) = y(x - vt)$. Calculate the cross correlation between signals $y(x_1)$ and $y(x_2)$.

Problem 25. Calculate the cross-correlation, ϕ_{12}, for the example in Fig. 11.21:

$$y_1(t) = \begin{cases} +1, & 0 < t < T/2 \\ -1, & T/2 < t < T \end{cases}$$

$$y_2(t) = \sin\left(\frac{2\pi t}{T}\right).$$

Both functions are periodic.

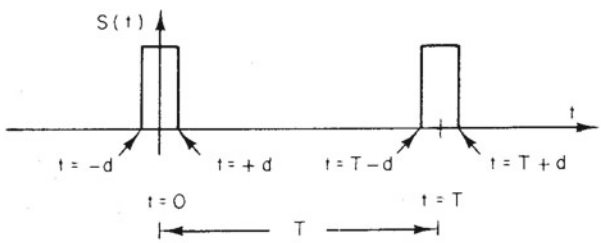
Problem 26. Suppose you measure some noisy signal every hour for several weeks. Explain how you could use the autocorrelation function to search for a *circadian rhythm*: a component of the signal that varies with a period of one day.

Section 11.8

Problem 27. Fill in the missing steps to show that the autocorrelation of $y_1(t)$ is given by Eq. 11.51.

Problem 28. Consider a square wave of amplitude A and period T.
(a) What are the coefficients in a Fourier-series expansion?
(b) What is the power spectrum?
(c) What is the autocorrelation of the square wave?
(d) Find the Fourier-series expansion of the autocorrelation function and compare it to the power spectrum.

Problem 29. The series of pulses shown are an approximation for the concentration of follicle-stimulating hormone (FSH) released during the menstrual cycle.



(a) Determine a_0, a_k, and b_k in terms of d and T.
(b) Sketch the autocorrelation function.
(c) What is the power spectrum?

Problem 30. Consider the following simplified model for the periodic release of follicle-stimulating hormone (FSH). At $t = 0$ a substance is released so the plasma concentration rises to value C_0. The substance is cleared so that $C(t) = C_0 e^{-t/\tau}$. Thereafter the substance is released in like amounts at times $T, 2T$, and so on. Furthermore, $\tau \ll T$.

(a) Plot $C(t)$ for two or three periods.

(b) Find general expressions for a_0, a_k, and b_k. Use the fact that integrals from 0 to T can be extended to infinity because $\tau \ll T$. Use the following integral table:

$$\int_0^\infty e^{-ax}\,dx = \frac{1}{a},$$

$$\int_0^\infty e^{-ax}\cos mx\,dx = \frac{a}{a^2+m^2},$$

$$\int_0^\infty e^{-ax}\sin mx\,dx = \frac{m}{a^2+m^2}.$$

(c) What is the "power" at each frequency?

(d) Plot the "power" for $k = 1, 10, 100$ for two cases: $\tau/T = 0.1$ and 0.01. Compare the results to the results of Problem 29.

(e) Discuss qualitatively the effect that making the pulses narrower has on the power spectrum. Does the use of Fourier series seem reasonable in this case? Which description of the process is easier—the time domain or the frequency domain?

(f) It has sometimes been said that if the transform for a given frequency is written as $A_k \cos(k\omega_0 t - \phi_k)$ that ϕ_k gives timing information. What is ϕ_1 in this case? ϕ_2? Do you agree with the statement?

Problem 31. Calculate the autocorrelation function and the power spectrum for the previous problem.

Section 11.9

Problem 32. Calculate the Fourier transform of $\exp[-(at)^2]$ using complex notation (Eq. 11.59). Hint: complete the square.

Problem 33. Figure 11.24 implies that two different functions can have the same autocorrelation, so that taking the autocorrelation is a one-way process. Show this by calculating the autocorrelation of $A\cos(\omega t)$ and comparing it to the autocorrelation of $A\sin(\omega t)$ given in Eq. 11.49.

Section 11.10

Problem 34. Prove that

$$\delta(t) = \delta(-t),$$

$$t\,\delta(t) = 0,$$

$$\delta(at) = \frac{1}{a}\delta(t).$$

Section 11.11

Problem 35. Rewrite Eqs. 11.61 in terms of an amplitude and a phase. Plot them.

Problem 36. Find the Fourier transform of

$$f(t) = \begin{cases} 1, & -a \le t \le a, \\ 0, & \text{everywhere else.} \end{cases}$$

Problem 37. Find the Fourier transform of

$$y = \begin{cases} e^{-at}\sin\omega_0 t, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

Determine $C(\omega)$, $S(\omega)$, and $\Phi'(\omega)$ for $\omega > 0$ if the term that peaks at negative frequencies can be ignored for positive frequencies.

Section 11.14

Problem 38. Here are some data.

t	y	t	y	t	y
2	1.39	14	5.01	26	0.91
3	0.67	15	0.75	27	1.32
4	−1.38	16	0.90	28	1.92
5	−0.76	17	−0.42	29	0.57
6	5.23	18	3.68	30	2.30
7	1.31	19	4.15	31	1.09
8	2.63	20	1.45	32	−0.71
9	1.03	21	−2.44	33	−1.72
10	4.62	22	4.44	34	4.22
11	1.98	23	−0.08	35	3.20
12	0.47	24	2.34	36	1.69

(a) Plot them.

(b) If you are told that there is a signal in these data with a period of 4 s, you can group them together and average them. This is equivalent to taking the cross correlation with a series of δ functions. Estimate the signal shape.

Section 11.15

Problem 39. Verify that Eqs. 11.80 and 11.81 are solutions of Eq. 11.79.

Problem 40. Equation 11.81 is plotted on log–log graph paper in Fig. 11.43. Plot it on linear graph paper.

Problem 41. If the frequency response of a system were proportional to $1/\left[1 + (\omega/\omega_0)^3\right]$, what would be the high frequency roll-off in decibels per octave for $\omega \gg \omega_0$?

Problem 42. Consider a signal $y = A\cos\omega t$. What is the time derivative? For a fixed value of A, how does the derivative compare to the original signal as the frequency is increased? Repeat these considerations for the integral of $y(t)$.

Section 11.16

Problem 43. Show that integration of Eq. 11.102 over all shift times is consistent with the integration of the δ function that is obtained in the limit $\tau_1 \to 0$.

Section 11.18

Problem 44. Show that the net clockwise rate of rotation of the Feynman ratchet is given by Eq. 11.103.

References

Acton FS (1990) Numerical methods that work. Mathematical Society of America, Washington DC

Adair RK, Astumian RD, Weaver JC (1998) Detection of weak electric fields by sharks, rays and skates. Chaos 8(3):576–587

Adair EC, Hobbie SE, Hobbie RK (2010) Single-pool exponential decomposition models: potential pitfalls in their use in ecological studies. Ecology 91(4):1225–1236

Anderka M, Declercq ER, Smith W (2000) A time to be born. Am J Pub Health 90(1):124–126

Astumian RD (1997) Thermodynamics and kinetics of a Brownian motor. Science 276:917–922

Astumian RD, Moss F (1998) Overview: the constructive role of noise in fluctuation driven transport and stochastic resonance. Chaos 8(3):533–538

Bevington PR, Robinson DK (2003) Data reduction and error analysis for the physical sciences, 3rd edn. McGraw-Hill, New York

Blackman RB, Tukey JW (1958) The measurement of power spectra. Dover, New York, pp 32–33

Bracewell RN (1990) Numerical transforms. Science 248:697–704

Bracewell RN (2000) Fourier transform and its applications, 3rd edn. McGraw-Hill, Boston

Cohen A (2006) Biomedical signals: origin and dynamic characteristics; frequency-domain analysis. In Bronzino JD (ed) The biomedical engineering handbook, vol 2, 3rd edn. CRC, Boca Raton, pp 1-1–1–22

Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comput 19:297–301

DeFelice LJ (1981) Introduction to membrane noise. Plenum, New York

Feynman RP, Leighton RB, Sands M (1963) The Feynman lectures on physics, vol 1, Chap 46. Addison-Wesley, Reading

Gammaitoni L, Hänggi P, Jung P, Marchesoni F (1998) Stochastic resonance. Rev Mod Phys 70(1):223–287

Gingl Z, Kiss LB, Moss F (1995) Non-dynamical stochastic resonance: theory and experiments with white and arbitrarily coloured noise. Europhys Lett 29(3):191–196

Glass L (2001) Synchronization and rhythmic processes in physiology. Nature 410(825):277–284

Guyton AC (1991) Textbook of medical physiology, 8th edn. Saunders, Philadelphia

Hämäläinen M, Harri R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV (1993) Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys 65(2):413–497

Kaiser IH, Halberg F (1962) Circadian periodic aspects of birth. Ann N Y Acad Sci 98:1056–1068

Kaplan D, Glass L (1995) Understanding nonlinear dynamics. Springer, New York

Lighthill, MJ (1958) An introduction to Fourier analysis and generalised functions. Cambridge University Press, Cambridge

Lybanon, M (1984) A better least-squares method when both variables have uncertainties. Am J Phys 52: 22–26

Mainardi, LT, Bianchi AM, Cerutti S (2006) Digital biomedical signal acquisition and processing. In: Bronzino JD (ed) The biomedical engineering handbook, vol 2, 3rd edn. CRC, Boca Raton, pp 2-1–2-24

Maughan WZ, Bishop CR, Pryor TA, Athens JW (1973) The question of cycling of the blood neutrophil concentrations and pitfalls in the statistical analysis of sampled data. Blood 41:85–91

Milnor WR (1972) Pulsatile blood flow. N Eng J Med 287:27–34

Nedbal L, Březina V (2002) Complex metabolic oscillations in plants forced by harmonic irradiance. Biophys J 83:2180–2189

Nyquist H (1928) Thermal agitation of electric charge in conductors. Phys Rev 32:110–113

Orear J (1982) Least squares when both variables have uncertainties. Am J Phys 50:912–916

Packard GC (2009) On the use of logarithmic transformations in allometric analyses. J Theor Biol 257:515–518

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, New York (reprinted with corrections, 1995)

Visscher PB (1996) The FFT: Fourier transforming one bit at a time. Comput Phys 10(5):438–443

Wiesenfeld K, Jaramillo F (1998) Minireview of stochastic resonance. Chaos 8(3):539–548

Images

12

Images are very important in the remainder of this book. They may be formed by the eye, a camera, an x-ray machine, a nuclear medicine camera, magnetic resonance imaging, or ultrasound. The concepts developed in Chap. 11 can be used to understand and describe image quality. The same concepts are also used to reconstruct computed tomographic or magnetic resonance images of the body. A very complete, advanced mathematical treatment of all kinds of images is found in a 1500-page book by Barrett and Myers (2004). A history of medical imaging has been written by Kevles (1997).

The convolution integral of Sect. 12.1 shows how the response of a linear system can be related to the input to the system and the impulse (δ-function) response of the system. It forms the basis for the rest of the chapter. The Fourier-transform properties of the convolution are also described in this section. Section 12.2 introduces quantitative ways to relate the image to the object, using the techniques developed in Chap. 11 to describe the blurring that occurs. Section 12.3 shows the importance of different spatial frequencies in an image and their effect on the quality of the image.

Sections 12.4 and 12.5 pose the fundamental problem of reconstructing slices from projections and introduce two techniques for solving it: the Fourier transform and filtered back projection. Section 12.6 provides a numerical example of filtered back projection for a circularly symmetric object.

This chapter is quite mathematical. The key understanding to take from it is the relationship between spatial frequencies and image quality in Sect. 12.3.

12.1 The Convolution Integral and Its Fourier Transform

12.1.1 One Dimension

We now apply the techniques developed in Chap. 11 to describe the formation of images. An image is a function of position, usually in two dimensions at an *image plane*. We start with the simpler case of an image extending along a line. Functions of time are easier to think about, so let us imagine a one-dimensional example that is a function of time: a high-fidelity sound system. A hi-fi system is (one hopes) *linear*, which means that the relationship between the output response and a complicated input can be written as a superposition of responses to more elementary input functions. The output might be the instantaneous air pressure at some point in the room; the input might be the air pressure at a microphone or the magnetization on a strip of tape.

It takes a certain amount of time for the signal to propagate through the system. In the simplest case the response at the ear would exactly reproduce the response at the input a very short time earlier. In actual practice the response at time t may depend on the input at a number of earlier times, because of limitations in the electronic equipment or echoes in the room. If the entire system is linear, the output $g(t)$ can be written as a superposition integral, summing the weighted response to inputs at other times. If $f(t')$ is the input and h is the weighting, the output $g(t)$ is

$$g(t) = \int_{-\infty}^{\infty} f(t')h(t,t')\,dt'. \tag{12.1}$$

Variable t' is a dummy variable. The integration is over all values of t' and it does not appear in the final result, which depends only on the functional forms of f and h. Note also that if f and g are expressed in the same units, then h has the dimensions of s^{-1}.

If input f is a δ function at time t'_0, then

$$g(t) = \int_{-\infty}^{\infty} \delta(t'-t'_0)h(t,t')\,dt' = h(t,t'_0). \tag{12.2}$$

We see that $h(t,t')$ is the *impulse response* of the system to an impulse at time t'. If the impulse response of a linear system is known, it is possible to calculate the response to any arbitrary input.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_12, © Springer International Publishing Switzerland 2015

If, in addition to being linear, the system responds to an impulse the same way regardless of when it occurs, the system is said to be *stationary*. In the hi-fi example, this means that no one is adjusting the volume or tone controls. For a stationary system the impulse response depends only on the *time difference* $t - t'$:

$$h(t, t') = h(t - t'), \tag{12.3}$$

and the superposition integral takes the form

$$g(t) = \int_{-\infty}^{\infty} f(t')h(t - t')\, dt'. \tag{12.4a}$$

This is called the *convolution integral*. It is often abbreviated as

$$g(t) = f(t) \otimes h(t). \tag{12.4b}$$

For the hi-fi system the function $h(t - t')$ is zero for all t' larger (later) than t; the response does not depend on future inputs. For the images we will be considering shortly, where the variables represent positions in the object and image, h can exist for negative arguments.

We saw an example of the impulse response in Sect. 11.15, where we found that the solution of the differential equation for the system was a step exponential, Eq. 11.84. For that simple linear system we can write

$$h(t - t') = \begin{cases} 0, & t < t' \\ (1/\tau_1)e^{-(t-t')/\tau_1}, & t > t'. \end{cases} \tag{12.5}$$

We have seen superposition integrals before: for one-dimensional diffusion (Eq. 4.73) and for the potential (Eq. 7.21) and magnetic field (Eq. 8.14) outside a cell.

There is an important relationship between the Fourier transforms of the functions appearing in the convolution integral, which was hinted at in Sect. 11.15. If the sine and cosine transforms of function h are denoted by $C_h(\omega)$ and $S_h(\omega)$, with similar notation for f and g, the relationships can be written

$$\begin{aligned} C_g(\omega) &= C_f(\omega)C_h(\omega) - S_f(\omega)S_h(\omega), \\ S_g(\omega) &= C_f(\omega)S_h(\omega) + S_f(\omega)C_h(\omega). \end{aligned} \tag{12.6a}$$

This is called the *convolution theorem*. If we were using complex exponential notation, the Fourier transforms would be related by

$$G(\omega) = F(\omega)H(\omega). \tag{12.6b}$$

The convolution of two functions in time is equivalent to multiplying their Fourier transforms.

Equations 12.6a are similar to the addition formulas for sines and cosines, which are of course used in the derivation. To derive them, we take the Fourier transforms of f and h:

$$f(t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[C_f(\omega)\cos\omega t' + S_f(\omega)\sin\omega t'\right] d\omega$$

$$h(t - t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[C_h(\omega)\cos\omega(t - t') + S_h(\omega)\sin\omega(t - t')\right] d\omega.$$

Then

$$\begin{aligned} g(t) &= \int_{-\infty}^{\infty} f(t')h(t - t')\, dt' \\ &= \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} dt' \left[\int_{-\infty}^{\infty} d\omega \left[C_f(\omega)\cos\omega t' + S_f(\omega)\sin\omega t'\right]\right. \\ &\left. \times \int_{-\infty}^{\infty} d\omega' \left[C_h(\omega')\cos\omega'(t - t') + S_h(\omega')\sin\omega'(t - t')\right]\right]. \end{aligned}$$

We can use the trigonometric addition formulas and the fact that $\sin(-\omega' t') = -\sin\omega' t'$ to rewrite and expand this expression, much as we did in the last chapter. Carrying out the integration over t' first and using the properties of integrals of the δ function gives

$$\begin{aligned} g(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \left[C_f(\omega)C_h(\omega) - S_f(\omega)S_h(\omega)\right]\cos\omega t \\ &+ \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \left[C_f(\omega)S_h(\omega) + S_f(\omega)C_h(\omega)\right]\sin\omega t. \end{aligned}$$

Comparison of this with Eqs. 11.57 proves Eq. 12.6a.

Fourier techniques need not be restricted to frequency and time. The quality and resolution of the image on the retina, an x-ray film, or a photograph are best described in terms of *spatial frequency*. The distance across the image in some direction is x, and a sinusoidal variation in the image would have the form $A(\lambda)\sin(2\pi x/\lambda - \phi)$. The spatial frequency, $1/\lambda$, is the number of cycles per unit length and is expressed in cycles per meter or cycles per millimeter. The *wave number* or *angular wave number is* $k = 2\pi/\lambda$, where λ is the wavelength. We can write the variation as $A(k)\sin(kx - \phi)$.

12.1.2 Two Dimensions

The convolution and Fourier transform in two dimensions are needed to analyze the response of a system that forms a two-dimensional image of a two-dimensional object. The object can be represented by function $f(x', y')$ in the *object plane*. The image is given by a function $g(x, y)$ in the image plane:

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x', y')h(x, x'; y, y')\, dx'dy'. \tag{12.7}$$

If the contribution of object point (x', y') to the image at (x, y) depends only on the relative distances $x - x'$ and $y - y'$, then the two-dimensional impulse response is $h(x - x', y - y')$, and the image is obtained by the two-dimensional convolution

$$g(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x', y')h(x - x', y - y')\, dx'dy' \tag{12.8a}$$

or

$$g(x, y) = f(x, y) \otimes\otimes h(x, y). \tag{12.8b}$$

The Fourier transform in two dimensions is defined by

$$\begin{aligned} f(x, y) = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} dk_x \\ \times \int_{-\infty}^{\infty} dk_y [C(k_x, k_y)\cos(k_x x + k_y y) \\ + S(k_x, k_y)\sin(k_x x + k_y y)]. \end{aligned} \tag{12.9a}$$

The coefficients are given by

$$C(k_x, k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y)\cos(k_x x + k_y y), \tag{12.9b}$$

$$S(k_x, k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y)\sin(k_x x + k_y y). \tag{12.9c}$$

The Fourier transforms of the functions in the convolution are related by equations similar to those for the one-dimensional convolution.

$$\begin{aligned} C_g(k_x, k_y) &= C_f(k_x, k_y)C_h(k_x, k_y) \\ &\quad - S_f(k_x, k_y)S_h(k_x, k_y), \\ S_g(k_x, k_y) &= C_f(k_x, k_y)S_h(k_x, k_y) \\ &\quad + S_f(k_x, k_y)C_h(k_x, k_y). \end{aligned} \tag{12.10}$$

With complex notation we would define the two-dimensional Fourier transform pair by

$$\begin{aligned} F(k_x, k_y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-i(k_x x + k_y y)}\, dxdy, \\ f(x, y) &= \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(k_x, k_y)e^{i(k_x x + k_y y)}dk_x dk_y, \end{aligned} \tag{12.11a}$$

and the convolution theorem would be

$$G(k_x, k_y) = F(k_x, k_y)H(k_x, k_y). \tag{12.11b}$$

12.2 The Relationship Between the Object and the Image

12.2.1 Point Spread Function

Suppose that an object in the $x'y'$ plane is described by a function $L(x', y')$ that varies from place to place on the object. The image is

$$E_{\text{image}}(x, y) = \iint L(x', y')h(x, y; x', y')\, dx'dy'. \tag{12.12}$$

Function h is called the *point spread function*. The point spread function tells how information from a point source at (x', y') spreads out over the image plane. It receives its name from the following. If we imagine that the object is a point described by $L(x', y') = L\delta(x' - x'_0)\delta(y' - y'_0)$, then integration shows that

$$E_{\text{image}} = h(x, y; x'_0, y'_0).$$

The point spread function has the same functional form as the image from a point source, just as did the impulse response in one dimension.

You can verify that the point spread function for an ideal imaging system with magnification m is

$$h(x, y; x', y') = m^2\delta(x - mx')\delta(y - my'). \tag{12.13}$$

The δ functions pick out the values $(x' = x/m, y' = y/m)$ in the object plane to contribute to the image at (x, y). You can make the verification by substituting Eq. 12.13 in Eq. 12.12 and using the properties of the δ function from Eq. 11.64.

This discussion assumes that *intensities* add. This is true when the oscillations of the radiant energy (such as the electric field for light waves) have random phases lasting for a time short compared to the measurement time. Such radiant energy is called *incoherent*.[1]

We have already seen that when the impulse response in a one-dimensional system depends on coordinate differences such as $t - t'$ (or $x - x'$ or $x - mx'$), the system is stationary. In this case it is also said to be *space invariant*: changing the position of the object changes the position of the image but not its functional form. Stationarity is easier to obtain in a system such as a hi-fi system than in an imaging system, but we usually assume that it holds in an imaging system as well. For a space-invariant system

$$E_{\text{image}}(x, y) = \iint L(x', y')h(x - mx', y - my')\, dx'\, dy'. \tag{12.14}$$

[1] These arguments also work for coherent radiation, where the phases are important, but the point spread function is for the amplitude of the wave instead of the square of the amplitude (intensity). The calculation then gives rise to interference and diffraction effects.

This is a two-dimensional convolution. The convolution theorem is

$$\begin{aligned} C_{\text{image}}(k_x, k_y) &= C_{\text{object}}(k_x, k_y)C_h(k_x, k_y) \\ &\quad - S_{\text{object}}(k_x, k_y)S_h(k_x, k_y), \\ S_{\text{image}}(k_x, k_y) &= C_{\text{object}}(k_x, k_y)S_h(k_x, k_y) \\ &\quad + S_{\text{object}}(k_x, k_y)C_h(k_x, k_y). \end{aligned} \tag{12.15}$$

12.2.2 Optical, Modulation, and Phase Transfer Functions

The *optical transfer function* (OTF) is the Fourier transform of the point spread function, $C_h(k_x, k_y)$ and $S_h(k_x, k_y)$. It is analogous to the transfer function for an amplifier (Sect. 11.15). The *modulation transfer function* (MTF) is the amplitude of the OTF:

$$\text{MTF}(k_x, k_y) = \left[C_h^2(k_x, k_y) + S_h^2(k_x, k_y)\right]^{1/2}. \tag{12.16}$$

The *phase transfer function* is

$$\text{PTF}(k_x, k_y) = \tan^{-1}\left(\frac{S_h(k_x, k_y)}{C_h(k_x, k_y)}\right). \tag{12.17}$$

Often the transfer functions are normalized by dividing them by their value at zero spatial frequency.

The modulation transfer function can be measured by using a set of objects for which L varies sinusoidally at different spatial frequencies. The property L cannot be negative and must be offset by a zero-frequency component:

$$L(x, y) = a + b\cos(k_x x + k_y y), \quad 0 < b < a. \tag{12.18}$$

The image is described by

$$\begin{aligned} E &= \text{MTF}(0, 0)a \\ &\quad + \text{MTF}(k_x, k_y)b\cos\left[k_x x + k_y y + \phi(k_x, k_y)\right]. \end{aligned} \tag{12.19}$$

The *modulation* of the object is defined to be

$$(\text{modulation}) = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} = \frac{(a+b) - (a-b)}{(a+b) + (a-b)} = \frac{b}{a}. \tag{12.20}$$

A similar expression defines the modulation of the image. The modulation transfer function is the ratio of the modulation of the image divided by the modulation of the object. The phase of the optical transfer function describes shifts of the phase of the image at each angular frequency along the appropriate axis. It is fully as important as the amplitude, since it describes the evenness or oddness of the image about its stated origin.

The modulation transfer function of an ideal system would be flat for all spatial frequencies. However, there is an upper limit imposed by diffraction, if nothing else. Figure 12.1 shows the point spread function and MTF for a diffraction-limited case. Figure 12.2 shows three possible modulation transfer functions for an imaging system. The upper one represents the diffraction limit. It has the same general shape as in Fig. 12.1. Curves 1 and 2 might be for real systems. While the second system transmits more of the highest spatial frequencies, it transmits less of the midrange frequencies, and its image would not have as much "punch" as the first system. Figure 12.3 shows the modulation transfer functions of several photographic films, with (a) being the most sensitive and (e) the least sensitive but with the highest resolution. Photographers are well aware of the trade-off between speed and resolution in film. Fast films are "more grainy" than slow films.

A complex imaging system may have several components, just as the hi-fi system did. If the system is linear, the modulation transfer function for the combined system is the product of the modulation transfer functions for each component. The optical transfer functions combine according to equations like Eq. 12.10.

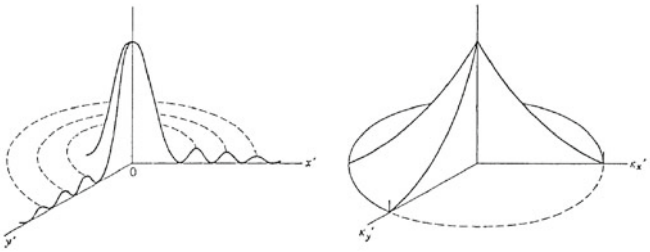


Fig. 12.1 The point spread function and modulation transfer function for a diffraction-limited circular aperture. (Source: Williams and Becklund 1972). Used by permission of the authors

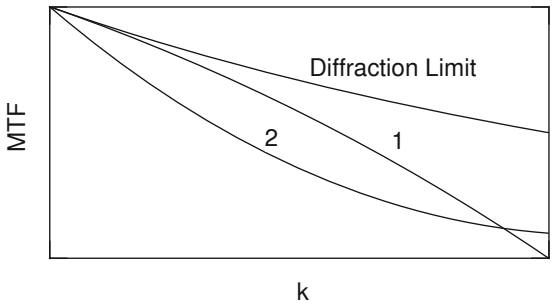


Fig. 12.2 Three possible modulation transfer functions. The *top* one is the diffraction limit for monochromatic light. (Compare it with Fig. 12.1.) Curve 2 is higher than curve 1 at the highest value of k shown, but an image produced by system 2 would not have as much "punch." It has less content at the middle spatial frequencies

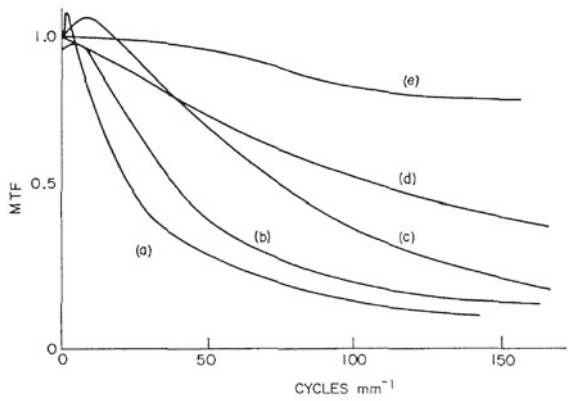


Fig. 12.3 Some representative modulation transfer functions for various photographic films, showing how the resolution decreases as the film sensitivity increases. Film (*a*) has the greatest sensitivity and worst resolution. Film (*e*) is the least sensitive ("slowest") and has the best resolution. (With permission from Shaw 1979)

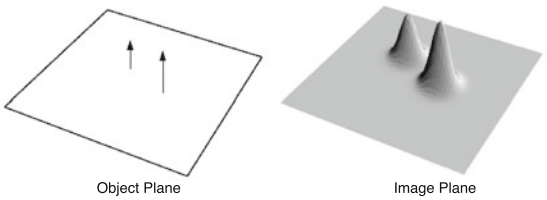


Fig. 12.4 The point spread function. Two impulse sources of different heights are shown in the object plane. The response to them is shown in the image plane

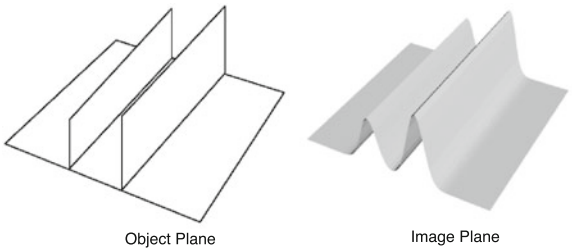


Fig. 12.5 The line spread function. Two line sources are shown in the object plane. The response to them is shown in the image plane

12.2.3 Line and Edge Spread Functions

The *line spread function* is the response of a system to a line object in the object plane. In general, the system is not isotropic and the line spread function depends on the orientation of the line. The Fourier transform of the line spread function along the y axis is $C_h(k_x, 0)$ and $S_h(k_x, 0)$. Figure 12.4 shows a geometrical interpretation of the point spread function. Figure 12.5 shows the line spread function. The *edge spread function* is the response to an object that is a step function. All of these functions are interrelated. A discussion of how one can be obtained from another is found in many places, including Chap. 9 of Gaskill (1978).

12.3 Spatial Frequencies in an Image

There are some universal relationships between the spatial frequencies present in an image and the character of the image. These relationships hold whether the image is a photograph, an x-ray film, a computed tomographic scan, an ultrasound or nuclear medicine image, or a magnetic resonance image. In this section we describe these general relationships, which we will use throughout the rest of the book.

The first general relationship concerns the size of an image and the lowest spatial frequency present. For simplicity, consider the x direction and the corresponding spatial frequencies k. The *object* is nonperiodic. But its *image* is represented by a Fourier series which has period L. We saw in Chap. 11 that if the lowest angular frequency present is ω_0, the period is $T = 2\pi/\omega_0$. The lowest spatial frequency present (other than zero) is $k_0 = 2\pi/L$. The series has harmonics with separation $\Delta k = k_0$. This leads to the fundamental relationship

$$L = \frac{2\pi}{k_0}. \tag{12.21}$$

The lowest spatial frequency present (which equals the separation of the spatial frequencies) is related to the size of the image L (the "field of view" or FOV).

The second general relationship concerns the spatial resolution in an image and the highest spatial frequency present. If the image has N discrete samples, then the sampling interval or spatial resolution is $\Delta x = L/N$. This allows (or requires) the determination of $N/2$ cosine coefficients and $N/2$ sine coefficients (see Sect. 11.4). The highest spatial frequency present is $k_{\max} = N\Delta k/2$. We obtain

$$\Delta x = \frac{L}{2}\frac{\Delta k}{k_{\max}} = \frac{\pi}{k_{\max}}. \tag{12.22}$$

The spatial resolution is inversely proportional to the highest spatial frequency present. As we saw for the Fourier series representing a square wave, the higher harmonics give fine detail and sharpness to the image.

To reiterate: *The lowest spatial frequency in the image determines the field of view. The lower the minimum spatial frequency, the larger the field of view. The highest spatial frequency in the image determines the resolution. The higher the maximum spatial frequency, the finer the resolution.*

Here are a number of pictures that show how changing the coefficients in certain regions of k space affect an image. Figure 12.6b shows a transverse scan of a head by magnetic resonance imaging. This is a normal image to compare with the following figures. It consists of 256 samples in each direction or 256 × 256 pixels. The magnitude of its Fourier transform is shown in Fig. 12.6a. Figure 12.7 shows the cosine and sine coefficients in the expansion.

Figures 12.8 and 12.9 show what happens when the high-frequency Fourier components are removed. In the first case

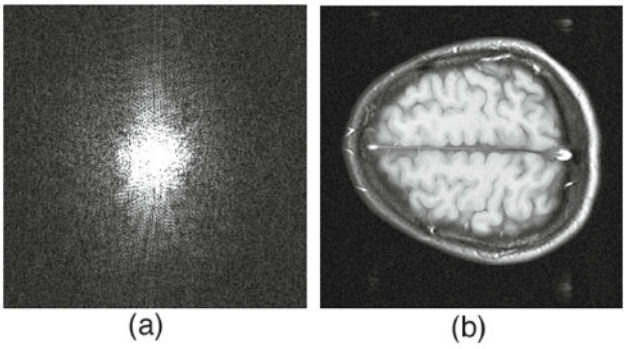


Fig. 12.6 A magnetic resonance imaging head scan: **a** The squared amplitude $C^2 + S^2$ in k space. **b** The image. This is a normal image to compare with the following figures. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

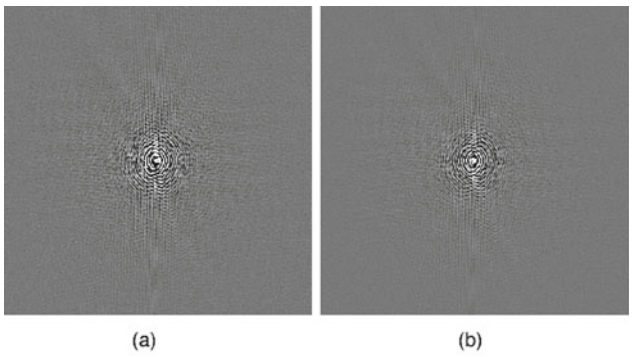


Fig. 12.7 The sine and cosine coefficients for the image in Fig. 12.6. **a** $C(k_x, k_y)$. **b** $S(k_x, k_y)$. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

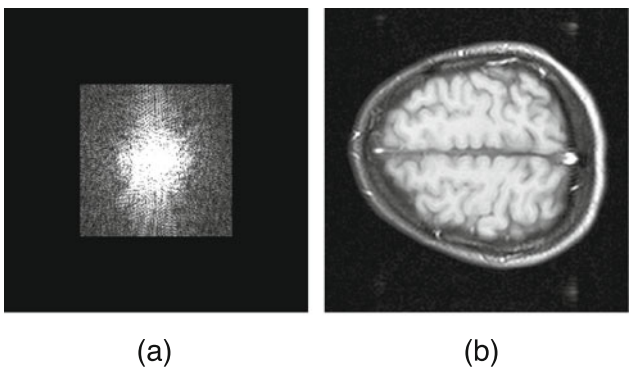


Fig. 12.8 The image that results when the high-frequency Fourier components above $k_{x\,\max}/2$ and $k_{y\,\max}/2$ are removed. Note the blurring compared to Fig. 12.6. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

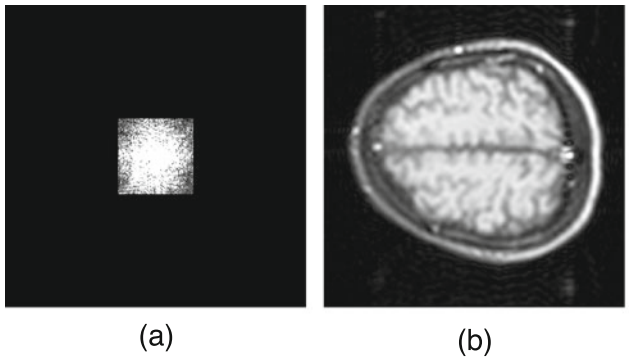


Fig. 12.9 The image that results when the high-frequency Fourier components above $k_{x\,\max}/4$ and $k_{y\,\max}/4$ are removed. The blurring is even greater. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

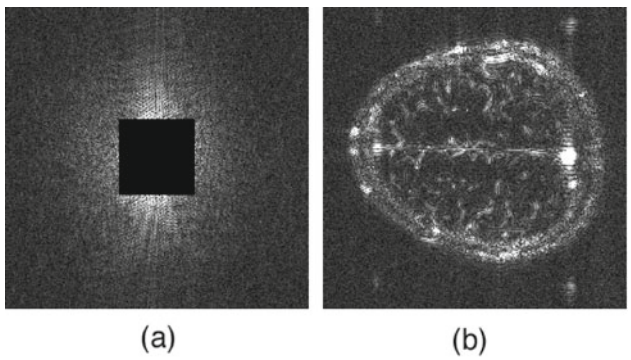


Fig. 12.10 The image that results when the low-frequency Fourier components below $k_{x\,\max}/4$ and $k_{y\,\max}/4$ are removed. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

they have been removed above $k_{x\,\max}/2$ and $k_{y\,\max}/2$. In the second they are removed above $k_{x\,\max}/4$ and $k_{y\,\max}/4$. Compare the blurring in these figures with the original image.

When the low-frequency coefficients are set to zero as in Fig. 12.10, only the high-frequency edges remain. In this case the Fourier components below $k_{x\,\max}/4$ and $k_{y\,\max}/4$ have been set to zero. (Keeping the same values of $k_{x\,\max}$ and Δk and removing the information on those coefficients keeps the field of view the same.)

Figure 12.11 shows the artifact that results from setting every other Fourier coefficient to zero: "ghost" images (Buonocore and Gao 1977). In the first case alternate Fourier coefficients have been removed in k_x space; in the second they have been removed in both k_x and k_y.

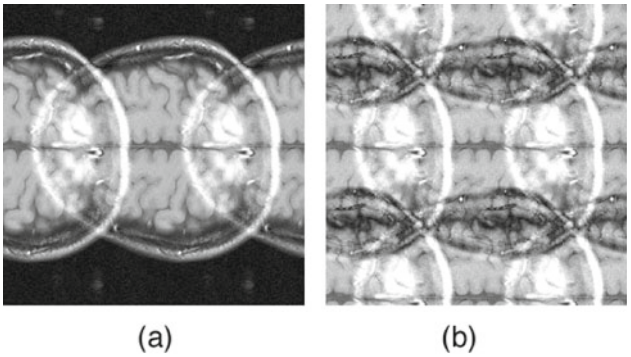


Fig. 12.11 The Fourier coefficients for every other value of k have been set to zero which leads to ghost images. **a** Every other value of k_x has been removed. **b** Every other value of both k_x and k_y has been removed. Prepared by Mr. Tuong Huu Le, Center for Magnetic Resonance Research, University of Minnesota. Thanks also to Prof. Xiaoping Hu

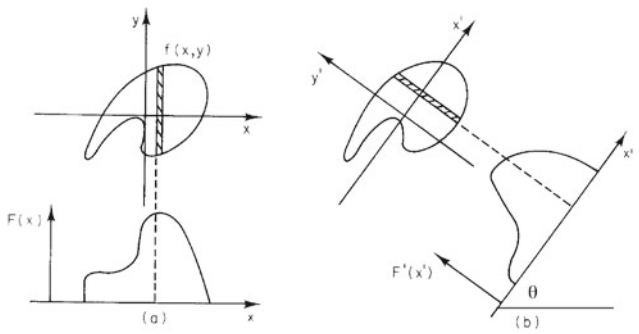


Fig. 12.12 **a** Function $F(x)$ is the integral of $f(x, y)$ over all y. **b** The scan is repeated at angle with the x axis

12.3.1 Summary

In summary: *The lowest spatial frequency in the image determines the field of view. The lower the minimum spatial frequency, the larger the field of view. Low spatial frequencies provide shape, contrast, and brightness.*

The highest spatial frequency in the image determines the resolution. The higher the maximum spatial frequency, the finer the resolution. High spatial frequencies provide resolution, edges, and sharp detail.

12.4 Two-Dimensional Image Reconstruction from Projections by Fourier Transform

The reconstruction problem can be stated as follows. A function $f(x, y)$ exists in two dimensions. Measurements are made that give *projections*: the integrals of $f(x, y)$ along various lines as a function of displacement perpendicular to each line. For example, integration parallel to the y axis gives a function of x,

$$F(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \tag{12.23}$$

as shown in Fig. 12.12. The scan is repeated at many different angles θ with the x axis, giving a set of functions $F(\theta, x')$, where x' is the distance along the axis at angle θ with the x axis. The problem is to reconstruct $f(x, y)$ from the set of functions $F(\theta, x')$. Several different techniques can be used. A detailed reference is the book by Cho et al. (1993).

We will consider two of these techniques: reconstruction by Fourier transform, where the Fourier coefficients are obtained from projections (in this section), and filtered back projection (Sect. 12.5).

The Fourier transform technique is easiest to understand. Consider Eqs. 12.9. If $k_y = 0$ in Eq. 12.9b, the result is

$$\begin{aligned} C(k_x, 0) &= \int_{-\infty}^{\infty} \cos(k_x x) dx \int_{-\infty}^{\infty} f(x, y) dy \\ &= \int_{-\infty}^{\infty} \cos(k_x x) F(\theta = 0, x) dx. \end{aligned} \tag{12.24}$$

Similarly

$$S(k_x, 0) = \int_{-\infty}^{\infty} \sin(k_x x) F(0, x) dx. \tag{12.25}$$

To state this in words: the Fourier transform of $F(0, x)$ determines the sine and cosine transforms of $f(x, y)$ along the line $k_y = 0$ (the k_x axis) in the spatial frequency plane. This is shown in Fig. 12.13.

A scan in another direction can be Fourier-transformed to give C and S at an angle θ with the k_x axis. The Fourier transform of the projection at angle θ is equal to the two-dimensional Fourier transform of the object, evaluated in the direction θ in Fourier transform space. This result is known as the *projection theorem* or the *central slice theorem* (Problem 20). The transforms of a set of projections at many different angles provide values of C and S throughout the $k_x k_y$ plane that can be used in Eq. 12.9a to calculate $f(x, y)$. In Chap. 18 we will find that the data from an MRI scan give the functions $C(k_x, k_y)$ and $S(k_x, k_y)$ directly.

In practice, the transforms are discrete. Using the notation that includes the redundant frequencies above $N/2$ and makes the coefficients half as large (Eqs. 11.27), the two-dimensional discrete Fourier transform (DFT) is[2]

$$f_{jk} = \sum_{l=0}^{N-1} \sum_{m=0}^{N-1} C_{lm} \cos\left[\frac{2\pi(jl + km)}{N}\right] \tag{12.26a}$$

[2] In this notation the low frequencies occur for low values of the indices l and m. Usually, as in Figs. 12.6, 12.7, 12.8, 12.9, 12.10,and 12.11, the indices are shifted so $k = 0$ occurs in the middle of the sum.

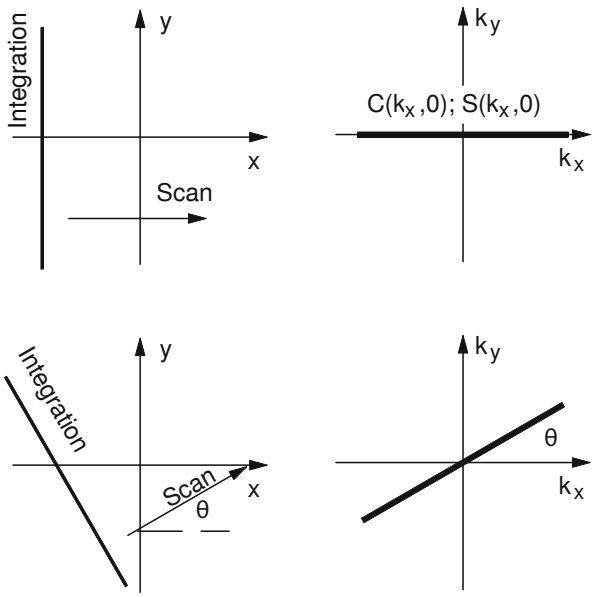


Fig. 12.13 The Fourier transform of $F(\theta = 0, x) = \int f(x, y)dy$ gives Fourier coefficients C and S along the k_x axis ($k_y = 0$). The Fourier transform of scans at other angles θ give C and S along corresponding lines in the $k_x k_y$ plane

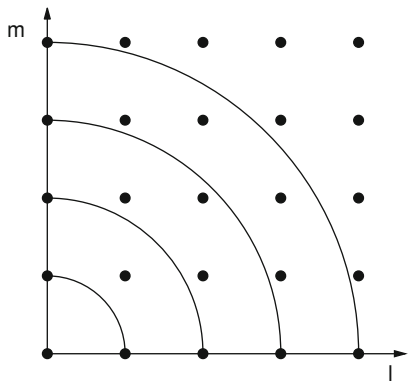


Fig. 12.14 The two-dimensional Fourier reconstruction requires values of C and S at the *lattice points* shown. The Fourier transforms of the projections $F(\theta, x)$ give the coefficients along the *circular arcs*. Interpolation is necessary to do the reconstruction

$$+\sum_{l=0}^{N-1}\sum_{m=0}^{N-1} S_{lm}\sin\left[\frac{2\pi(jl+km)}{N}\right].$$

The coefficients are given by

$$C_{lm} = \frac{1}{N^2}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} f_{jk}\cos\left[\frac{2\pi(jl+km)}{N}\right], \quad (12.26b)$$

$$S_{lm} = \frac{1}{N^2}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1} f_{jk}\sin\left[\frac{2\pi(jl+km)}{N}\right]. \quad (12.26c)$$

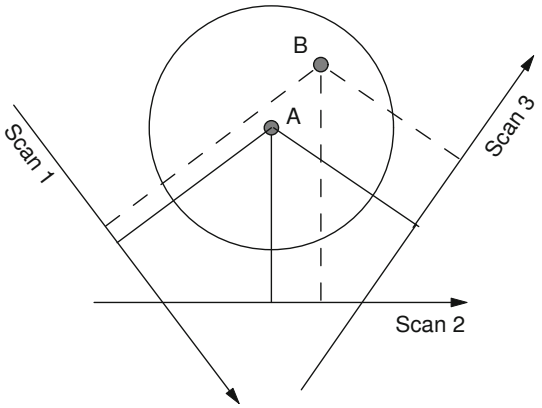


Fig. 12.15 The principle of back projection. Each point in the image is generated by summing all values of $F(\theta, x')$ that projected through that point. For point A at the center of rotation, the appropriate value of x' is the same at each angle. For other points such as B, the value of x' is different at each angle

Making a DFT of the projections gives values for C and S that lie on the circles in Fig. 12.14. But taking the inverse transform to calculate the reconstructed image requires values at the lattice points. They are obtained by interpolation (see Problem 23). The details of how the interpolation is made are crucial when using the Fourier transform reconstruction technique.

12.5 Reconstruction from Projections by Filtered Back Projection

Filtered back projection is more difficult to understand than the direct Fourier technique.[3] It is easy to see that every point in the object contributes to some point in each projection. The converse is also true. In a back projection every point in each projection contributes to some point in the reconstructed image. This can be seen from Fig. 12.15, which shows two points A and B and three projections. For point A, which is at the center of rotation, the relevant value of x' is the same in each projection, while for point B the value of x' is different in each projection.

A very simple procedure would be to construct an image by back-projecting every projection. The back projection $f_b(x, y)$ at point (x, y) is the sum of $F(\theta, x')$ for every projection or scan, using the value of x' that corresponds to the original projection through that point. That is, for Fig. 12.15, the back projection at point A would be the sum of the three values for which the solid projection lines intersect the scans,

[3] A simple experiment on back-projection using a laser pointer is described by Delaney and Rodriguez (2002).

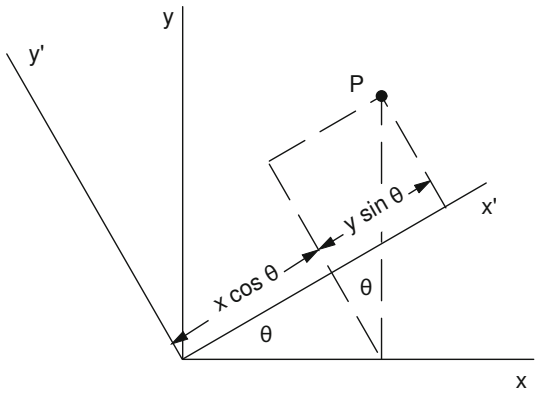


Fig. 12.16 By considering components of the coordinates of point P in both coordinate systems, one can derive the transformation equations, Eqs. 12.27 and 12.28

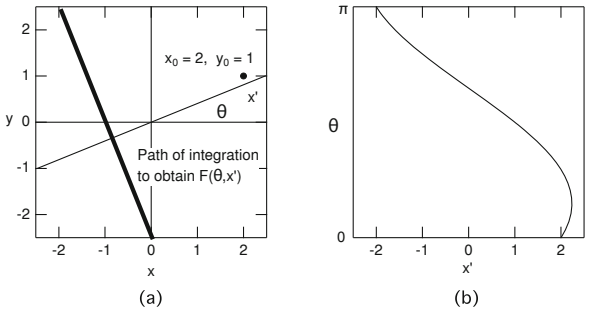


Fig. 12.17 An object and its sinogram. The object is a δ function at (x_0, y_0). **a** The object and the path of a projection at angle θ. **b** A sinogram of the object $F(\theta, x')$. The value of F would be plotted on an axis perpendicular to the $x'\theta$ plane. The line shows the values of θ and x' for which F is nonzero

while for point B it would be the sum of the values where the three dashed lines strike the scans. This gives a rather crude image, but we will see how to refine it.[4]

Figure 12.16 shows how to relate the values of x' and y' for a projection at angle θ to the object or image coordinates x and y. The transformations are

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta, \\ y' &= -x\sin\theta + y\cos\theta, \end{aligned} \tag{12.27}$$

and the inverse transformations are

$$\begin{aligned} x &= x'\cos\theta - y'\sin\theta, \\ y &= x'\sin\theta + y'\cos\theta. \end{aligned} \tag{12.28}$$

The projection at angle θ is integrated along the line y':

$$\begin{aligned} F(\theta, x') &= \int f(x, y)\, dy' \\ &= \int f(x'\cos\theta - y'\sin\theta, x'\sin\theta + y'\cos\theta) dy'. \end{aligned} \tag{12.29}$$

The process of calculating $F(\theta, x')$ from $f(x, y)$ is sometimes called the *Radon transformation*. When $F(\theta, x')$ is plotted with x' on the horizontal axis, θ on the vertical axis, and F as the brightness or height on a third perpendicular axis, the resulting picture is called a *sinogram*. For example, the projection of $f(x, y) = \delta(x - x_0)\delta(y - y_0)$ is $F(\theta, x') = \delta(x' - (x_0\cos\theta + y_0\sin\theta))$. A plot of this object and its sinogram is shown in Fig. 12.17.

The definition of the back-projection is

$$f_b(x, y) = \int_0^{\pi} F(\theta, x')\, d\theta, \tag{12.30}$$

where x' is determined for each projection by using Eq. 12.27. The limits of integration are 0 and π since the projection for $\theta + \pi$ repeats the projection for angle θ.

We will now show that the image $f_b(x, y)$ obtained by taking projections of the object $F(\theta, x')$ and then back-projecting them is equivalent to taking the convolution of the object with the function $h(x - x', y - y') = 1/r$, where r is the distance in the xy plane from the object point to the image point. Function h depends only on the distance between the object and image points. This is discussed in greater detail by Barrett and Myers (2004, p. 280). To simplify the algebra, we find the back projection at the origin. We want the set of projections for $x' = 0$ as a function of scan angle θ. They are, from Eq. 12.29,

$$F(\theta, 0) = \int_{-\infty}^{\infty} f(-y'\sin\theta, y'\cos\theta)\, dy'. \tag{12.31}$$

In terms of angle $\theta' = \theta + \pi/2$ which is the angle from the x axis to the y' axis,

$$F(\theta', 0) = \int_{-\infty}^{\infty} f(y'\cos\theta', y'\sin\theta')\, dy'.$$

The arguments of f look very much like components of a vector, with magnitude r' and components $r'\cos\theta'$ and $r'\sin\theta'$. This suggests expressing the integral in polar coordinates. Since y' is a dummy variable, call it r'. In terms of r' and θ' the projection is

[4] To see why it is crude, suppose the original object is a disk at the origin. Every projection will be the same because of the symmetry in angle. Every back projection will lay down a contribution to the image along a stripe. Even though the reconstructed image will be largest where the original circle was, the image will have nonzero values throughout the image plane. We will see this example in Sect. 12.6.

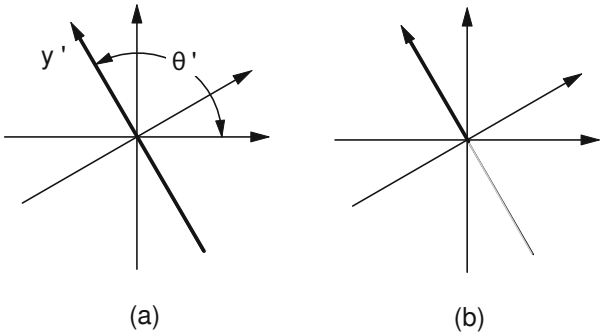


Fig. 12.18 Integration for the back projection is over y' from $-\infty$ to $+\infty$, as shown in **a**. This can be converted to an integral from 0 to ∞ if the angular integration is taken from 0 to 2π, as shown in **b**

$$F(\theta', 0) = \int_{-\infty}^{\infty} f(r', \theta')\, dr'. \tag{12.32}$$

Inserting this expression in Eq. 12.30 gives for the back projection

$$f_b(0,0) = \int_0^{\pi} F(\theta', 0) d\theta' = \int_{-\infty}^{\infty}\int_0^{\pi} f(r', \theta')\, dr'\, d\theta'. \tag{12.33}$$

Figure 12.18a shows how y' (that is, r') is integrated from $-\infty$ to ∞ while θ' goes from 0 to π. For the purposes of Eq. 12.33 the limits of integration can be changed as in Fig. 12.18b. Variable r' can range from 0 to ∞ while θ' goes from 0 to 2π. Then the expression for f_b looks even more like an integration in polar coordinates:

$$f_b(0,0) = \int_0^{\infty}\int_0^{2\pi} f(r', \theta')\, dr'\, d\theta'.$$

There is still one difference between this and polar coordinates. The element of area, which is $dx'dy'$ in Cartesian coordinates, is $r'dr'd\theta'$ in polar coordinates. Therefore, let us rewrite this as

$$f_b(0,0) = \int_0^{\infty}\int_0^{2\pi} \left(\frac{f(r', \theta')}{r'}\right) r'\, dr'\, d\theta'. \tag{12.34}$$

We now change to the Cartesian variables x' and y'. The back-projected image at the origin is

$$f_b(0,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{f(x', y')}{(x'^2 + y'^2)^{1/2}}\, dx'\, dy'. \tag{12.35}$$

For an arbitrary point (x, y) the result is similar:

$$f_b(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{f(x', y')}{\left[(x - x')^2 + (y - y')^2\right]^{1/2}}\, dx'\, dy'. \tag{12.36}$$

We have shown that the image obtained by taking projections of the object $F(\theta, x')$ and then back projecting them is equivalent to taking the convolution of the object with the function $h(x - x', y - y') = 1/r$, where r is the distance in the xy plane from the object point to the image point.

The back-projected image is not a faithful reproduction of the object. But it is possible to manipulate the projections $F(\theta, x')$ to produce a function $G(\theta, x')$ whose back projection is the desired $f(x, y)$. This is the process of *filtering* before making the back projection. To find the relationship between F and the desired function G, note that there is some function $g(x, y)$ that we do not know, but which, when projected and then back projected, yields the desired function $f(x, y)$. That is,

$$f(x, y) = g_b(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{g(x', y')}{\left[(x - x')^2 + (y - y')^2\right]^{1/2}}\, dx'\, dy'. \tag{12.37}$$

Equations 12.10 relate the Fourier coefficients of f, g, and $h(r) = 1/r$:

$$C_f(k_x, k_y) = C_g(k_x, k_y)C_h(k_x, k_y) - S_g(k_x, k_y)S_h(k_x, k_y),$$
$$S_f(k_x, k_y) = C_g(k_x, k_y)S_h(k_x, k_y) + S_g(k_x, k_y)C_h(k_x, k_y).$$

These can be solved for

$$S_g = \frac{C_h S_f - S_h C_f}{C_h^2 + S_h^2}, \qquad C_g = \frac{C_h C_f + S_h S_f}{C_h^2 + S_h^2}. \tag{12.38}$$

One can show by direct integration (see Problem 31) that the Fourier transform of $h(r) = 1/r$ is

$$C_h(k_x, k_y) = 2\pi(k_x^2 + k_y^2)^{-1/2}, \qquad S_h(k_x, k_y) = 0, \tag{12.39}$$

so that

$$C_g(k_x, k_y) = \frac{1}{2\pi}(k_x^2 + k_y^2)^{1/2} C_f(k_x, k_y), \qquad S_g(k_x, k_y) = \frac{1}{2\pi}(k_x^2 + k_y^2)^{1/2} S_f(k_x, k_y). \tag{12.40}$$

If function $g(x, y)$ were known and were projected to give $G(\theta, x')$, then back-projecting G would give the desired $f(x, y)$. The final step is to relate $G(\theta, x')$ and $F(\theta, x')$ so that we do not have to know $g(x, y)$. To establish this relationship, consider a projection on the x axis. Equations 12.24 and 12.25 show that

$$F(0, x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[C_f(k_x, 0)\cos(k_x x) + S_f(k_x, 0)\sin(k_x x)\right] dk_x,$$

while

$$G(0,x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[C_g(k_x,0)\cos(k_x x) + S_g(k_x,0)\sin(k_x x)\right] dk_x.$$

Equations 12.40 relate the Fourier coefficients for F and G. For $k_y = 0$, $(k_x^2 + k_y^2)^{1/2} = |k_x|$. Therefore

$$G(0,x) = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} \left[C_f(k_x,0)\cos(k_x x) + S_f(k_x,0)\sin(k_x x)\right] |k_x|\, dk_x. \quad (12.41)$$

This result is independent of the choice of axis, so it must be true for any projection. There is a function $h(x)$ which can be convolved with any $F(\theta, x)$ to give the desired function $G(\theta, x)$. Equation 12.41 shows that

$$C_g(k_x,0) = C_f(k_x,0)\,|k_x|/2\pi,$$
$$S_g(k_x,0) = S_f(k_x,0)\,|k_x|/2\pi.$$

Comparison with Eqs. 12.9 shows that

$$C_h = |k_x|/2\pi, \qquad S_h = 0.$$

Therefore

$$h(x) = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} |k_x| \cos(k_x x)\, dk_x.$$

Because the integrand is an even function, we can multiply by 2 and integrate from zero to infinity. The integral to infinity does not exist. However, there is some maximum spatial frequency, roughly the reciprocal of the resolution we want, which we call $k_{x\,\max}$. We can therefore cut the integral off at this maximum spatial frequency and obtain

$$\begin{aligned} h(x) &= \frac{1}{2\pi^2}\int_0^{k_{x\,\max}} k_x \cos(k_x x)\, dk_x \\ &= \frac{1}{2\pi^2}\left[\frac{\cos(k_x x)}{x^2} + \frac{k_x \sin(k_x x)}{x}\right]_0^{k_{x\,\max}} \\ &= \frac{k_{x\,\max}^2}{(2\pi)^2}\left[2\ \mathrm{sinc}(\xi) - \mathrm{sinc}^2(\xi/2)\right], \end{aligned} \quad (12.42)$$

where $\xi = k_{x\,\max} x$ and $\mathrm{sinc}(\xi) = \sin(\xi)/\xi$. The function $h(x)$ is plotted in Fig. 12.19. Using a sharp high-frequency cutoff introduces some problems, which are described below and which are easily overcome.

To summarize: *If each projection F is convolved with the function h of Eq. 12.42 and then back-projected, the back-projected image is equal to the desired image.*

Figure 12.20 summarizes the two methods of reconstructing an image from projections.

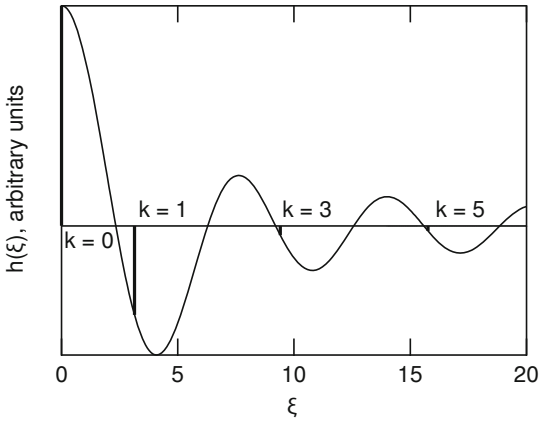


Fig. 12.19 The weighting function $h(x)$ of Eq. 12.42. The *bars* show the nonzero values for the example in Sect. 12.6

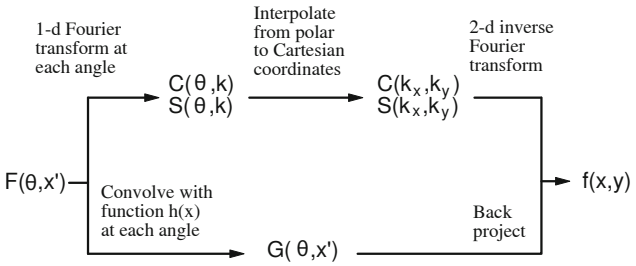


Fig. 12.20 A summary of the two methods for reconstructing an image

12.6 An Example of Filtered Back Projection

It is not difficult to write a computer program to perform filtered back projection if execution speed is not a concern. For our example we will use an object with circular symmetry, so that every projection is equivalent and only one projection needs to be convolved with the weighting function h. Because of the circular symmetry the back projection is needed only along one diameter. The program shown in Fig. 12.21 was used to reconstruct the image.

The "top-hat" function is used as the object:

$$f(x,y) = \begin{cases} 1, & x^2 + y^2 < a^2 \\ 0, & \text{otherwise.} \end{cases} \quad (12.43)$$

The projection is independent of θ: $F(x) = 2(a^2 - x^2)^{1/2}$ for $x^2 < a^2$. Procedure `CalcF` evaluates $F(x)$ for 100 points. Variables x and i are related by $x = 2i/N - 1$, so that x ranges from -1 to 1 as index i goes from 0 to 100. The value of a is 0.5.

The convolution is done by procedure `Convolve`, which uses convolving function h to operate on function F to produce G. The discrete form of h is obtained from Eq. 12.42 by the following argument, originally due to Ramachandran

```
// Circular Back Projection
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#define N 100  //Number of data points
const int n = N;
const double pi = 3.141592654;
double F[N], Image[N],
//Projection of object and image
// along a line [i]
  G[N]; //Convolved Projection
void CalcF(double *F)
/*Calculates the projection of a
circle of radius a centered at N/2.*/
{
  int i,i1,i2;
  double a = 0.5, x;
  i1 = (int)(50-a*((float)n)/2.0);
  i2 = (int)(50+a*((float)n)/2.0);
  for (i=0;i<n;i++)
  {
    x = 2*(i+1)/(float) n-1.0;
    F[i] = 0;
    if (i+1 > i1 && i+1< i2)
      F[i] = F[i]+2*sqrt(pow(a,2)-
pow(x,2));}
}
double H(int k)
{
  if (k==0)
    return  pow((float)n,2)/16.0;
  else if(k%2 == 1 || (-k)%2 == 1)
    return -pow(((float) n)/
    (pi*(float) k),2)/4.0;
  else
    return  0;
}
void Convolve(double*y, double*G)
{
  int i,j;
  double temp;
  for (i=0;i<n;i++)
  {
    temp = 0;
    for (j=0;j<n;j++)
      if (y[j] != 0)
      temp = temp+H(i-j)*y[j];
    G[i] = temp/((float) n);
  }
}
void BackProject(double*G,
  double*Image)
{
  const int MaxProj = 180;
  int i,j,l;
  double x,xinterp,
         temp,c;
  for(i=0;i<n;i++)
    Image[i] = 0;
  for (j=0;j<MaxProj;j++)
  {
    c = cos(pi*((float) j)/180.0);
    for (i=0;i<n;i++)
    {
      x = ((float)n)/2.0+(i+1
        -((float) n)/2.0)*c;
      l = (int) x;
      xinterp = x-l;
      if (l<=1)
        temp = G[0];
      else if(l>=n)
        temp = G[n-1];
      else
temp=G[l-1]+xinterp*(G[l]-G[l-1]);
      Image[i] = Image[i]+temp;
    }
  }
  for(i=0;i<n;i++)
    Image[i] = Image[i]*pi/
      ((float) MaxProj);
}
void PrintData(int n1, int n2,
  double*x, FILE *fp, char *title)
{
  int i,j;
  fprintf(fp,"\n\n%s\n",title);
  j = 0;
  for (i = n1 -1; i < n2; i++)
  {
    if (j%10 == 0)
      fprintf(fp,"\n%2i",i+1);
    fprintf(fp,"\t%8.3f",x[i]);
    j++;
  }
  fprintf(fp,"\t%8.3f",x[i]);
}
void main()
{
  FILE *ofp;
  if (!(ofp =
fopen("OutputFile","w")))
  {
    printf("cannot open output
file\n");
    exit(1);
  }
  fprintf(ofp,"\n");
  CalcF(F);
  PrintData(1,n,F,ofp,"Projected
Object
F");
  Convolve(F,G);
  PrintData(1,n,G,ofp,"Convolved
Projection G");
  BackProject(G,Image);
  PrintData(1,n,Image,ofp,"Back-
Projected
Image");
  fclose(ofp);
}
```

Fig. 12.21 The program used to make a filtered back projection of a circularly symmetric function

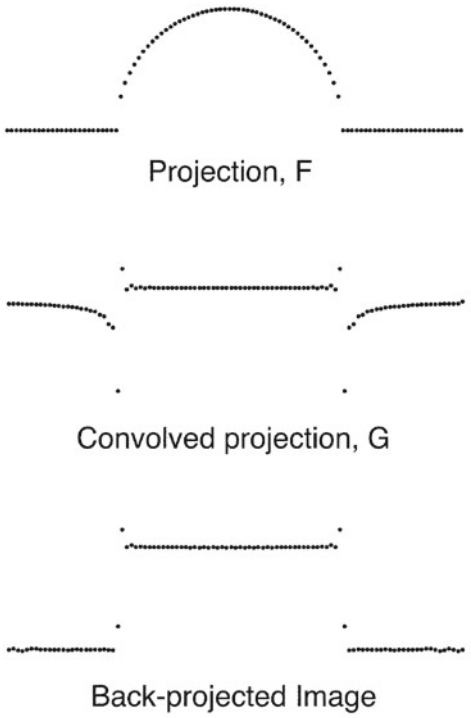


Fig. 12.22 Reconstruction of a circularly symmetric image by filtered back projection. **a** The projection $F(x)$. **b** The convolved projection $G(x)$. **c** The image from back-projecting the convolved data

and Lakshminarayanan (see Cho et al. 1993, p. 80). Variable x is considered on the interval $(-1, 1)$, so the period is 2 and $\omega_0 = \pi$. The maximum spatial frequency is $k_{x\,\max} = N\pi/2$. The value of x in the weighting function $h(x)$ depends on the value of index $k = i - j$: $x_i - x_j = 2(i - j)/N = 2k/N$. Therefore $\xi = k_{x\,\max}x = (N\pi/2)(2/N)k = \pi k$, where k is an integer. From Eq. 12.42 we obtain

$$h(k) = \begin{cases} N^2/16, & k = 0 \\ 0, & k \text{ even} \\ -N^2/4k^2\pi^2 & k \text{ odd.} \end{cases} \qquad (12.44)$$

Procedure `Convolve` replaces the integral of Eq. 12.4a by a sum. The factor dx in the integral becomes $1/N$ in the sum.

Procedure `BackProject` forms the image from G. One hundred eighty projections are done in 1° increments from 0 to 179. The value of x is determined from $x = i\cos\theta$, but it is shifted so that the rotation takes place about $i = 50$. Unless x is at the end points, the value of G is obtained by linear interpolation. The value of $\Delta\theta$ used to convert the integral to a sum is $\pi/180$.

Procedure `PrintData` writes the data for the plots shown in Fig. 12.22. One can see from inspection of Fig. 12.22 how the convolution converts the semicircular projection F into a function G that is flat-topped over the



Fig. 12.23 Reconstruction by simple back projection without convolution. The object is the same as in Fig. 12.22

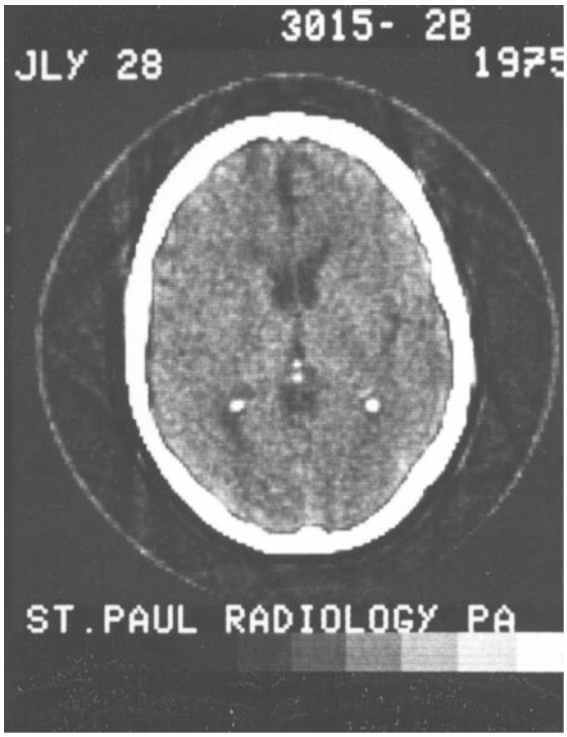


Fig. 12.24 An early CT brain scan, showing ringing inside the skull. Photograph courtesy of St. Paul Radiology Associates, St. Paul, MN

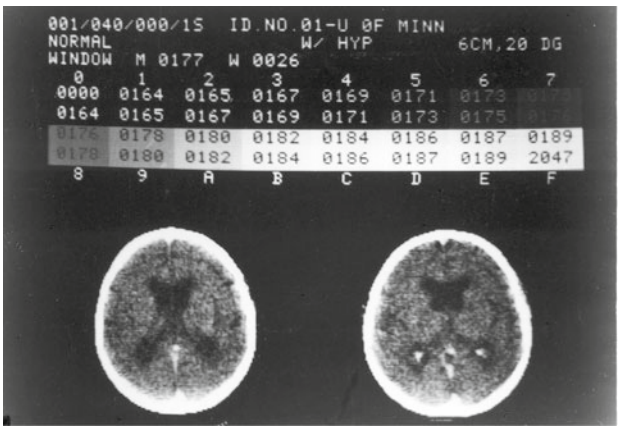


Fig. 12.25 Brain scans using a gradual high-frequency cutoff to eliminate ringing. Photograph courtesy of Prof. J. T. Payne, Department of Diagnostic Radiology, University of Minnesota

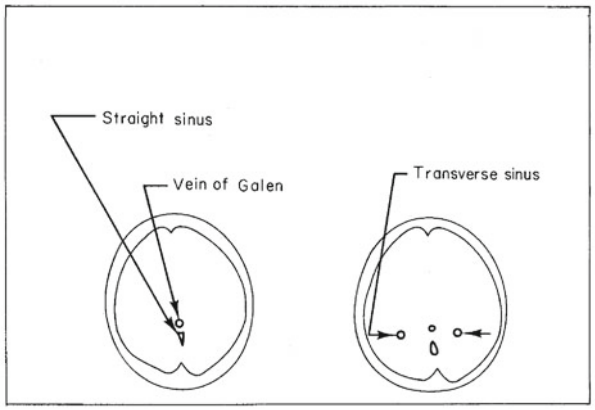


Fig. 12.26 Anatomic features shown in Fig. 12.25

region of nonvanishing f and has a negative contribution in the wings. Figure 12.23 shows what the image looks like if the back projection is done without first performing the convolution.

One can also see from Fig. 12.22 that ringing is introduced at the sharp edges. This is characteristic of the sharp high-frequency cutoff at k_x (similar to the Fourier series representation of a square wave with only a finite number of terms). Early computer tomography (CT) scans created with the convolution function presented here showed a dark band just inside the skull where there was an abrupt change in $f(x, y)$ upon going from bone to brain (Fig. 12.24). A gradual high-frequency cutoff changes the details of $h(k)$ and eliminates this ringing (Fig. 12.25). Fig. 12.26 shows the anatomic details of Fig. 12.25.

Symbols Used in Chap. 12

Symbol	Use	Units	First used page
a, b	Constants		348
a	Radius of "top-hat" function	m	355
b'	Length of object	m	346
f, g	Arbitrary functions		345
f_b, g_b	Back-projected images of f, g		353
h	Point spread function; impulse response for convolution		345
i	$\sqrt{-1}$		347
j, k	Subscript indices for data		352
k, k_x, k_y	Spatial frequencies	m^{-1}	346
l, m	Subscript indices for Fourier coefficients		351
m	Magnification		347
t, t'	Time or arbitrary variable		345
x, y, x', y'	Distance; coordinates in image or object plane; rotated coordinate system for image reconstruction	m	346
A	Amplitude		346
C_f	Fourier cosine transform of function f		346
D	Length of image	m	349
E	Function describing an image		347
F	Projection of function f		351
F, G, H	Complex Fourier transforms of f, g, h		347
L	Property describing an object		347
L	Width of image or Field of View (FOV)	m	349
N	Total number of data points; number of discrete values in one dimension of an image		349
S_f	Fourier sine transform of function f		346
T	Period	s	346
$\delta(t)$	Dirac delta function	s^{-1}	345
λ	Wavelength	m	346
ϕ	Phase		346
θ, θ'	Angle		351
τ_1	Time constant	s	346
ω, ω_0	Angular frequency	(radian) s^{-1}	346
ξ	Dummy variable		353

Problems

Section 12.1

Problem 1. Compare Eq. 12.4a to Eqs. 4.73 and 7.21 and deduce the impulse response for those two systems.

Problem 2. Except for the minus sign, Eq. 12.4a is the same integral that defines the cross-correlation function. There are some important differences, however. Show that the convolution function is commutative—interchanging the order of variables gives the same result—but that the cross-correlation function is not.

Problem 3. (a) Use the convolution integral, Eq. 12.4a, to calculate the convolution $g(t)$ of the function $h(t - t')$ in Eq. 12.5 with

$$f(t) = \begin{cases} 1, & 0 < t < T, \\ 0, & \text{otherwise} \end{cases}.$$

Plot $f(t)$ and $g(t)$.

(b) Calculate the Fourier transform of $g(t)$, $h(t - t')$, and $f(t)$ from part (a), and show that they obey Eq. 12.6a.

Problem 4. Fill in the details in the derivation of Eq. 12.6a.

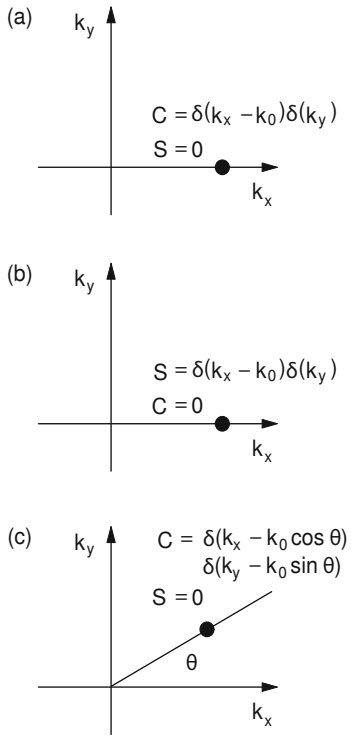
Problem 5. Use the convolution integral to calculate $g(x)$ from $h(x - x') = a/[a^2 + (x - x')^2]$ and $f(x) = \cos(kx)$.

Interpret this physically as a spatial frequency filter. Hint:

$$\int_{-\infty}^{\infty} \frac{\cos(ky)dy}{y^2+b^2} = \frac{\pi}{b}e^{-kb},$$
$$\int_{-\infty}^{\infty} \frac{\sin(ky)dy}{y^2+b^2} = 0.$$

Problem 6. If you are familiar with complex variables, use the definition of the Fourier transform in Eq. 12.11a to prove the convolution theorem, Eq. 12.11b.

Problem 7. What are the two-dimensional images whose Fourier transforms are shown?



Problem 8. Calculate the two-dimensional Fourier transform of the function

$$f(x,y) = \begin{cases} 1, & -a/2 < x < a/2, \quad -b/2 < y < b/2, \\ 0, & \text{otherwise.} \end{cases}$$

Plot $f(x,y)$ vs x and y and $C_f(k_x,k_y)$ vs k_x and k_y for $a = 2b$.

Problem 9. Calculate the two-dimensional Fourier transform of the function

$$f(x,y) = \text{sech}\left(\frac{x}{a}\right)\text{sech}\left(\frac{y}{b}\right).$$

You may need the relationship

$$\int_0^{\infty} \text{sech}(uz)\cos(vz)dz = \frac{\pi}{2u}\text{sech}\left(\frac{\pi v}{2u}\right).$$

Problem 10. Calculate the two-dimensional Fourier transform of the function

$$f(x,y) = \begin{cases} 1, & \sqrt{x^2+y^2} < a, \\ 0, & \sqrt{x^2+y^2} > a. \end{cases}$$

Hint: convert to polar coordinates in both the xy and k_xk_y planes, and use the facts that

$$J_0(u) = \frac{1}{2\pi}\int_0^{2\pi} \cos(u\cos v)dv,$$
$$\int uJ_0(u)du = uJ_1(u),$$

where J_0 and J_1 are Bessel functions of order zero and order one. Bessel functions are tabulated and have known properties, similar to trigonometric functions. See Abramowitz and Stegun (1972), p. 360.

Section 12.2

Problem 11. Complete the verification of Eq. 12.13 suggested in the text.

Problem 12. Find the Fourier transform of the point spread function for the ideal imaging system, Eq. 12.13.

Problem 13. Use Eq. 12.15 to show that the sum of the squares of the Fourier coefficients of the image is equal to the sum of the squares of the Fourier coefficients of the object times the square of the modulation transfer function, for a given set of spatial frequencies (k_x,k_y).

Problem 14. Write the modulation of the image in terms of the variables in Eq. 12.19.

Problem 15. How does magnification m change the spatial frequencies in going from object to image? Since one is concerned about seeing detail in the object, resolution and spatial frequencies are usually converted to object coordinates in medical imaging, while they are left in terms of the detector coordinates in photography.

Section 12.3

Problem 16. This problem shows how increasing the detail in an image introduces high-frequency components. Find the continuous Fourier transform of the two functions

$$f_1(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 < x < 1, \\ 0, & x > 1 \end{cases}$$

$$f_2(x) = \begin{cases} 0, & x < 0, \\ \sqrt{3/2}, & 0 < x < 1/3, \\ 0, & 1/3 < x < 2/3, \\ \sqrt{3/2}, & 2/3 < x < 1, \\ 0, & x > 1. \end{cases}$$

Plot $a(k_x) = \left[C^2(k_x) + S^2(k_x)\right]^{1/2}$ for each function using a spreadsheet or plotting package, for the range $-45 < k_x < 45$. Compare the features of each plot. Both functions have the same value of $\int_{-\infty}^{\infty} f^2(x)dx$.

Problem 17. To see the blurring effect shown in Fig. 12.8 consider a one-dimensional problem. Let $y_0 = 1$ and $y_j = 0$ for $j = 1, 2, \ldots, 7$. Use Eqs. 11.27b and 11.27c to calculate the discrete Fourier transform of this function a_k and b_k for $k = 0, 1, 2, \ldots, 7$. Then remove the high frequencies by setting $a_k = b_k = 0$ for $k = 2, 3, 4, \ldots, 6$, as was done in Fig. 12.8. (Note that the $k = 7$ point is equivalent to $k = -1$ and therefore acts like a "low" frequency.) Use Eq. 11.27a to calculate new values of y_j. Do you get a blurred image?

Problem 18. To see the edge effect in Fig. 12.10, consider the one-dimensional function defined in Problem 17. After calculating the discrete Fourier transform a_k and b_k, remove the low frequencies by setting $a_k = b_k = 0$ for all k except $k = 3, 4, 5$ as was done in Fig. 12.10. (Note that the points $k = 6$ and 7 are equivalent to $k = -2$ and $k = -1$ and therefore act like "low" frequencies.) Use Eq. 11.27a to calculate new values of y_j. Do you get an image with edge effects?

Problem 19. To see the ghost image effect in Fig. 12.11, consider the one-dimensional function described in Problem 17. After calculating the discrete Fourier transform a_k and b_k, set $a_k = b_k = 0$ for all odd values of k as was done in Fig. 12.11. Use Eq. 11.27a to calculate new values of y_j. Do you get a "ghost image?"

Section 12.4

Problem 20. Prove the central slice theorem analytically. Consider the cosine term of the 2-dimensional Fourier transform $C(k_x, k_y)$ in Eq. 12.9b. Rotate to the primed coordinates given by Eq. 12.28. Note that the area element $dxdy$ transforms to $dx'dy'$. Express C as a function of polar coordinates in k-space, $k_x = k\cos\theta$ and $k_y = k\sin\theta$. Show that

$$C(\theta, k) = \int_{-\infty}^{\infty} F(\theta, x')\cos(kx')dx',$$
$$S(\theta, k) = \int_{-\infty}^{\infty} F(\theta, x')\sin(kx')dx'.$$

Problem 21. Suppose that $f(x, y)$ is independent of y. Find expressions for $C(k_x, k_y)$ and $S(k_x, k_y)$ and insert them in the expression for $f(x, y)$ to verify that $f(x, y)$ is recovered. You will need Eqs. 11.66.

Problem 22. Suppose that the object is a point at the origin, so that $f(x, y) = \delta(x)\delta(y)$. Find the projection $F(x)$ and the transform functions $C(k_x, 0)$ and $S(k_x, 0)$. Use these results to reconstruct the image using the Fourier technique.

Problem 23. Figure 12.14 shows that taking the Fourier transform of the projection $F(\theta, x')$ gives the Fourier coefficients $C(k, \theta)$ at points along circular arcs in frequency space. In order to get these coefficients at equally spaced points in x and y, interpolation is necessary. One simple method is to use *bilinear interpolation* (Press et al. 1992). Suppose you know the Fourier coefficients at points $r_i = i\Delta r$, $\theta_j = j\Delta\theta$, and you want to get the Fourier coefficients at points $x_n = n\Delta x$, $y_m = m\Delta y$. For a given x_n, y_m, convert to polar coordinates to get r and θ, then find the four known points that "surround" the desired point. The value of the coefficient is

$$\begin{aligned} C(x_n, y_m) = \frac{1}{\Delta r\Delta\theta}\big[&C(r_i, \theta_j)(r_{i+1} - r)(\theta_{j+1} - \theta) \\ &+C(r_{i+1}, \theta_j)(r - r_i)(\theta_{j+1} - \theta) \\ &+C(r_i, \theta_{j+1})(r_{i+1} - r)(\theta - \theta_j) \\ &+C(r_{i+1}, \theta_{j+1})(r - r_i)(\theta - \theta_j)\big]. \end{aligned}$$

Suppose $C(r, \theta) = \sin(r)/r$, which is also called $\mathrm{sinc}(r)$. If C is known at points with $\Delta r = 0.5$ and $\Delta\theta = \pi/8$, evaluate C at point $x = 2$, $y = 3$ using bilinear interpolation. Compare this result to the exact value of $C = \mathrm{sinc}((x^2 + y^2)^{1/2})$. Try this for other points (x_n, y_m).

Section 12.5

Problem 24. Derive Eqs. 12.27 and 12.28.

Problem 25. An object is described by the function $f(x, y) = e^{-(x^2+y^2)/b^2}$.

(a) Find the Fourier transform $C(k_x, k_y)$ and $S(k_x, k_y)$ directly from Eqs. 12.9 b and c.

(b) Find the projection $F(\theta, x')$ using Eq. 12.29. Then take the 1-dimensional Fourier transform of $F(\theta, x')$ using the equations

$$C(\theta, k) = \int_{-\infty}^{\infty} F(\theta, x')\cos kx'dx'$$
$$S(\theta, k) = \int_{-\infty}^{\infty} F(\theta, x')\sin kx'dx'.$$

Use $k = \left(k_x^2 + k_y^2\right)^{1/2}$ to express C and S in terms of k_x and k_y. Your answer should be the same as part (a).

Use the following integral table:

$$\int_{-\infty}^{\infty} e^{-az^2}\,dz = \sqrt{\frac{\pi}{a}}$$
$$\int_{-\infty}^{\infty} e^{-az^2}\cos bz\,dz = \sqrt{\frac{\pi}{a}}e^{-b^2/4a}$$
$$\int_{-\infty}^{\infty} e^{-az^2}\sin bz\,dz = 0.$$

Problem 26. Assume you have just measured the projection function $F(\theta, x') = \pi^{1/2}be^{-(x'-a\cos\theta)^2/b^2}$. (For this problem, ignore the fact that your measuring device would only give F at discrete values of θ and x'.)

(a) Find $f(x, y)$ using the Fourier method. You may need the integrals from Problem 25.

(b) Qualitatively sketch plots of the object $f(x, y)$ and the sinogram $F(\theta, x')$. Use a gray scale to indicate the magnitude of f and F.

Problem 27. Repeat Problem 26 using

$$F(\theta, x') = \frac{a\sqrt{\pi}}{2}e^{-x'^2/a^2}\left[1 + \cos^2\theta\left(2\frac{x'^2}{a^2} - 1\right)\right].$$

Look up any integrals you need.

Problem 28. Suppose an object is a point at the origin, $f(x, y) = \delta(x)\delta(y)$. The projection is also a point: $F(\theta, x') = \delta(x')$. Calculate the back projection $f_b(x, y)$ (without filtering) using Eq. 12.30. To solve the problem, use this property of δ functions:

$$\delta(g(u)) = \sum_i \frac{\delta(u - u_i)}{|dg/du|_{u=u_i}},$$

where the u_i are the points such that $g(u_i) = 0$. Note that the back projection is not a point. Back projection without filtering does not recover the object.

Problem 29. This problem is an extension of Problem 28, but the object is no longer at the origin. Let $f(x, y) = \delta(x - x_0)\delta(y - y_0)$.

(a) Calculate $F(\theta, x')$. You may need the following properties of the δ function: $\int \delta(b - z)\delta(z - a)dz = \delta(b - a)$, $\delta(az) = \delta(z)/|a|$.

(b) Use the function $F(\theta, x')$ you found in part (a) to calculate the back projection $f_b(x, y)$ using Eq. 12.30. You will need the property of the δ function given in Problem 28.

(c) Show that $f_b(x, y)$ is equivalent to the convolution of $f(x, y)$ with the function $1/\sqrt{(x - x')^2 + (y - y')^2}$.

Problem 30. Here is an easy way to show that the back projection $f_b(x, y)$ cannot be equivalent to the object $f(x, y)$. If $f(x, y)$ is dimensionless, determine the units of $F(\theta, x')$ and $f_b(x, y)$. Do $f(x, y)$ and $f_b(x, y)$ have the same units?

Problem 31. Consider the Fourier transform of $1/r$. The coefficients are given by

$$C(k_x, k_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{dx\,dy\cos(k_x x + k_y y)}{(x^2 + y^2)^{1/2}},$$
$$S(k_x, k_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{dx\,dy\sin(k_x x + k_y y)}{(x^2 + y^2)^{1/2}}.$$

Transform to polar coordinates ($x = r\cos\theta$, $y = r\sin\theta$). Show from symmetry considerations of the angular integral that $S = 0$. Use the facts about the Bessel functions in Problem 10 and

$$\int_0^{\infty} J_0(kr)dr = 1/k$$

to derive Eqs. 12.39. The function $J_0(x)$ is a *Bessel function of order zero*. It is tabulated and has known properties, similar to a trigonometric function. See Abramowitz and Stegun (1972, p. 360).

Problem 32. An object consists of three δ functions at $(0, 2)$, $(\sqrt{3}, -1)$, and $(-\sqrt{3}, -1)$. Draw the sinogram of the object.

Problem 33. Let $f(x, y) = 1/[(x - a)^2 + y^2 + b^2]$. Calculate $F(\theta, x')$. Qualitatively sketch plots of the object $f(x, y)$ and the sinogram $F(\theta, x')$. Use a gray scale to indicate the magnitude of f and F.

Problem 34. Let $f(x, y) = x/(x^2 + y^2)^2$. Calculate $F(\theta, x')$. Qualitatively sketch plots of the object $f(x, y)$ and the sinogram $F(\theta, x')$. Use a gray scale to indicate the magnitude of f and F. Hint:

$$\int \frac{du}{\left(u^2 + v^2\right)^2} = \frac{u}{2v^2\left(u^2 + v^2\right)} + \frac{1}{2|v|^3}\tan^{-1}\left(\frac{u}{v}\right).$$

Problem 35. Consider the object $f(x, y) = a/\sqrt{a^2 - x^2 - y^2}$ for $\sqrt{x^2 + y^2} < a$, and 0 otherwise.

(a) Plot $f(x, 0)$ vs x.

(b) Calculate the projection $F(\theta, x')$. Plot $F(0, x')$ vs x'.

(c) Use the projection from part (b) to calculate the back projection $f_b(x, y)$. Plot $f_b(x, 0)$ vs x.

(d) Compare the object and the back projection. Explain qualitatively how they differ.

Section 12.6

Problem 36. Verify that

$$F(\theta, x) = \begin{cases} 2\sqrt{a^2 - x^2}, & |x| < a \\ 0, & |x| > a \end{cases}$$

is the projection of the function in Eq. 12.43.

Problem 37. Verify Eqs. 12.44.

Problem 38. Modify the program of Fig. 12.21 and run it without the convolution.

Problem 39. Modify the program of Fig. 12.21 to reconstruct an annulus instead of a top-hat function.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions with formulas, graphs and mathematical tables. Government Printing Office, Washington, DC

Barrett HH, Myers KJ (2004) Foundations of image science. Wiley-Interscience, New York

Buonocore MH, Gao L (1977) Ghost artifact reduction for echo planar imaging using image phase correction. Magn Reson Med 38: 89–100.

Cho Z-h, Jones JP, Singh M (1993) Foundations of Medical Imaging. Wiley, New York

Delaney C, Rodriguez J (2002) A simple medical physics experiment based on a laser pointer. Am J Phys 70(10):1068–1070

Gaskill JD (1978) Linear systems, fourier transforms, and optics. Wiley, New York

Kevles BH (1997) Naked to the bone: medical imaging in the twentieth century. Rutgers University Press, New Brunswick

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing. 2nd ed. reprinted with corrections, 1995. Cambridge University Press, New York

Ramachandran GN, Lakshminarayanan AV (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. Proc Nat Acad Sci USA 68:2236–2240

Shaw R (1979). Applied optics and optical engineering. Academic, New York, pp 121–154. (Photographic Detectors. Ch. 5 Vol. **7**).

Williams CS, Becklund OA (1972) Optics: a short course for engineers and scientists. Wiley, New York

Sound and Ultrasound

13

Sound (or *acoustics*) plays two important roles in our study of physics in medicine and biology. First, animals hear sound and thereby sense what is happening in their environment. Second, physicians use high-frequency sound waves (*ultrasound*) to image structures inside the body. This chapter provides a brief introduction to the physics of sound and the medical uses of ultrasound for imaging and therapy. A classic textbook by Morse and Ingard (1968) provides a more thorough coverage of theoretical acoustics, and books such as Hendee and Ritenour (2002) describe the medical uses of ultrasound in more detail.

In Sect. 13.1, we derive the fundamental equation governing the propagation of sound: the wave equation. Section 13.2 discusses some properties of the wave equation, including the relationship between frequency, wavelength, and the speed of sound. The acoustic impedance and its relevance to the reflection of sound waves are introduced in Sect. 13.3. Section 13.4 describes the intensity of a sound wave and introduces the decibel intensity scale. The ear and hearing are described in Sect. 13.5. Section 13.6 discusses the attenuation of sound waves. Physicians use ultrasound imaging for medical diagnosis, as described in Sect. 13.7.

13.1 The Wave Equation

In Chap. 1, we assumed that solids and liquids are incompressible. If a long rod was truly incompressible, a displacement of one end would instantly result in an identical displacement of the other end. In fact, the displacement does not propagate instantaneously. It travels at the speed of sound in the rod.

The propagation of sound involves small displacements of each volume element of the medium from its equilibrium position. In this section, we consider sound waves propagating along the x-axis. The results can be generalized to three dimensions (see Morse and Ingard 1968). We first consider an elastic rod, then a fluid in which viscous effects are not important, and finally, shear waves.

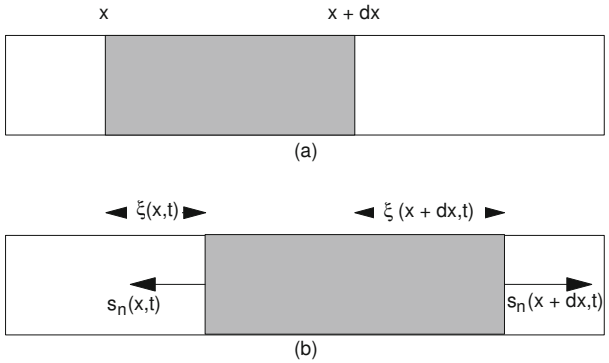


Fig. 13.1 An elastic rod. **a** The rod in its equilibrium position. **b** Each point on the rod has been displaced from its equilibrium position by an amount ξ which depends on x and t. As a result there is a normal stress s_n which also depends on x and t

13.1.1 Plane Waves in an Elastic Rod

The simplest case to consider is an elastic rod which is forced to move longitudinally at one end. This results in the propagation of a sound wave along the rod. We set up a coordinate system where x measures distance along the rod from a fixed origin when no sound wave is traveling along the rod. We also assume that the disturbance of the rod depends only on the position along the rod, x, and not on y or z, which are perpendicular to x. A wave in three dimensions that depends only on one dimension is called a *plane wave*.

When the sound travels along the rod, the material at point x is displaced from its undisturbed position by a small amount $\xi(x, t)$, as shown in Fig. 13.1. The material originally at $x + dx$ is displaced by amount $\xi(x + dx, t)$. Since

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_13, © Springer International Publishing Switzerland 2015

$\xi(x + dx, t)$ is in general different from $\xi(x, t)$, there is a strain in the rod (Eq. 1.24)

$$\epsilon_n(x, t) = \frac{\Delta l}{l} = \frac{\xi(x + dx, t) - \xi(x, t)}{dx} = \frac{\partial \xi}{\partial x}. \quad (13.1)$$

Young's modulus, E, relates the stress in the rod, s_n, to the strain, ϵ_n (Eq. 1.25):

$$s_n(x, t) = E\epsilon_n(x, t) = E\frac{\partial \xi}{\partial x}. \quad (13.2)$$

The difference between the stress at each end, multiplied by the cross-sectional area of the rod, S, provides a net force that accelerates the shaded volume element in Fig. 13.1. The net force on the volume element is

$$F_{\text{net}} = S\,[s_n(x + dx, t) - s_n(x, t)] = S\frac{\partial s_n}{\partial x}dx = SE\frac{\partial \epsilon_n}{\partial x}dx,$$

$$F_{\text{net}} = SE\frac{\partial^2 \xi}{\partial x^2}dx. \quad (13.3)$$

The mass of the shaded volume is $\rho S dx$, where ρ is the density, and the acceleration of the volume is $\partial^2\xi/\partial t^2$. [Since we are not subtracting a value at one end from the value at the other, and since we are taking the limit as $dx \to 0$, we can ignore changes in ξ in the interval $(x, x + dx)$]. Therefore, Newton's second law becomes

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{\rho}{E}\frac{\partial^2 \xi}{\partial t^2}. \quad (13.4)$$

This is the *wave equation*, and it is seen in many contexts, from the vibrations of a string to the propagation of electromagnetic waves. It is usually written as

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 \xi}{\partial t^2}, \quad (13.5)$$

where c is the speed of propagation of sound in the rod. In this case

$$c = \sqrt{\frac{E}{\rho}}. \quad (13.6)$$

As Young's modulus becomes very large or the density of the rod becomes very small, the speed with which a disturbance travels from one end of the rod to the other becomes larger and larger.

13.1.2 Plane Waves in a Fluid

Now we consider a sound wave propagating in a fluid, where shear can be neglected. We also neglect viscous effects. Changes in the fluid caused by the sound wave depend only on x and t.[1] To make it easier to imagine the situation, suppose the fluid is confined in a tube. Then, we can construct a figure very similar to Fig. 13.1. A small volume of fluid at rest extends from position x to $x + dx$, with cross-sectional area S as shown in Fig. 13.2a. The force pushing on the left side of the volume is SP_0, and the force on the right is $-SP_0$.[2] Here, P_0 is the pressure when the fluid is undisturbed by a sound wave. In equilibrium, there is no net force on the volume element.

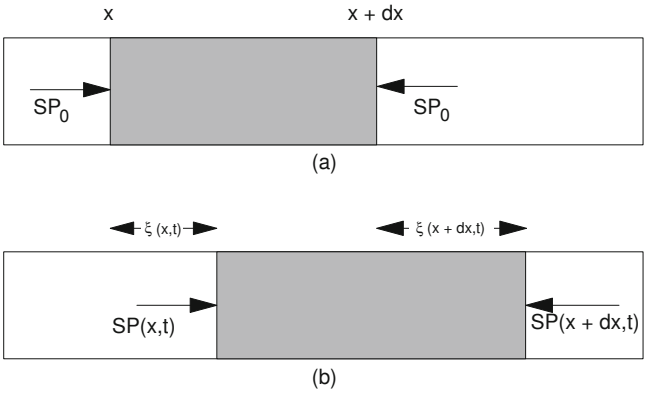


Fig. 13.2 Sound propagates in one dimension in a fluid in a tube of cross-sectional area S. **a** In equilibriuim the pressure is p_0 and the force on the shaded volume of fluid has magnitude $p_0 S$ on each end. **b** When the sound is propagating, the forces on each end are as shown

When the fluid element is displaced, as in Fig. 13.2b, the net force to the right on the fluid element is

$$F_{\text{net}} = S\,[P(x, t) - P(x + dx, t)] = -S\frac{\partial P}{\partial x}dx. \quad (13.7)$$

The change of pressure from the equilibrium value P_0 is related to the change of volume of the fluid by the compressibility, κ (Eq. 1.32):

$$P - P_0 = p = -\frac{1}{\kappa}\frac{dV}{V_0} = -\frac{1}{\kappa}\frac{d\xi}{dx}, \quad (13.8)$$

from which

$$F_{\text{net}} = \frac{S}{\kappa}\frac{\partial^2 \xi}{\partial x^2}dx. \quad (13.9)$$

To obtain the mass, we use the volume $S dx$ times the equilibrium density ρ_0. We multiply by the acceleration of the fluid element, $\partial^2\xi/\partial t^2$, to obtain

$$\frac{\partial^2 \xi}{\partial x^2} = \rho_0\kappa\frac{\partial^2 \xi}{\partial t^2}. \quad (13.10)$$

[1] We might be looking at a wave whose properties depend on all three coordinates, x, y, and z, but where, in the region we are studying, the dependence on y and z is very slight. This is like the 1-D electrostatic approximations in Chap. 6.

[2] See Sect. 1.12; we ignore any forces arising from viscosity, gravity, or surface tension.

This is the wave equation, Eq. 13.5, with

$$c = \sqrt{\frac{1}{\kappa\rho_0}}. \tag{13.11}$$

In both of these cases, the wave equation has been written in terms of the displacement of elements of the rod or the fluid from their equilibrium positions. It is also possible to show that the pressure, fluid density, and velocity of the fluid element also satisfy the wave equation. The pressure is discussed in Problem 2. The velocity of the fluid due to the sound wave is

$$v = \frac{d\xi}{dt}. \tag{13.12}$$

Another important relationship is obtained by combining Eq. 13.12 with Eq. 13.8 and interchanging the order of differentiation (Appendix N):

$$\frac{\partial v}{\partial x} = -\kappa\frac{\partial p}{\partial t}. \tag{13.13}$$

Equation 13.8 and 13.10 can also be used to show that

$$\frac{\partial v}{\partial t} = -\frac{1}{\rho_0}\frac{\partial p}{\partial x}. \tag{13.14}$$

Finally, since the density is $\rho = M/V$, we can show that

$$\frac{d\rho}{\rho_0} = \kappa dp. \tag{13.15}$$

In this section, we have considered Young's modulus E and compressibility κ. Remember from Chap. 3 that we can compress a gas at a constant temperature, and we can also do it adiabatically, in which case there is no heat flow and the temperature rises as the gas is compressed. The compressibility is different in these two cases. When static measurements of these parameters are made, there is usually time for the system being studied to remain isothermal. The pressure changes in a sound wave usually occur so rapidly that there is not time for heat to flow, and the adiabatic compressibility must be used. Values of Young's modulus are also different for isothermal and adiabatic stresses and strains.

13.1.3 Shear Waves

Sound in a fluid is a *longitudinal wave*, which means that the fluid moves in the same direction that the wave propagates. A fluid cannot support a shear stress, but shear stresses can exist in tissue, which results in another type of acoustic wave, called a *transverse wave* or *shear wave*, where the tissue moves perpendicular to the direction the wave propagates. Consider the tissue in Fig. 13.3. When a shear wave travels through the tissue, the material at point x is displaced

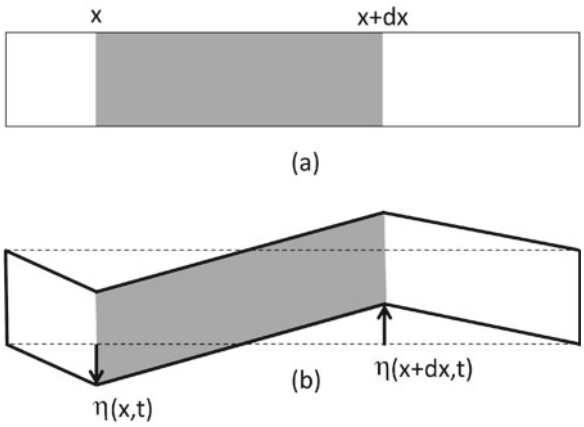


Fig. 13.3 A shear wave in a rod viewed from the top. The deflection is in the plane of the paper

by a small amount $\eta(x,t)$ in the transverse direction. The shear strain is (Eq. 1.27)

$$\epsilon_s = \frac{\eta(x+dx,t) - \eta(x,t)}{dx} = \frac{\partial\eta}{\partial x},$$

and the shear modulus, G, relates the stress and strain (Eq. 1.28)

$$s_s = G\epsilon_s = G\frac{\partial\eta}{\partial x}.$$

The difference between the stress at each end of the shaded volume in Fig. 13.3, multiplied by the cross-sectional area S, provides the net force

$$F_{net} = S[s_s(x+dx,t) - s_s(x,t)] = S\frac{\partial s_s}{\partial x}dx = SG\frac{\partial^2\eta}{\partial x^2}dx.$$

The mass of the shaded volume is $\rho S dx$, and the acceleration of the volume is $\partial^2\eta/\partial t^2$. Newton's second law becomes

$$\frac{\partial^2\eta}{\partial x^2} = \frac{\rho}{G}\frac{\partial^2\eta}{\partial t^2}$$

so the speed of the shear wave is

$$c_{\text{shear}} = \sqrt{\frac{G}{\rho}}.$$

Shear moduli in soft tissue are in the order of $G = 4$ kPa, implying that the shear wave speed is about 2 m s^{-1}, compared to 1500 m s^{-1} for longitudinal acoustic waves.

13.2 Properties of the Wave Equation

The parameter c in the wave equation has units of speed. To appreciate its physical interpretation, consider the departure from the undisturbed pressure $p(x,t) = P(x,t) - P_0 =$

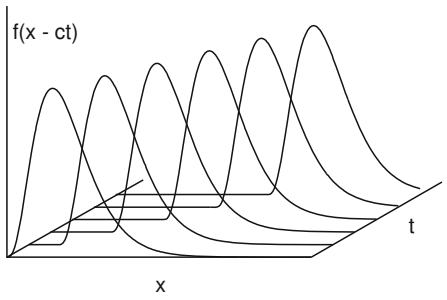


Fig. 13.4 A wave $f(x - ct)$ traveling to the right with speed c

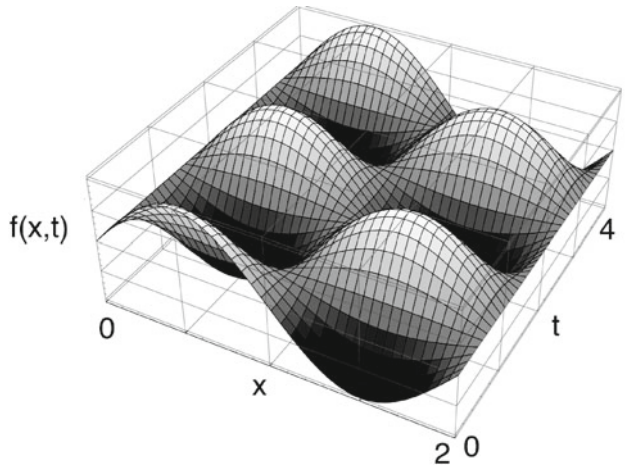


Fig. 13.5 A standing wave $f(x,t) = \sin \pi x \cos \pi t$, plotted for $0 < x < 2$ and $0 < t < 4$

$f(x - ct)$, where f is any function. This solution obeys the wave equation (see Problem 5). It is called a *traveling wave*. A point on $f(x - ct)$, for instance its maximum value, corresponds to a particular value of the argument $x - ct$. To travel with the maximum value of $f(x - ct)$, as t increases, x must also increase in such a way as to keep $x - ct$ constant. This means that the pressure distribution propagates to the right with speed c, as shown in Fig. 13.4. Solutions $p(x,t) = g(x + ct)$, where g is any function, also are solutions to the wave equation, corresponding to a wave propagating to the left

The wave speed c is one of the most important parameters governing the propagation of sound waves. The density of water is about $\rho_0 = 1000$ kg m^{-3} and the compressibility of water is approximately 5×10^{-10} Pa^{-1}, so the speed of sound in water is about 1400 m s^{-1}. The speed of sound in tissues is slightly higher: 1540 m s^{-1} is often taken as an average speed of sound in soft tissue. The speed of sound in air is about 344 m s^{-1}. See Denny (1993) for a more detailed comparison of the speed of sound in air and water.

One very useful traveling wave is $p(x,t) = p_0 \sin\left[\frac{2\pi}{\lambda}(x - ct)\right] = p_0 \sin\left[2\pi(\frac{x}{\lambda} - \frac{t}{T})\right] = p_0 \sin(kx - \omega t)$. The pressure distribution oscillates sinusoidally with frequency

$$f = c/\lambda \tag{13.16}$$

cycles per second (Hz) or angular frequency $\omega = 2\pi f$ (radians) s^{-1}. Equation 13.16 relates the frequency and wavelength. For instance, middle C has a frequency of 261.63 Hz. In air, the wavelength is (344 m s^{-1})/(261.63 Hz) = 1.315 m. The *wave number* is

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c}. \tag{13.17}$$

Standing waves such as

$$p(x,t) = p \cos(\omega t) \sin(kx) \tag{13.18}$$

are also solutions to the wave equation. An example is shown in Fig. 13.5. The standing wave in Eq. 13.18 has nodes fixed in space where $\sin(kx)$ is zero. Standing waves can occur, for example, in an organ pipe and in the ear canal (Problem 7).

A standing wave can also be written as the sum of two sinusoidal traveling waves, one to the left and one to the right. Conversely, two standing waves can be combined to give a traveling wave (Problem 8).

Since the fluid velocity v obeys the wave equation, it can also be represented as a sinusoidal wave. It is important to realize that the fluid oscillates back and forth. The fluid itself does not propagate with the wave. What propagates is the disturbance in the fluid. Sound in a fluid is a *longitudinal wave*, which means that the fluid oscillates along the same axis that the disturbance propagates (in this case, both move in the x direction). Other types of waves exist in nature, such as electromagnetic waves studied in Chap. 14. Electromagnetic waves are *transverse waves*, because the electric field oscillates in a direction perpendicular to the direction of wave propagation. Fluids cannot support significant shear stresses and only propagate longitudinal waves.

13.3 Acoustic Impedance

13.3.1 Relationships Between Pressure, Displacement and Velocity in a Plane Wave

For a plane wave traveling to the right, the pressure, displacement, and speed of the fluid have simple relationships. If the pressure change is

$$p(x,t) = p_0 \sin(kx - \omega t), \tag{13.19}$$

one can use Eqs. 13.8 and 13.12 to show that the fluid displacement is

$$\xi = \xi_0 \cos(kx - \omega t), \tag{13.20}$$

the fluid velocity is

$$v = v_0 \sin(kx - \omega t), \tag{13.21}$$

and the amplitudes are related by

$$\xi_0 = p_0 \frac{\kappa}{k} = p_0 \frac{\kappa\lambda}{2\pi} = p_0 \frac{\kappa c}{\omega}, \tag{13.22}$$

$$v_0 = \frac{p_0}{\rho_0 c} = \frac{p_0}{Z}. \tag{13.23}$$

The quantity $Z = \rho_0 c = \sqrt{\rho_0/\kappa}$ is called the *acoustic impedance* of the medium.[3] The acoustic impedance of water is about $(10^3 \text{ kg m}^{-3})(1400 \text{ m s}^{-1}) = 1.4 \times 10^6$ Pa s m^{-1}. The acoustic impedance of air is about 400 Pa s m^{-1}, so $Z_{\text{air}} \ll Z_{\text{water}}$ (Denny 1993).

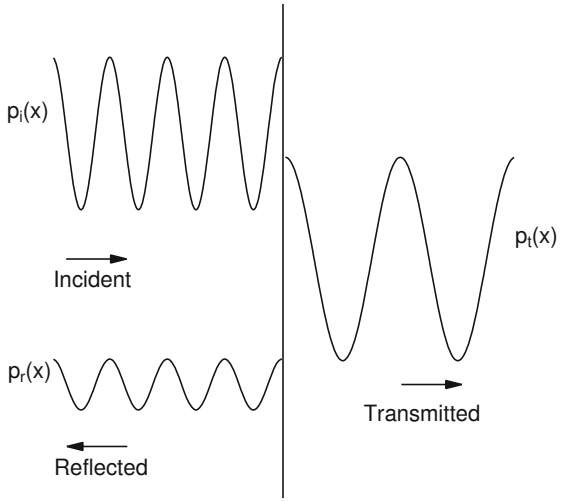


Fig. 13.6 A sound wave with pressure amplitude p_i traveling to the right is incident on a boundary separating tissue 1 on the left from tissue 2 on the right. Each tissue has a different density ρ_0 and compressibility κ. $Z_2 = 2Z_1$. Part of the wave is transmitted to the right with amplitude p_t, and part is reflected to the left with amplitude p_r. The drawing shows one instant in time

13.3.2 Reflection and Transmission of Sound at a Boundary

Consider next what happens at the boundary between two different media. Suppose a traveling wave is propagating to the right in a fluid with sound speed c_1 and acoustic impedance Z_1. At $x = 0$, it encounters a second fluid, with speed c_2 and impedance Z_2. In general, the interaction of the incoming wave with the boundary between the first and second fluids results in a reflected wave traveling to the left in fluid 1 and a transmitted (or refracted) wave traveling to the right in fluid 2 (Fig. 13.6). The acoustic impedances determine how much of the incoming wave is reflected and how much is transmitted. The waves must oscillate with the same frequency in both media. The pressure at the boundary must be the same in each medium, and the fluid velocity must also be continuous across the boundary. Let $p_i(x,t) = p_i \sin\left[\frac{\omega}{c_1}(x - c_1 t)\right]$, $p_r(x,t) = p_r \sin\left[\frac{\omega}{c_1}(x + c_1 t)\right]$, and $p_t(x,t) = p_t \sin\left[\frac{\omega}{c_2}(x - c_2 t)\right]$ be the incoming, reflected, and transmitted pressures. The velocities are related to the pressures by the acoustic impedances. At the boundary, the pressure and the velocity must be continuous. In fluid 1, the amplitude of the pressure is $p_i + p_r$, and in fluid 2, it is p_t. In fluid 1, the amplitude of the velocity is $(p_i - p_r)/Z_1$, and in fluid 2, it is p_t/Z_2. (The minus sign arises because the reflected wave is traveling to the left.) Therefore

$$p_i + p_r = p_t \tag{13.24}$$

and

$$(p_i - p_r)/Z_1 = p_t/Z_2. \tag{13.25}$$

We can solve these two equations for p_r and p_t in terms of p_i :

$$p_r = \frac{Z_2 - Z_1}{Z_2 + Z_1} p_i, \tag{13.26}$$

$$p_t = \frac{2Z_2}{Z_2 + Z_1} p_i. \tag{13.27}$$

The *intensity* I of a sound wave is a measure of the power per unit area (W m^{-2}). The instantaneous power per unit area transmitted by the wave in Eq. 13.19 at some point is

$$I(t) = p(t)v(t) = p_0 v_0 \sin^2 \omega t. \tag{13.28}$$

The average power per unit area is

$$I = \frac{1}{2} p_0 v_0 = \frac{1}{2} \frac{p_0^2}{Z}. \tag{13.29}$$

Problems 13–15 show that the *reflection* and *transmission coefficients* are

$$R = \frac{I_r}{I_i} = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1}\right)^2, \tag{13.30}$$

[3] Strictly speaking, the acoustic impedance is the ratio $Z = p_0/v_0$, and carries information about both the amplitude ratio and the relative phase of the pressure and velocity. If the waves are in phase, Z is said to be *resistive*; if they are $\pi/2$ out of phase, Z is said to be *reactive*. The *characteristic acoustic impedance* is a property of the medium: $Z_0 = \rho_0 c$. Both have units Pa m s^{-1} or kg m^{-2} s^{-1}. For a plane wave, the impedance is resistive and $Z = Z_0$. For other waves, such as standing waves, there is a reactive component.

and

$$T = \frac{I_t}{I_i} = \frac{4Z_1Z_2}{(Z_1 + Z_2)^2}, \tag{13.31}$$

and that $R + T = 1$.

If the acoustic impedance of the two fluids is the same, $Z_1 = Z_2$, there is no reflected wave and the entire incoming wave is transmitted. If $Z_1 \ll Z_2$ (e.g., sound going from air to water), almost all of the sound is reflected.

13.4 Comparing Intensities: Decibels

13.4.1 The Decibel

When comparing two intensities, the range of differences is often so great that a logarithmic comparison scale is used. We first saw the decibel when discussing the frequency response of a linear system in Chap. 11. Intensity levels in dB have meaning only in terms of ratios:

$$\text{Intensity Difference (dB)} = 10 \log_{10}\left(\frac{I_2}{I_1}\right). \tag{13.32}$$

The intensity difference can also be written in terms of pressure (or displacement or velocity) ratios:

$$\begin{aligned}\text{Intensity Difference (dB)} &= 10\log_{10}\left(\frac{I_2}{I_1}\right)\\ &= 10\log_{10}\left(\frac{p_2}{p_1}\right)^2\\ &= 20\log_{10}\left(\frac{p_2}{p_1}\right).\end{aligned} \tag{13.33}$$

This assumes that p_1 and p_2 are measured in the same medium, so the acoustic impedance does not change. If the intensity of a wave falls to 1% of its original value, the intensity difference is $10\log_{10}(0.01) = -20$ dB.

13.4.2 Measuring Hearing Response

In auditory acoustics, intensities are measured with respect to a reference intensity $I_0 = 10^{-12}$ W m^{-2}. This is the intensity of the faintest sound that a person can typically hear:

$$\text{Intensity level} = 10\log_{10}\left(\frac{I}{I_0}\right). \tag{13.34}$$

A sound that is ten times as intense as the threshold for hearing has an intensity level of 10 dB. A sound with an average intensity $I = 1$ W m^{-2} is perceived as painful, so the threshold for pain has an intensity level of about 120 dB. Table 13.1 gives the intensity in decibels for some common sounds.

Table 13.1 Approximate intensity levels of various sounds

Sound	Intensity (W m^{-2})	Level (dB, A weighting)
Rocket launch pad	10^5	170
	10^4	160
	10^3	150
	10^2	140
F-84 jet at takeoff, 25 m from the tail; Large pneumatic riveting machine (1 m); Boiler shop (maximum level); Peak sound level at a rock concert	10	130
Sound that produces pain	1	120
Woodworking shop	10^{-1}	110
Near a pneumatic drill ("jack hammer")	10^{-2}	100
Inside a motor bus	10^{-3}	90
Urban dwelling near heavy traffic	10^{-4}	80
Busy street	10^{-5}	70
Speech at 1 m	10^{-6}	60
Office	10^{-7}	50
Average dwelling	10^{-8}	40
Maximum background sound level tolerable in a broadcast studio	10^{-9}	30
Whisper; maximum background sound level tolerable in a motion picture studio	10^{-10}	20
	10^{-11}	10
Minimum perceptible sound	10^{-12}	0

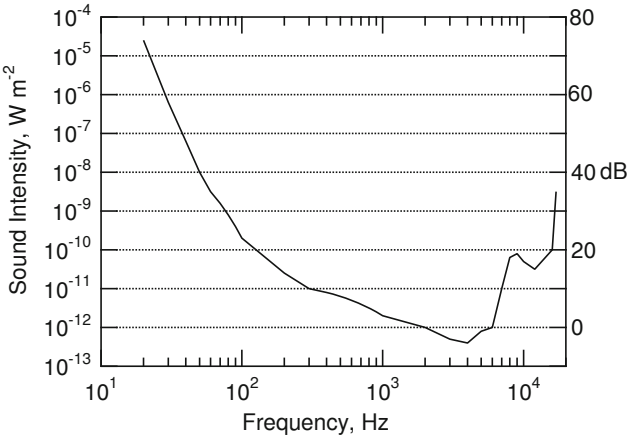


Fig. 13.7 Hearing response (MAF) curve for a young adult

The sensitivity of the ear depends on frequency. A typical hearing response curve for a young person is shown in Fig. 13.7. The *minimum auditory field* (MAF) is measured with a loudspeaker; the slightly different *minimum auditory pressure* (MAP) is measured with headphones. The ear is most sensitive to sounds between about 100 and 5000 Hz. A sound at 20 Hz will not be perceived to be as loud as one at 1000 Hz with the same intensity. Commercial sound level meters typically have two weightings. The "C" weighting has almost the same sensitivity at all frequencies. The "A" weighting more nearly mimics the response of the normal

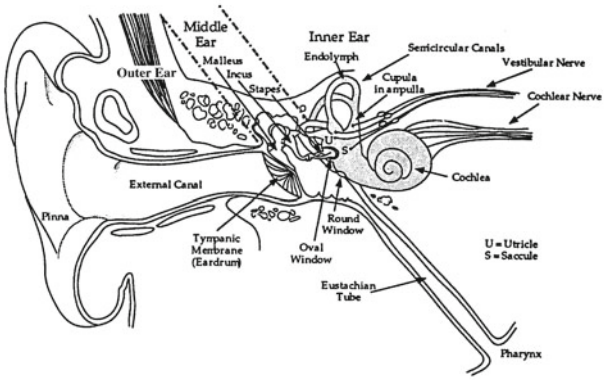


Fig. 13.8 A cross section of the ear. From Cameron et al. 1999. Used by permission

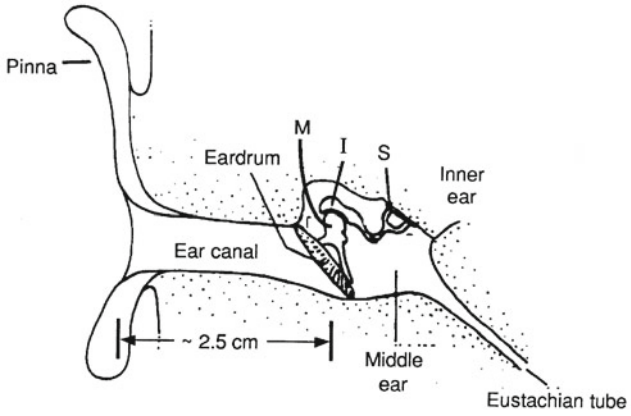


Fig. 13.9 Details of the middle ear. From Cameron et al. 1999. Used by permission

ear. Sounds with the same level when the meter is on "A" weighting will be perceived as having the same loudness.

13.5 The Ear and Hearing

A cross-section of the ear is shown in Fig. 13.8. The ear can be thought of as having three different sections, each with a unique purpose: the *external ear* gathers sound, the *middle ear* transfers energy from the air (low acoustic impedance) to the liquid of the inner ear (high acoustic impedance), and the *inner ear* transforms the signal into nerve impulses going to the brain.

The external ear consists of the *pinna*, the visible part of the ear, and an air-filled tube called the *ear canal*.

The middle ear is a small chamber filled with air that contains three small bones, or *ossicles* shown in Fig. 13.9. It is separated from the ear canal by the *ear drum*. The bone in contact with the ear drum is called the *malleus* (it is shaped a bit like a mallet or hammer). The next bone is the *incus* (from the Latin for an anvil, which it resembles slightly). The third, in contact with the *oval window* to the inner ear, is the *stapes* (again from the Latin, for stirrup.) The *eustachian tube* leads from the middle ear to the mouth and throat (*nasopharynx*). Since the ear is sensitive to very small pressure changes, the eustachian tube serves the important function of keeping the pressure on both sides of the ear drum the same for slow changes, such as when we climb stairs or the weather changes. The walls of the eustachian tube are often collapsed together. Swallowing helps to open them up and equalize the pressure if necessary.

Sound arrives at the ear as a vibration in air. Sound energy must enter the inner ear in order to be converted into a nerve signal to the brain. Yet, the inner ear is filled with liquid. The acoustic impedance of the liquid in the inner ear is about 3500 times larger than the acoustic impedance of air. This means that without the impedance transformation by the middle ear, the intensity in the inner ear would be only about $1/1000$ of the intensity in air—a loss of about 30 dB (Problem 14).

The middle ear transforms the impedance by two mechanisms. The first is a simple area change. The ear drum vibrates in response to the pressure changes in the sound wave. If a sound wave with pressure amplitude p_{air} impinges on the ear drum of area $S_{\text{ear drum}}$, the total excess force on the ear drum is $F = p_{\text{air}} S_{\text{ear drum}}$. For this simplest model, assume that the three bones behave like a single rigid rod and there are no effects of the boundary at the circumference of the ear drum. Then, the bones transmit this force to the membrane at the oval window, which has area $S_{\text{oval window}}$. The pressure induced in the liquid in the inner ear is then $p_{\text{inner ear}} = F/S_{\text{oval window}} = p_{\text{air}} S_{\text{ear drum}}/S_{\text{oval window}}$. The area of the ear drum is about $S_{\text{ear drum}} = 64\ \text{mm}^2$, while the area of the base of the stapes is $3.2\ \text{mm}^2$ (Newman 1957). Therefore, $p_{\text{inner ear}} = 20 p_{\text{air}}$. Actually, the ear drum and the membrane at the oval window are not connected by a simple rigid rod. The malleus, incus, and stapes are pivoted in such a way that they serve as a set of levers multiplying the force at the oval window by an additional factor of 1.3. Therefore, the total pressure amplification by the middle ear is 26. This corresponds to a 28 dB change in sound intensity, which almost compensates for the 30 dB loss going from air to the liquid of the inner ear. The bones of the middle ear have muscles that change their stiffness, so they can reduce the amount of pressure amplification to protect the inner ear from very loud, low-frequency noises.

The inner ear contains three *semicircular canals*, which help control our sense of balance, and the *cochlea*, which changes the sound to nerve impulses. All are filled with liquid. The cochlea is a small spiral about the size of the tip

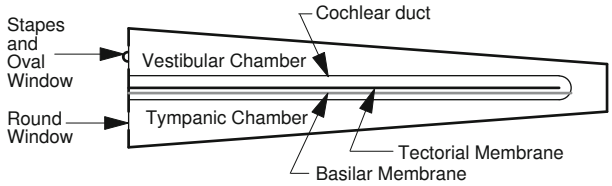


Fig. 13.10 A schematic representation of the cochlea

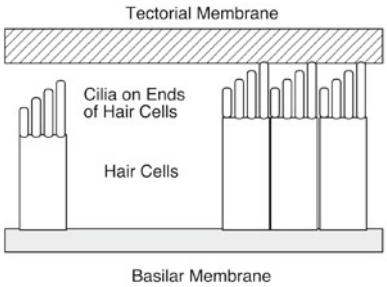


Fig. 13.11 A cross section of the cochlea. The hair cells are deformed as the basilar membrane moves relative to the tectorial membrane

of your little finger. Unwound, it is about 3 cm long. Figure 13.10 shows it schematically. There are three chambers. The *vestibular chamber* connects to the stapes in the middle ear through the oval window. At the other end of the cochlea the vestibular chamber connects to the *tympanic chamber*. The *round window* opens onto the middle ear and allows the pressure to be equalized at low frequencies. The third chamber is the *cochlear duct*.

When the stapes moves the oval window, it generates a sound wave that travels through the liquid in the cochlea. This produces a displacement of the *basilar membrane* in the third chamber, the cochlear duct. Two types of *hair cells* sit on the basilar membrane: one row of inner hair cells and three rows of outer hair cells. The hair cells in turn have very fine "hairs" on them, called *cilia*. The cilia of the outer hair cells touch another membrane, the *tectorial membrane*, but the cilia of the inner hair cells do not. A cross-section of this is shown schematically in Fig. 13.11. When the basilar and tectorial membranes are displaced by the sound wave, the cilia on the inner hair cells move in the liquid that fills the region between the membranes. It is just as if you submerged your head in a swimming pool and shook it back and forth. Your head would move in the water, but the motion of your hair would be altered as the water "dragged" it. As a result of this motion of the cilia in the liquid, the inner hair cells generate nerve impulses that then travel to the brain and provide our sensation of sound. The mechanism was discussed briefly in Sect. 9.9.

Figure 13.10 shows that the size of the cochlea changes along its length. So does its stiffness. Different locations along the cochlea therefore oscillate at different frequencies: points near the oval window (base or left side of Fig. 13.10) respond to high frequencies, while points near the apex (right side of Fig. 13.10) respond to low frequencies. The physics and neurophysiology of pitch perception and music is fascinating and complex (Sacks 2007; Hartmann 2013). For instance, some people have *absolute pitch*. They can identify a pitch (say, F-sharp) in isolation, just as normal people can identify a uniform color (say, yellow). Most people can only perceive relative pitch: one note has a higher pitch than another (for example, an E is a major third above a C).

The cochlear implant was mentioned in Chap. 7 as a way to use functional electrical stimulation to partially restore hearing. A row of electrodes is inserted along the cochlea to stimulate the nerves that are usually excited by the hair cells. Some pitch perception can be restored by performing a Fourier analysis of a sound and stimulating neurons at different places along the cochlea.

13.6 Attenuation

A plane wave of sound propagating through a medium is *attenuated*: there is a decrease in intensity because of dissipative factors such as viscosity and heat conduction, which we did not include in Sect. 13.1. The attenuation is exponential. The *amplitude attenuation coefficient*[4] is defined by

$$\alpha = -\frac{1}{p}\frac{dp}{dx}, \tag{13.35}$$

where x is the distance the wave travels in the medium. The sound pressure amplitude decays exponentially:

$$p(x) = p(0)e^{-\alpha x}. \tag{13.36}$$

Since the intensity is proportional to p^2,

$$I(x) = I(0)e^{-2\alpha x}. \tag{13.37}$$

The *intensity attenuation coefficient* is $\mu = 2\alpha$. In acoustics, the attenuation is usually expressed in decibels per meter, which is then independent of whether μ or α is used.[5]

The wave equation for acoustics is an approximation, because the basic equations of fluid dynamics are nonlinear. Therefore, effects that we have ignored, such as waveform distortion, the generation of harmonics, and increased attenuation may occur, particularly at high sound intensities.

[4] ICRU 61 (1998).

[5] Sometimes the attenuation coefficient is expressed in nepers m^{-1}, in which case the natural logarithm of the intensity or pressure ratio is used.

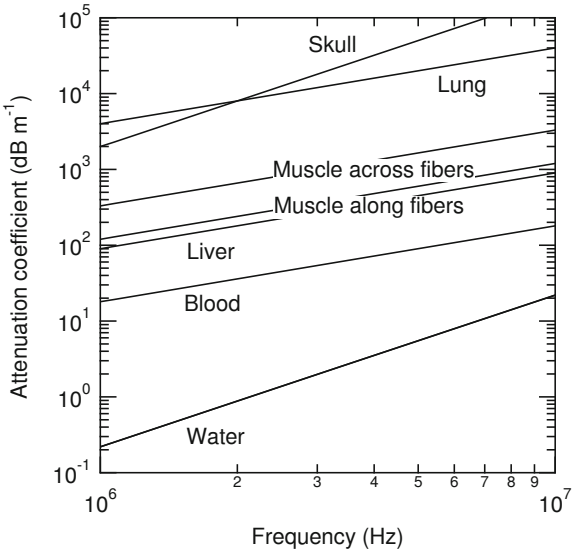


Fig. 13.12 Representative values of the attenuation coefficient for ultrasound

In air, the attenuation depends on the frequency of the sound and the temperature and humidity of the air (Lindsay and Beyer 1989; Denny 1993). Sound that we can hear (in the frequency range of 20 Hz to 20 kHz) is attenuated by about 0.1–10 dB km^{-1}. Water transmits sound better than air, but its attenuation is an even stronger function of frequency. It also depends on the salt content. At 1000 Hz, sound attenuates in fresh water by about 4×10^{-4} dB km^{-1}. The attenuation in sea water is about a factor of ten times higher (Lindsay and Beyer 1989). The low attenuation of sound in water (especially at low frequencies) allows aquatic animals to communicate over large distances (Denny 1993).

The attenuation of sound depends strongly on frequency. Figure 13.12 shows some representative values. As a rule of thumb, at ultrasonic frequencies the attenuation is proportional to frequency, with the constant of proportionality being 100 dB m^{-1} MHz^{-1}. There are large variations in attenuation in tissue, depending on the age of the subject and other factors. Values can be found in Appendix A of ICRU 61 (1998).

There can also be scattering of the sound from some object, just as there is for light. The total scattering cross-section for the object is defined by

$$\sigma_s = \frac{W_s}{I_0}, \tag{13.38}$$

where W_s is the total power scattered and I_0 is the incident intensity. As in Chap. 14, the differential scattering cross-section can also be defined. Scattering can be increased by using an ultrasound contrast agent (Faez et al. 2013).

13.7 Diagnostic Uses of Ultrasound

Ultrasound has several uses in medicine. The most common is to provide diagnostic images that complement those made with x-rays, nuclear medicine, and magnetic resonance.[6] Ultrasound does not provide the image quality of these other methods, and it is susceptible to artifacts (see Problems 28–31), but it can be performed in real time, at low cost, with a small instrument at the patient's bedside. In general, the different medical imaging techniques compete with one another. Each has its own advantages and disadvantages (Glide-Hurst et al. 2010).

The highest frequency sounds that we can hear ($\approx$ 15 kHz) have a wavelength in water of 0.1 m. One property of waves is that diffraction limits our ability to produce an image. Only objects larger than or approximately equal to the wavelength can be imaged effectively. This property is what limits light microscopes to resolutions equal to about the wavelength of visible light, 500 nm. If we used audible sound to form images, our resolution would be limited to about 0.07 m, which would be a poor image indeed. To overcome this difficulty, higher frequencies (ultrasound) are used. Typically, diagnostic ultrasound uses frequencies on the order of 1 to 15 MHz, corresponding to wavelengths of 1.4 to 0.1 mm in tissue. Higher frequencies would result in even shorter wavelengths, but higher frequency sound has increased attenuation, which ultimately sets an upper bound to the useful frequency.

13.7.1 Ultrasound Transducers

Ultrasound is typically produced using a *piezoelectric transducer*. A piezoelectric material converts a stress (or pressure) into an electric field, and vice versa. A high-frequency oscillating voltage applied across a piezoelectric material creates a sound wave at the same frequency. Conversely, an oscillating pressure applied to a piezoelectric material creates an oscillating voltage across it. Measurement of this voltage provides a way to record ultrasonic waves. Thus, the same piezoelectric material can serve as both source and detector. One piezoelectric material often used in medical transducers is lead zirconate titanate (PZT). Its density is 7.5×10^3 kg m^{-3}, the speed of sound in the material is 4065 m s^{-1}, and the acoustic impedance is 30×10^6 Pa s m^{-1}. About half of the electrical energy is converted to sound energy, and vice versa.

[6] See Kremkau (2006); Carson and Fenster (2008) or Wells (2006).

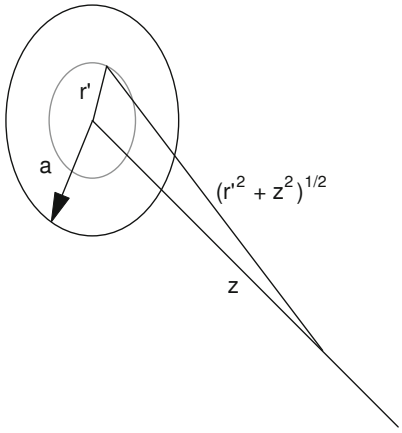


Fig. 13.13 Coordinate system for calculating the intensity of sound radiated from a transducer of radius a. The z axis passes through the center of the transducer and is perpendicular to it

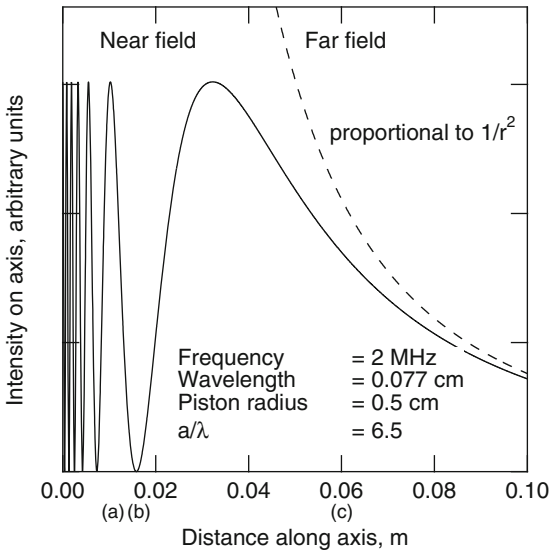


Fig. 13.14 The sound intensity on the axis of a circular transducer. The sound frequency is 2 MHz, and the transducer radius is 0.5 cm. Both near- and far-field regions are shown. Labels (**a**), (**b**), and (**c**) show the positions of the transverse radial scans in Fig. 13.15

There are some important features of the radiation pattern from a transducer that we review next. Consider a circular transducer or piston, the surface of which is oscillating back and forth in a fluid. Both faces set up disturbances in the fluid; however, we consider the radiation from only one face, since the transducer is placed in a holder which prevents radiation from the rear surface. We can easily calculate the intensity along the z-axis, which we set up perpendicular to the piston and passing through its center, as shown in Fig. 13.13.

The displacement of the face of the transducer, ξ, is the same as the displacement of the fluid in contact with it. The entire face of the piston, and therefore the fluid immediately in front of it, vibrates with a fluid velocity $d\xi/dt = v_0 \cos \omega t$.[7] Each small element of the vibrating fluid creates a wave that travels radially outward, the points of constant phase being expanding hemispheres. The amplitude of each spherical wave decreases as $1/r$, the intensity falling as $1/r^2$. We want the pressure at a point z on the axis of the transducer. It is obtained by summing up the effect of all the spherical waves emanating from the face of the transducer. At time t the phase of the wave is the same as the phase of the wave leaving the annular ring $r'dr'$ at the earlier time $t - r/c$:

$$p \propto \frac{d\xi(z,t)}{dt} \propto \int_0^a 2\pi r' dr' \frac{\cos[\omega(t - r/c)]}{r}.$$

This is easily evaluated by changing variables. Since $r^2 = r'^2 + z^2$, $2r dr = 2r' dr'$:

$$p \propto 2\pi \int_{r=z}^{r=\sqrt{a^2+z^2}} r dr \frac{\cos[\omega(t - r/c)]}{r}$$

$$= \frac{2\pi}{k}\left[\sin[\omega(t - \frac{1}{c}\sqrt{a^2+z^2})] - \sin[\omega(t - z/c)]\right].$$

To find the average intensity, we square and average over one period. The result is

$$I \propto \sin^2\left[\frac{\omega}{2c}\left(z - \sqrt{a^2+z^2}\right)\right]. \tag{13.39}$$

The result is plotted in Fig. 13.14 for a fairly typical but small transducer ($a = 0.5$ cm, $f = 2$ MHz).

There are several important features of Fig. 13.14. Close to the transducer there are large oscillations in intensity along the axis; there are corresponding oscillations perpendicular to the axis, as shown in Fig. 13.15. The maxima and minima form circular rings. This is called the *near field* or *Fresnel zone*. Further away the intensity falls as $1/r^2$, in the *far field* or *Fraunhoffer zone*. The depth of the Fresnel zone is approximately a^2/λ. For the example shown (2 MHz, transducer diameter 1 cm), the depth is about 3 cm; for a larger transducer or higher-frequency ultrasound, it would be greater.

In the far field, approximations can be made to simplify the calculation. The intensity is then given by

$$I \propto \frac{1}{r^2}\left(\frac{J_1(ka\sin\theta)}{ka\sin\theta}\right)^2. \tag{13.40}$$

Function $J_1(x)$ is the *Bessel function of order 1*. It is found in math tables and is available in many spreadsheets. The angular dependence of the far-field intensity is plotted in Fig. 13.16. If you want the ultrasound to be transmitted

[7] We use $d\xi/dt$ because it is in phase with the excess pressure.

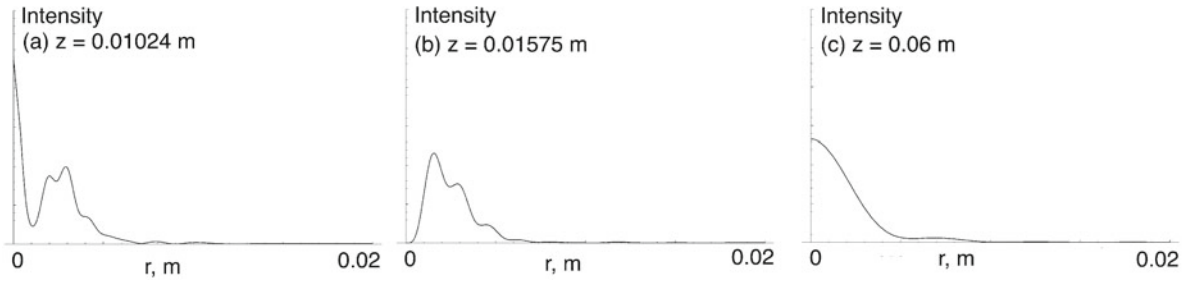


Fig. 13.15 Scans across the beam from the transducer shown in Fig. 13.14. **a** In the near field at an on-axis maximum 0.01024 m from the transducer. **b** In the near field at an on-axis minimum 0.01575 m from the transducer. **c** In the far field 0.060 m from the transducer

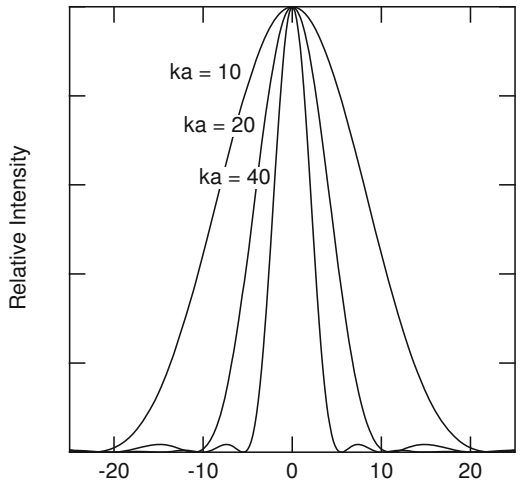


Fig. 13.16 The far-field intensity as a function of angle, calculated from Eq. 13.40. The value $ka = 10$ corresponds to 1 MHz and transducer radius $a = 0.25$ cm. The value $ka = 20$ corresponds to $f = 2$ MHz and $a = 0.25$ cm or $f = 1$ MHz and $a = 0.5$ cm. Value $ka = 40$ corresponds to 4 MHz and $a = 0.25$ cm or $f = 2$ MHz and $a = 0.5$ cm, the case examined in Fig. 13.14

mainly in the direction normal to the face of the transducer, select a transducer with a larger radius. If the transducer is used as a detector, a larger transducer is more selective in the direction perpendicular to its face. By shaping the face of the transducer, it is possible to bring the beam to a focus at some particular depth. This improves the spatial resolution and increases the strength of the returning echo. Ultrasound imaging may be done in the near field, the far field, or the transition region. Modern transducers typically consist of an array of transducers which may lie on a straight or curved line. They can be driven in such a way as to produce waves that come to a focus, or that travel in an off-axis direction (see Hendee and Ritenour 2002 or Fig. 13.18).

The impedance of a typical transducer is about 30×10^6 Pa s m^{-1}, so it is necessary to have an impedance-matching material between the transducer and the patient's skin (see Problem 16).

13.7.2 Pulse Echo Imaging

Most ultrasonic imaging is based on a pulse-echo technique. A short pulse (typically 0.5 μs in duration with a central frequency of about 5 MHz) is applied to the tissue by a piezoelectric transducer. The pulse travels with a speed of about $c = 1540$ m s^{-1} (or 1.54 mm μs^{-1}). Whenever it approaches a boundary between two tissues having different acoustic impedances, part of the incident pulse is reflected as an echo, which can be detected by the same piezoelectric transducer. The longer the time Δt between the generation and detection of the pulse, the farther away the reflecting boundary. In general, the distance from the source to the boundary is $\Delta x = c\Delta t/2$. Multiple boundaries produce multiple echoes, with each echo corresponding to a different distance from the source to boundary. A plot of echo intensity versus time is called an *A scan*. An A scan of the eye is shown in Fig. 13.17. As the attenuation is high, it is customary to increase the gain of the receiving amplifier as the echo time increases.

To form a two-dimensional image, it is necessary to scan in many different directions. In a *B scan* the brightness of the screen corresponds to the intensity of the echo, plotted versus position in the body in the plane of the scan. The B-scan transducer sends a narrow beam into the body. The direction of the beam is rapidly changed to cover a fan-shaped region of the body. This can be done with an oscillating or rotating transducer head (often containing three transducers), with an array of transducers that are pulsed sequentially, or with a *phased array* of transducers that are pulsed together. The operation of sequential pulsing or a phased array can be understood by referring to Fig. 13.18. The basic principle of using

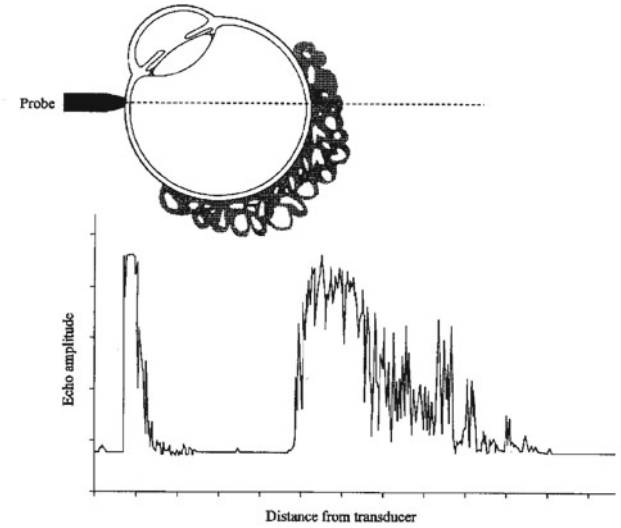


Fig. 13.17 An A scan of the eye. From ICRU 61, p. 2. Used by permission

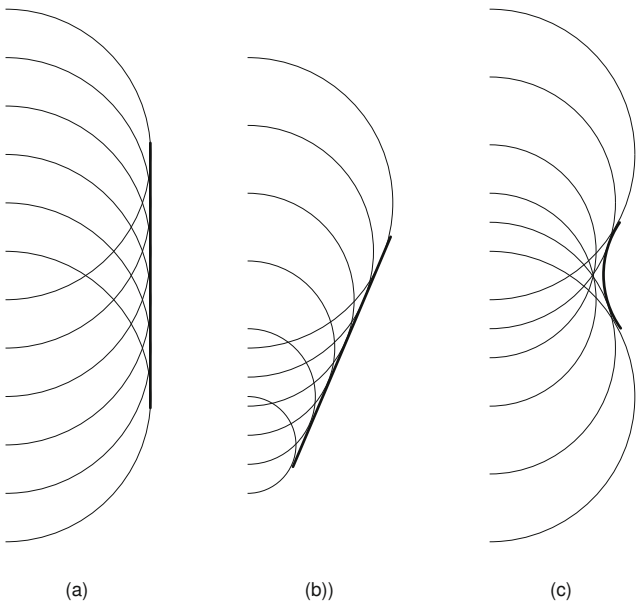


Fig. 13.18 How a phased array or delayed-pulse array works. Five transducers have been pulsed; the semi-circles show the propagating lines of constant phase from each one. The thick lines show the advancing wave front. In (**a**) all five transducers have been pulsed at the same time. The signals from each transducer add along the plane wave front traveling to the right. In (**b**) the top transducer was pulsed first. Each lower transducer was pulsed at successively later times, so the pulses have not traveled as far. This steers the beam downward. In (**c**) the outer transducers were pulsed first. As one goes inward, each transducer was pulsed later than the one before. This focuses the beam. The same technique can be used to steer or focus the sensitivity to the scattered wave during detection



Fig. 13.19 A B scan of a 16-week fetus

multiple one-dimensional (x) echo scans along different lines through the body is explored in Problems 24 and 25.

Two-dimensional ultrasound is widely used in diagnostic medicine; for instance in monitoring the fetus during pregnancy. Figure 13.19 shows a typical ultrasound image of a fetus.

Other imaging methods include motion or *M mode* to observe the beating heart as a function of time, and detecting sound backscattered from structures in an organ that are smaller than a wavelength.

As the tissue response to high intensity ultrasound is nonlinear, harmonics of the original ultrasound pulse are generated in the tissue; the second-harmonic signals are used to form *harmonic images*.

13.7.3 The Doppler Effect

When the source of an ultrasound wave is moving, the frequency of the wave observed by a stationary receiver is different than the frequency of the source. This phenomenon is called the *Doppler effect*. When the source is moving toward the receiver, the frequency is higher, and when the source moves away from the receiver, the frequency is lower.

To see why this happens, consider the source moving to the right with speed v_s in a fluid for which the speed of sound is c. At $t = 0$, the source emits the crest of a wave with period T (frequency $f = 1/T$). The wave travels to the right. This crest takes a time $t = L/c$ to reach a stationary receiver a distance L away. At $t = T$, one period later, another crest is emitted by the source. This crest takes less time to reach the receiver because the source has moved closer to the receiver. Specifically, the distance from source to receiver is now $L - v_s T$, so the crest reaches the receiver at $t = T + (L - v_s T)/c$. The time T' between crests reaching the receiver is $T' = T + (L - v_s T)/c - L/c = T(1 - v_s/c)$. The frequency

observed by the receiver is

$$f' = 1/T' = \frac{f}{1 - v_s/c}. \tag{13.41}$$

If the source is moving toward the receiver with a speed equal to 10% of the speed of sound, then f' is about 11% higher than f. When the source is moving away from the receiver, $f' = f/(1 + v_s/c)$ (see Problem 33). It is not difficult to include the effect of motion of the reflecting surface at an angle with the ultrasound beam.

In medical ultrasound applications, the detected wave is often a reflection from moving tissue, such as red blood cells. In this case, the relationship between the frequency f produced by a stationary source and the frequency f' received by the stationary receiver after reflection from an object moving away from it at speed v_o is (see Problem 34)

$$f' = f\frac{1 - v_o/c}{1 + v_o/c}. \tag{13.42}$$

The difference in frequency between f and f' contains information about the speed of the object (Problem 36). Doppler ultrasound is used in medicine to measure speed, such as the speed of moving blood cells. Often the Doppler shift is measured for a pulse of ultrasound, so that one can be sure of the depth at which the Doppler shift occurred. A distribution of red cell velocities can be measured by looking at the Doppler shift frequency spectrum. In *color flow imaging* the velocity information from Doppler imaging is superimposed on a B-scan ultrasound image.

13.7.4 Elastography

Several techniques are being developed to measure the elastic properties of tissue. For example, an A-mode signal is measured with and without a static force on the tissue; the slight changes in signal reflect changes in tissue density.

In *shear wave elastography*, the shear wave propagates so slowly that images of the displacement $\eta(x, t)$ can be measured using traditional ultrasound techniques. The distribution of wave speeds can be determined from these images, and then the distribution of the shear modulus can be calculated. This method has been used to analyze breast cancer tumors, which tend to have a higher shear modulus than the surrounding healthy breast tissue (Berg et al. 2012).

13.7.5 Safety

The skin intensities used in diagnostic ultrasound range from 0.1 W m^{-2} for an obstetric examination to 25,000 W m^{-2} for some procedures that image the heart or blood vessels. These intensities occur over a small area of the body and for a limited period of time. Many studies have been done to see if any harm results from these sound intensities. No harmful effects have been found.

13.8 Therapeutic Uses of Ultrasound

The primary potential causes of harm from ultrasound are also used for therapy. They are *diathermy,* the heating of the tissue because of the energy deposited, and *cavitation*, a process in which very high intensity sound waves cause tiny bubbles of steam to form and then collapse violently. Cavitation requires intensities of 3.5×10^7 W m^{-2} or more.

High-intensity-focused ultrasound is made possible by phased arrays. It is possible to accurately locate the focal spot by applying lower intensity pulses that only heat the tissue a few degrees. The temperature in the tissue is mapped using magnetic resonance imaging, described in Chap. 18. Once the focal region is in the desired target, a longer pulse is applied to heat the tissue to the desired temperature. The technique is called magnetic resonance guided focused ultrasound (MRgFUS) or magnetic resonance guided high intensity focused ultrasound (MRgHIFUS). This technique is used for fat reduction (Saedi and Kaminer 2013), breast (Merckel et al. 2013), and prostate cancer, relief of pain from metastatic cancer, and neurosurgery.

The neurosurgical use of focused ultrasound has an interesting history. It was first proposed in the 1940s, but the large impedance difference between skull and soft tissue meant that a portion of the skull had to be removed in order for the ultrasound to reach the brain. Focus was achieved with a plastic lens in front of the transducer. A special water-filled container coupled the transducer to the surface of the brain. The development of phased transducer arrays made it possible to focus without a special lens and also to make corrections for the ultrasound waves passing through the skull (Clement and Hynynen 2002). There is high energy absorption at the skull, so transducer arrays covering a large part of the skull are used, and the water that couples the array to the scalp is cooled. Some patients have been treated for chronic neuropathic pain (Jeanmonod et al. 2012). Other uses are reviewed by Monteith et al. (2013) and Ellis et al. (2013).

Another use of ultrasound is *lithotripsy,*[8] the destruction of kidney stones using sharply focused ultrasound. Lithotripsy uses extremely intense, pulsed ultrasound waves. The peak intensity is about 3.8×10^8 W m^{-2}. The sound is intense enough so that bubbles of steam form and then collapse. When they collapse near the surface of the stone they

[8] *Litho-* means stone.

"hammer" on the stone. With repeated blows, the stone shatters. These smaller pieces may pass in the urine, avoiding surgery.

Symbols Used in Chap. 13

Symbol	Use	Units	First used page
a	Transducer radius	m	372
c	Speed of sound	m s^{-1}	364
c_{shear}	Speed of shear wave	m s^{-1}	365
f, g, h	Arbitrary functions		366
f, f'	Frequency	Hz	366
k	Wave number	m^{-1}	366
l	Length	m	364
p	Excess pressure	Pa	364
s_n	Normal stress	Pa	364
s_s	Shear stress	Pa	365
r, r'	Position	m	372
t	Time	s	363
v	Fluid or particle velocity	m s^{-1}	365
v_s, v_o	Velocity of source, observer	m s^{-1}	375
x, y, z	Position	m	363
E	Young's modulus	Pa	364
F	Force	N	364
G	Shear modulus	Pa	365
I	Intensity	W m^{-2}	367
J_1	Bessel function of order 1		372
L	Distance	m	374
M	Mass	kg	365
P	Pressure	Pa	364
R	Reflection coefficient		367
S	Area	m^2	364
T	Transmission coefficient		367
T	Period	s	375
V	Volume	m^3	364
W_s	Power scattered	W	371
Z	Acoustic impedance	Pa s m^{-1} or kg m^{-2} s^{-1}	367
α	Amplitude attenuation coefficient	m^{-1}	370
κ	Compressibility	Pa^{-1}	364
ϵ_n	Normal strain		364
ϵ_s	Shear strain		365
λ	Wavelength	m	366
η	Displacement from equilibrium in a shear wave	m	365
μ	Intensity attenuation coefficient	m^{-1}	370
ρ	Density	kg m^{-3}	364
σ	Scattering cross section	m^2	371
θ	Angle		372
ξ	Displacement from equilibrium	m	363
ω	Angular frequency	s^{-1}	366

Problems

Section 13.1

Problem 1. Show that $1/\sqrt{\rho_0 \kappa}$ has units of speed.
Problem 2. Show that the pressure p satisfies the wave equation. Hint: Use Eqs. 13.13 and 13.14. Differentiate to obtain $\partial^2 p/\partial x^2$ and $\partial^2 p/\partial t^2$. Also use the fact that when multiple partial derivatives are taken, the order of differentiation can be interchanged (Appendix N).
Problem 3. Show that v and ρ also satisfy the wave equation.
Problem 4. Derive Eq. 13.15.

Section 13.2

Problem 5. Use the chain rule, with $u = x - ct$, to show that $f(x - ct)$ obeys the wave equation for any function f. Show that $g(x + ct)$ also obeys the wave equation.
Problem 6. Calculate the wavelength in air for the lowest audible frequency (20 Hz for most people) and the highest audible frequency (20 kHz for most young people).
Problem 7. The ear canal is about 2.5 cm long. It is open to the air at one end and closed by the ear drum at the other. This can cause a standing wave to form, which has a pressure node (zero amplitude) at the opening and pressure maximum at the ear drum. What is the longest wavelength of a standing wave that is set up? What frequency does this correspond to? Compare this to the most sensitive frequency of the ear (Fig. 13.7).
Problem 8. Use the trigonometric identity $\sin(a \pm b) = \sin a \cos b \pm \cos a \sin b$ to show that a traveling wave can be written as the sum of two out-of phase standing waves, and that a standing wave can be written as the sum of two oppositely-propagating traveling waves.

Section 13.3

Problem 9. Derive the relationships between p_0, ξ_0, and v_0 (Eqs. 13.22 and 13.23), where p_0, ξ_0, and v_0 are the amplitudes of a sinusoidally varying plane wave.
Problem 10. For the following five tissues, calculate the density and compressibility (data are from Hendee and Ritenour 2002).

Tissue	Z (Pa s m^{-1})	c (m s^{-1})
Fat	1.38×10^6	1475
Brain	1.55×10^6	1560
Blood	1.61×10^6	1570
Muscle	1.65×10^6	1580
Bone	6.10×10^6	3360

Problem 11. Show that the intensity of a sound wave (Eq. 13.29) can be written as $\frac{1}{2}ZV^2$, as $\frac{1}{2}PV$, or as $\frac{1}{2}\frac{P^2}{Z}$.

Problem 12. The threshold for audible sound is $10^{-12}\,\mathrm{W\,m^{-2}}$. Use Eq. 13.29 to convert this to the amplitude of the pressure oscillation in air, using $Z_{\mathrm{air}} = 400\,\mathrm{Pa\,s\,m^{-1}}$. Compare this to $10^5\,\mathrm{Pa}$ (atmospheric pressure), and to $5 \times 10^{-6}\,\mathrm{Pa}$ (which is on the order of the amplitude of random pressure variations in the air due to thermal motion). Are the pressure oscillations small? Perform the same analysis for the threshold for pain, $I = 1\,\mathrm{W\,m^{-2}}$.

Problem 13. When an incident sound wave in fluid 1 encounters the boundary with fluid 2, the reflection coefficient, R, is defined as the fraction of the incident intensity that is reflected. Derive an expression for R in terms of Z_1 and Z_2. Use the data in Problem 10 to calculate what fraction of the incident intensity is reflected at the boundary going from muscle to fat. Do the same for the boundary going from fat to muscle.

Problem 14. When an incident sound wave in fluid 1 encounters the boundary with fluid 2, the transmission coefficient, T, is defined as the fraction of the incident intensity that is transmitted. Derive an expression for T in terms of Z_1 and Z_2. Hint: recall that fluids 1 and 2 are different, so that the value of Z in Eq. 13.29 is different for the incident and transmitted waves.

Problem 15. Use the results of Problems 13 and 14 to show that $R + T = 1$.

Problem 16. (a) Show that when sound goes from a transducer with $Z_{\mathrm{transducer}} = 30 \times 10^6\,\mathrm{Pa\,s\,m^{-1}}$ to tissue with $Z_{\mathrm{tissue}} = 1.5 \times 10^6\,\mathrm{Pa\,s\,m^{-1}}$, the transmission coefficient is $T = 0.18$.

(b) Show that a coupling medium between the transducer and tissue will maximize the overall transmission if $Z_{\mathrm{coupling}} = \sqrt{Z_{\mathrm{transducer}} Z_{\mathrm{tissue}}}$. Show that in that case the transmission is $T = 0.36$. Ignore interference effects ($\lambda \gg$ the thickness of the coupling medium).

Section 13.4

Problem 17. If the intensity of a sound wave falls to half its original value, what is the change in dB?

Section 13.5

Problem 18. A sound wave with intensity of $10^{-12}\,\mathrm{W\,m^{-2}}$ is the threshold for hearing. Convert that to a pressure amplitude P. Convert the pressure amplitude to a displacement amplitude using Eq. 13.22, with $f = 1\,\mathrm{kHz}$, $\kappa_{\mathrm{air}} = 10^{-5}\,\mathrm{Pa^{-1}}$, and $c_{\mathrm{air}} = 344\ \mathrm{m\ s^{-1}}$. Compare your result with the size of an atom, which is on the order of 0.1 nm. Surprised?

Problem 19. The ear can just hear sound at about 1000 Hz at a level that corresponds to a pressure change of 2×10^{-5} Pa. Atmospheric pressure is 10^5 Pa. Since atmospheric pressure is due to collisions of molecules with the eardrum, there are pressure fluctuations because of fluctuations in the number of collisions in time Δt. We can expect that $\Delta p/p$ is about 1/(number of collisions)$^{1/2}$. Suppose that the eardrum has area S and that when detecting a signal at 1000 Hz it averages over a time interval of 0.5 ms. The number of collisions per unit area per unit time is given by $nv/4$, where n is the number of air molecules per unit volume and v is an average velocity of $482\,\mathrm{m\,s^{-1}}$. The radius of the eardrum is 4.5 mm. Find $\Delta p/p$.

Problem 20. People use many cues to estimate the direction a sound came from. One is the time delay between sound arriving at the left and right ears. Estimate the maximum time delay. Ignore any diffraction effects caused by the head.

Section 13.6

Problem 21. Find the conversion between α in $\mathrm{dB\ m^{-1}}$ and $\mathrm{m^{-1}}$ (as in $I = I_0 e^{-\alpha x}$).

Section 13.7

Problem 22. An ultrasound pulse used in medical imaging has a frequency of 5 MHz and a pulse width of 0.5 μs. Approximately how many oscillations of the sound wave occur in the pulse? The number of oscillations is sometimes called the quality, Q, of the pulse. A pulse with little damping has $Q \gg 1$, whereas a heavily damped pulse has $Q \approx 1$. Is the ultrasound pulse heavily damped?

Problem 23. A heavily damped pulse does not represent a single frequency. Consider a pulse $p(t)$ having the shape

$$p(t) = e^{-(t/\tau)^2} \cos \omega_0 t.$$

Using the techniques developed in Sect. 11.9, calculate the Fourier transform of this pulse. Determine the shape of the power spectrum. How is the parameter τ related to the width of the power spectrum? What is the central frequency of the power spectrum?

Problem 24. Suppose you send a short ultrasound pulse into the body at $t = 0$, and observe echoes at $t = 31$, 79, and 95 μs. How far from the source are the three tissue boundaries? Assume $c = 1540\ \mathrm{m\ s^{-1}}$ in each tissue, and ignore attenuation. Draw a line corresponding to the x-axis ($x = 0$ is the source location), and draw a dot at the position corresponding to each boundary. You have just created an A scan, where each dot represents a boundary.

Problem 25. Suppose you emit an ultrasound pulse in the x direction from a source at each of eight different positions y. Each pulse receives a series of echoes, as shown in the table below (Echo times are in μs.):

y (mm):	0	10	20	30	40	50	60	70
Echo 1	35	37	39	40	45	47	48	49
Echo 2	97	98	58	56	57	96	91	90
Echo 3			71	73	71			
Echo 4			99	99	98			

Draw an x–y coordinate system ($x = 0$ is location of the source) and put a bright spot corresponding to each echo. Assume $c = 1540$ m s^{-1} in each tissue, and ignore attenuation. You have just created a two-dimensional ultrasound image.

Problem 26. Assume the attenuation is proportional to frequency, and is given by 100 dB m^{-1} MHz^{-1}. If you use a 5-MHz ultrasound wave to image a surface 30 mm below the surface of the skin, the measured echo is what fraction of the original intensity? Ignore impedance differences at the surface of the skin and assume that 100 % of the wave is reflected by the surface, so that the reduction of the echo intensity is caused entirely by attenuation. Remember that you must consider the round-trip distance traveled by the wave. Express you answer in dB.

Problem 27. The intensity of echoes depends on not only the nature of the boundary they reflect from, but also the distance to the boundary. Consider a boundary that reflects 50 % of the incident wave intensity. Compare the intensity of the echoes recorded by the detector for boundaries 10, 20, and 30 mm from the source. Assume an attenuation coefficient of 500 dB m^{-1}. Ignore any inverse-square fall-off. Clinical ultrasound imaging devices often use a technique called *time gain compensation* to selectively amplify later echoes, thereby correcting for the effect of attenuation that you just calculated.

Problem 28. The depth resolution of an ultrasound image depends on the speed of sound and the duration of the ultrasound pulse. A pulse having a duration of 0.5 μs has what spatial width (assume $c = 1540$ m s^{-1})? Structures smaller than the spatial pulse width are difficult to resolve using ultrasound imaging.

Problem 29. Ultrasound images are often generated using a series of ultrasound pulses, with echoes detected from each pulse. Images are obtained more quickly if the time between pulses is short. However, if this time is too short, echoes from consecutive pulses overlap, making the ultrasound signal difficult to interpret. Assume the deepest structure you wish to image is 80 mm from the source, and the speed of sound is 1540 m s^{-1}. What is the minimum time between pulses you can use without overlapping echoes? How many pulses per second does this correspond to? If you need to use 256 pulses in order to build up a two-dimensional image, how many images can you generate per second? Can you generate images at the video rate (30 frames per second)?

Problem 30. Suppose that an ultrasound wave is traveling to the right in muscle, toward a 3-mm thick layer of fat. (Use the data in Problem 10 for the acoustic properties of these tissues.) Part of the wave reflects off the left surface of the fat (echo 1), but part is transmitted and then reflects off the right surface, producing a wave traveling to the left. Part of this is detected as echo 2, but part of this left-traveling wave undergoes two additional reflections, traveling back and forth through the fat before being detected as echo 3. Echo 3 is called a *reverberation echo* and is one source of artifact in an ultrasound image. You can have more than one, since the wave can reflect back and forth between the left and right surfaces multiple times. Calculate the time between the first three echoes, and the relative intensities of each one (ignore attenuation).

Problem 31. Assume a fat-muscle boundary is 50 mm below the tissue surface. Calculate the intensity of the reflected wave, ignoring attenuation, using the data in Problem 10. Now, assume there is a bone that lies in the region from 20 to 30 mm below the surface, with the fat-muscle boundary still 50 mm below the surface. Calculate the intensity of the wave reflected from the fat-muscle boundary, accounting for the front and back bone surfaces, ignoring attenuation. If the minimum measurable intensity is −25 dB, will the fat-muscle boundary be observable in each case? In general, surfaces behind a bone do not appear in ultrasound images. The bone casts an *acoustic shadow*.

Problem 32. Verify Eq. 13.39. Show that for $z \gg a$, the intensity falls off as z^{-2}.

Section 13.7.3

Problem 33. Show that when a source of sound waves is moving away from the receiver, the frequency of the source, f, and the frequency measured by the receiver, f', are related by $f' = f/(1 + v_s/c)$.

Problem 34. Suppose a stationary source sends ultrasound waves to the right. They are reflected from an object moving to the right with speed v_o, and then are recorded by the stationary receiver (the receiver and source are at the same location). Derive the relationship in Eq. 13.42 between the frequency of the source, f, and the frequency recorded by the receiver, f', using the following steps.

(a) Find the time t_1 when the receiver records a signal that was emitted by the source at $t = 0$, traveled a distance L, was reflected, and then returned to the receiver.

(b) Find the time t_2 when the receiver records a signal that was emitted by the source at $t = T$, traveled a distance $L + \Delta L$, was reflected, and then returned to the receiver.

(c) Relate the distance ΔL to the speed of the object.
(d) Solve for $T' = t_2 - t_1$.
(e) Determine f' in terms of f.

Problem 35. Show that if $v_o \ll c$, Eq. 13.42 reduces to $f' = f(1 - 2v_o/c)$.

Problem 36. Solve Eq. 13.42 for v_o as a function of f'/f. This allows you to measure the emitted and received frequencies and determine the speed of the object.

References

Berg WA, Cosgrove DO, Doré CJ et al. (2012) Shear-wave elastography improves the specificity of breast US. Radiology 262:435–449

Cameron JR, Skofronick JG, Grant RM (1999) Physics of the body. Medical Physics Publishing, Madison.

Carson PL, Fenster A (2008) Anniversary paper: evolution of ultrasound physics and the role of medical physicists and the AAPM and its journal in that evolution. Med Phys 36:411–428

Clement GT, Hynynen K (2002) A non-invasive method for focusing ultrasound through the human skull. Phys Med Biol 47:1219–1236

Denny MW (1993) Air and water. Princeton University Press, Princeton

Ellis S, Reike V, Kohl M, Westphalen AC (2013) Clinical applications for magnetic resonance guided high intensity focused ultrasound (MRgHIFU): present and future. J Med Imaging Radiat Oncol 57:391–399

Faez T, Emmer M, Kooiman K, Versluis M, van der Steen AFW, de Jong N (2013) 20 years of ultrasound contrast agent modeling. IEEE Trans Ultrason Ferroelectr Freq Control 60:7–20

Glide-Hurst CK, Maidment ADA, Orton CG (2010) Point/counterpoint: ultrasonography is soon likely to become a viable alternative to x-ray mammography for breast cancer screening. Med Phys 37:4526–4529

Hartmann WM (2013) Principles of musical acoustics. Springer, New York

Hendee WR, Ritenour ER (2002) Medical imaging physics, 4th ed. Wiley-Liss, New York

ICRU (1998) Tissue substitutes, phantoms and computational modeling in medical ultrasound. International Commission on Radiation Units and Measurements Report 61. ICRU, Bethesda, MD

Jeanmonod D, Werner B, Morel A, Michels L, Zadicario E, Schiff G, Martin E (2012) Transcranial magnetic resonance imaging-guided focused ultrasound: noninvasive central lateral thalamotomy for chronic neuropathic pain. Neurosurg Focus 32(1):E1–E6

Kremkau FW (2006) Diagnostic ultrasound: principles and instruments. Elsevier Saunders, St. Louis

Lindsay RB, Beyer RT (1989) Acoustics. Ch. 2. In: Anderson HL (ed in chief) A physicist's desk reference. American Institute of Physics, New York

Monteith S, Sheehan J, Medel R, Wintermark M, Eames M, Snell J, Kassell NF, Elias WJ (2013) Potential intracranial applications of magnetic resonance-guided focused ultrasound surgery. J Neurosurg 118:215–221

Merckel LG, Bartels LW, Köhler MO, van den Bongard HJ, Deckers R, Mali WP, Binkert CA, Moonen CT, Gilhuijs KG, van den Bosch MA (2013) MR-guided high-intensity focused ultrasound ablation of breast cancer with a dedicated breast platform. Cardiovasc Intervent Radiol 36(2):292–301. doi:10.1007/s00270-012-0526-6

Morse PM, Ingard KU (1968) Theoretical acoustics. McGraw-Hill, New York

Newman EB (1957) Speech and hearing. In: Gray DE (coordinating ed) American institute of physics handbook. McGraw-Hill, New York, p 3–123

Sacks O (2007) Musicophilia: tales of music and the brain. Knopf, New York

Saedi N, Kaminer M (2013) New waves for fat reduction: high-intensity focused ultrasound. Semin Cutan Med Surg 32(1):26–30.

Wells PNT (2006) Ultrasound imaging. Phys Med Biol 51:R83–R98

Atoms and Light

14

This chapter describes some of the biologically important properties of infrared, visible, and ultraviolet light. X rays are discussed in Chaps. 15 and 16. A brief discussion of geometrical optics accompanies the description of image formation in the eye and errors of refraction.

Section 14.1 considers the particle properties of light (photons), while Sect. 14.2 looks at the wave properties of electrons. Photons can be emitted or absorbed when single atoms change energy levels, and they have certain frequencies characteristic of the atom, as described in Sect. 14.3. Molecules have additional energy levels shown in Sect. 14.4. Biological examples include spectrophotometry, photodissociation, immunofluorescence, infrared spectroscopy, and Raman scattering. There is an extensive literature about these; the discussion here is quite brief.

Section 14.5 describes the scattering and absorption of radiation, processes that are important in the rest of this chapter and in Chaps. 15–17. The probability of scattering or absorption is measured by the cross section, which is also introduced here. Photons may scatter many times in a substance without being absorbed. This leads to the concept of turbid media such as milk or clouds. In some cases the process can be modeled accurately with the diffusion approximation developed in Sect. 14.6. Biological examples of infrared scattering (including Raman scattering) are described in Sect. 14.7

Photons can be absorbed and emitted by some substances in a continuous range of frequencies or wavelengths. This happens when many atoms interact with each other and blur the energy levels, as in liquids and solids. This leads to the concept of thermal radiation described in Sect. 14.8. Examples of thermal radiation are infrared radiation by the skin and ultraviolet radiation by the sun. The former is discussed in Sect. 14.9.

Blue and ultraviolet light are used for therapy, as described in Sect. 14.10. They can also be harmful, particularly to skin and eyes.

Lasers are used to heat tissue, often rapidly enough to do surgery as water in the tissue suddenly boils. Models of this process include the bioheat equation that is developed in Sect. 14.11.

Section 14.12 describes the problem of radiometry: measuring radiation. All of the important quantities are defined, and the corresponding photometric and actinometric quantities are also introduced.

Section 14.13 describes how the eye focuses an image on the retina and the correction of simple errors of refraction. A final example of the photon nature of light is given in Sect. 14.14: the statistical limit to dark-adapted vision—shot noise—which is important when the eye is operating in its most sensitive mode.

We can only provide a brief introduction to the role optics and light play in biology. For more details with many fascinating examples, see Johnsen (2011).

14.1 The Nature of Light: Waves and Photons

Light travels in a vacuum with a velocity $c = 3 \times 10^8\,\mathrm{m\,s^{-1}}$ (to an accuracy of 0.07 %). When light travels through matter, its speed is less than this and is given by

$$c_n = \frac{c}{n}, \tag{14.1}$$

where n is the *index of refraction* of the substance. The value of the index of refraction depends on both the composition of the substance and the color of the light.

A controversy over the nature of light existed for centuries. Sir Isaac Newton explained many properties of light with a particle model in the seventeenth century. In the early nineteenth century, Thomas Young performed some interference experiments that could be explained only by assuming that light is a wave. By the end of the nineteenth century, nearly all known properties of light, including many of

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_14, © Springer International Publishing Switzerland 2015

its interactions with matter, could be explained by assuming that light consists of an electromagnetic wave. By an electromagnetic wave, we mean that

1. Light can be produced by accelerating an electric charge.
2. Light has an electric and a magnetic field associated with it; the force that the light exerts on a charged particle is given by Eq. 8.2, $\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$. The force due to the magnetic field is usually very small.
3. The velocity of light traveling in a vacuum is given by electromagnetic theory as $c = 1/\sqrt{\epsilon_0 \mu_0}$, where parameters ϵ_0 and μ_0 are measured in the laboratory for "ordinary" electric and magnetic fields.

In the early twentieth century, light was discovered to have *both* particle properties and electromagnetic wave properties at the same time. This rather disconcerting discovery was followed in a few years by the discovery that matter, which had been thought to consist of particles, also has wave properties.

A traveling wave of light can be described by a function of the form $f(x - c_n t)$, which represents a disturbance traveling along the x axis in the positive direction. If the wave is sinusoidal, then the period T, frequency ν,[1] and wavelength λ are related by

$$\nu = \frac{1}{T}, \qquad c_n = \lambda \nu. \tag{14.2}$$

As light moves from one medium into another where it travels with a different speed, the frequency remains the same. The wavelength changes as the speed changes.

Each particle of light or *photon* has energy E. The energy of each photon (a "particle" concept) is related to its frequency (a "wave" concept) by

$$E = h\nu = \frac{hc_n}{\lambda}. \tag{14.3}$$

The proportionality constant h is called *Planck's constant*. It has the numerical value[2]

$$h = 6.63 \times 10^{-34}\,\mathrm{J\,s} = 4.14 \times 10^{-15}\,\mathrm{eV\,s}. \tag{14.4}$$

We use the number "h stroke" or "h bar":

$$\hbar = \frac{h}{2\pi} = 1.05 \times 10^{-34}\mathrm{J\,s} = 0.66 \times 10^{-15}\,\mathrm{eV\,s}. \tag{14.5}$$

In terms of the angular frequency $\omega = 2\pi\nu$,

$$E = \hbar\omega. \tag{14.6}$$

[1] We used f for frequency in earlier chapters because this is customary when discussing noise. Here we adopt ν for frequency, the notation most often used in atomic physics.

[2] The electron volt (eV) is a unit of energy. $1\mathrm{eV} = 1.6 \times 10^{-19}$ J. It is the energy acquired by an electron that moves through a potential difference of 1 V.

Table 14.1 The regions of the electromagnetic spectrum and their boundaries

Name	Wavelength	Frequency (Hz)	Energy (eV)
Radio waves			
	1 m	300×10^6	1.24×10^{-6}
Microwaves			
	1 mm	300×10^9	1.24×10^{-3}
Extreme infrared			
	15 μm	20×10^{12}	0.083
Far infrared			
	6 μm	50×10^{12}	0.207
Middle infrared			
	3 μm	100×10^{12}	0.414
Near infrared			
	750 nm	400×10^{12}	1.65
Visible			
	400 nm	750×10^{12}	3.11
Ultraviolet			
	12 nm	24×10^{15}	100
X rays, γ rays			

Table 14.2 The visible electromagnetic spectrum

Color	Wavelength (nm)	Frequency (THz)	Energy (eV)
	750	400	1.65
Red			
	610	490	2.03
Orange			
	590	510	2.10
Yellow			
	570	530	2.17
Green			
	500	600	2.48
Blue			
	450	670	2.76
Violet			
	400	750	3.11

The electromagnetic spectrum includes radio waves, microwaves, infrared, visible, and ultraviolet light, x rays, and γ (gamma) rays. Table 14.1 shows the wavelengths that separate these arbitrary regions, together with the frequencies and the energies of the photons. Visible-light photons have an energy of a few electron volts. X rays are 10^4–10^7 times more energetic, while γ rays, which come from atomic nuclei, are often even more energetic but may have energies overlapping x-ray energies. The only difference between x rays and γ rays is their source.

The property of light that we associate with color is the frequency or the energy of each photon. Visible light covers a narrow range of frequencies, about an octave (a factor of 2). Table 14.2 shows the wavelengths and frequencies dividing the colors of the visible spectrum. The frequencies are in the 400–750 THz range.

Most of the effects discussed in this chapter, particularly those dealing with emission and absorption, can be explained by assuming that light is made up of photons. Phenomena

such as interference, diffraction, and polarization require the wave theory.

14.2 Electron Waves and Particles: The Electron Microscope

Like light, matter can have both wave and particle properties. The French physicist, Louis de Broglie, derived a quantum mechanical relationship between a particle's momentum p and wavelength:

$$\lambda = \frac{h}{p} \tag{14.7}$$

(Eisberg and Resnick 1985). For example, a 100-eV electron has a speed of $5.9 \times 10^6\ \mathrm{m\,s^{-1}}$ (about 2 % the speed of light), a momentum of $5.4 \times 10^{-24}\ \mathrm{kg\,m\,s^{-1}}$, and a wavelength of 12 nm.

The electron microscope takes advantage of the short wavelength of electrons to produce exquisite pictures of very small objects. Diffraction limits the spatial resolution of an image to about a wavelength. For a visible light microscope, this resolution is on the order of 500 nm (Table 14.2). For the electron microscope, however, the wavelength of the electron limits the resolution. A typical electron energy used for imaging is about 100 keV, implying a wavelength much smaller than an atom (however, practical limitations often limit the resolution to about 1 nm). Table 1.2 shows that viruses appear as blurry smears in a light microscope, but can be resolved with considerable detail in an electron microscope. In 1986, Ernst Ruska shared the Nobel Prize in Physics "for his fundamental work in electron optics, and for the design of the first electron microscope."

Electron microscopes come in two types. In a transmission electron microscope (TEM), electrons pass through a thin sample. In a scanning electron microscope (SEM), a fine beam of electrons is raster scanned across the sample, and secondary electrons emitted by the surface are imaged. In both cases, the image is formed in vacuum and the electron beam is focused using a magnetic lens.

14.3 Atomic Energy Levels and Atomic Spectra

The simplest system that can emit or absorb light is an isolated atom. An atom is isolated if it is in a monatomic gas. In addition to translational kinetic energy, isolated atoms have specific discrete internal energies, called *energy levels*. An atom can change from one energy level to another by emitting or absorbing a photon with an energy equal to the energy

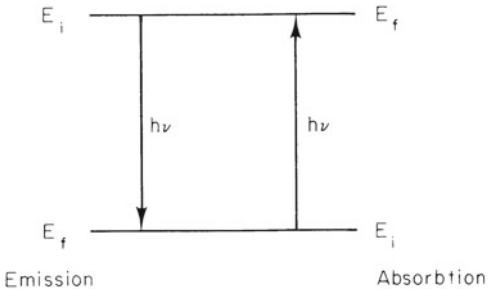


Fig. 14.1 A system can change from one energy to another by emitting or absorbing a photon. The photon has an energy equal to the difference in energies of the two levels

difference between the levels. Let the energy levels be labeled by $i = 1, 2, 3, \ldots$, with the energy of the ith state being E_i. There is a lowest possible internal energy for the atom; when the atom is in this state, no further energy loss can take place. If E_i is greater than the lowest energy, then the atom can lose energy by emitting a photon of energy $E_i - E_f$ and exist in a lower energy state E_f (Fig. 14.1).

It is possible, using techniques of quantum mechanics, to calculate the energies of the levels with reasonable accuracy (and in some cases with spectacular accuracy). For our purposes, we need to only recognize that energy levels exist and know their approximate values. You may be familiar with the model of the hydrogen atom developed by Niels Bohr, in which the energy of the nth level is given by

$$E_n = -\left(\frac{1}{4\pi\epsilon_0}\right)^2 \frac{m_e e^4}{2\hbar^2 n^2}, \quad n = 1, 2, 3, \ldots. \tag{14.8}$$

The energy is in joules when the electron mass m_e is in kilograms, the electronic charge e is in coulombs, and $\hbar$ is in J s. The Coulomb's law constant $1/4\pi\epsilon_0$ is given in Eq. 6.2. Dividing the energy in joules by e gives the energy in electron volts:

$$E_n = -\frac{13.6}{n^2} \quad \text{(in eV)}. \tag{14.9}$$

The energy-level diagram in Fig. 14.2 shows these energies and some transitions between them. In other cases, the energy depends not only on the integer $n = 1, 2, 3, 4, \ldots$, but on additional quantum numbers as well.

Figure 14.3 plots the spectrum for hydrogen versus wavelength, along with some of the energy levels of hydrogen. Letters $a, b, c, \ldots$ mark lines in the spectrum and the associated transitions.

In general, the internal energy of an atom depends on the values of five quantum numbers for each electron in the atom. The quantum numbers are

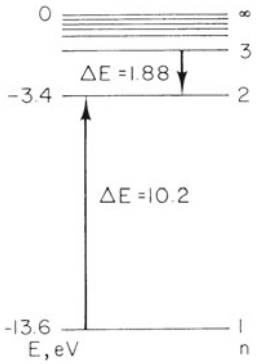


Fig. 14.2 Energy levels in a hydrogen atom. Transitions are shown corresponding to the emission and absorption of light

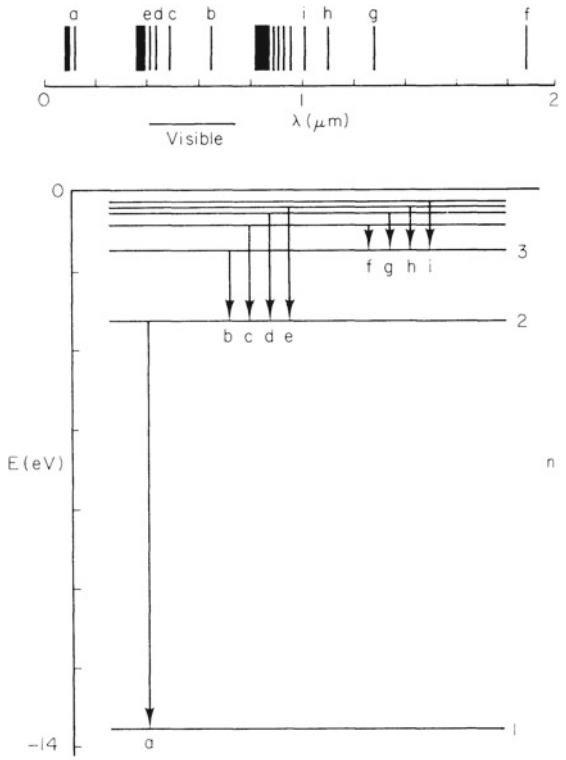


Fig. 14.3 The spectrum for hydrogen plotted versus wavelength and the energy levels for hydrogen. Some spectral lines and the corresponding transitions have been labeled

$n = 1, 2, 3, \ldots$	the principal quantum number
$l = 0, 1, 2, \ldots, n-1$	the orbital angular momentum quantum number
$s = \frac{1}{2}$	the spin quantum number
$m_l = -l, -(l-1), \ldots, l-1, l$	"z component" of the orbital angular momentum
$m_s = -\frac{1}{2}, \frac{1}{2}$	"z component" of the spin.

Sometimes the last two quantum numbers, m_l and m_s, are replaced by two other quantum numbers, j and m_j. The allowed values of j and m_j are

$j = l - \frac{1}{2}$ or $l + \frac{1}{2}$ except that $j = \frac{1}{2}$ when $l = 0$	total angular momentum quantum number
$m_j = -j, -(j-1), \ldots, j-1, j$	"z component" of total angular momentum

Whether one uses m_l and m_s or j and m_j, each electron is described by five quantum numbers, one of which is always $\frac{1}{2}$. There are four quantum numbers that can change, corresponding to the three space degrees of freedom and the spin associated with m_s. The internal energy of the atom is the sum of the kinetic and potential energies of each electron. The energy of each electron depends on the values of its quantum numbers. It is influenced by the electric field generated by the nucleus and all the other electrons. There are also magnetic interactions between electrons and between each electron and the nucleus, because the moving charges generate magnetic fields.

No two electrons in an atom can have the same values for all their quantum numbers, a fact known as the *Pauli exclusion principle.*

The *ionization energy* is the smallest amount of energy required to remove an electron from the atom when the atom is in its ground state. For hydrogen the ionization energy is 13.6 eV. In contrast, it takes only 5.1 eV to remove the least tightly bound electron from a sodium atom.

An atom can receive energy from an external source, such as a collision with another atom or some other particle. It can also absorb a photon of the proper energy. Absorbing just the right amount of energy allows one of its electrons to move to a higher energy level, as long as that level is not already occupied. The atom can then get rid of this excess energy by radiating a photon, with the excited electron falling to an unoccupied state with lower energy. This change is usually consistent with the following *selection rules*, which can be derived using quantum mechanics:

$$\Delta l = 1, \quad \Delta j = 0, \pm 1. \tag{14.10}$$

14.4 Molecular Energy Levels

In addition to internal energy, an atom can have kinetic energy of translation with three degrees of freedom. The translational kinetic energy is also quantized, but as long as the atom is not confined to a very small volume, the levels are so closely spaced that the translational kinetic energy can be regarded as continuous.

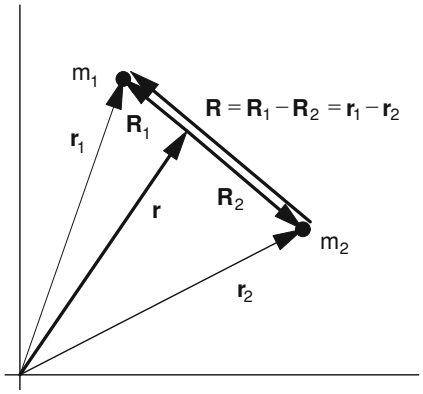


Fig. 14.4 A diatomic molecule. Vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ are the positions of the atoms measured in the laboratory. Vectors $\mathbf{R}_1$ and $\mathbf{R}_2$ are coordinates in the center-of-mass system. Vector $\mathbf{r}$ is the position of the center of mass

Two atoms together have six degrees of translational freedom, because each can move in three-dimensional space. However, if the atoms are bound together, their motions are not independent One can speak of the three degrees of freedom for translation of the molecule as a whole (center-of-mass motion) and also the vector displacement of one atom from the other. This is shown in Fig. 14.4. Vector $\mathbf{r}$ locates the center of mass of the two atoms. It is located at a point such that $m_1\mathbf{R}_1 = -m_2\mathbf{R}_2$.

Consider two particles of mass m_1 and m_2. Their positions with respect to some fixed origin are $\mathbf{r}_1$ and $\mathbf{r}_2$. The velocity of each particle is $\mathbf{v}_i = d\mathbf{r}_i/dt$. The kinetic energy of the ith particle is $T_i = m_i(\mathbf{v}_i \cdot \mathbf{v}_i)/2$. Define the center of mass by

$$\mathbf{r} = \frac{m_1\mathbf{r}_1 + m_2\mathbf{r}_2}{m_1 + m_2}$$

and the vectors from the center of mass to each particle by

$$\mathbf{R}_1 = \mathbf{r}_1 - \mathbf{r} = \frac{m_2(\mathbf{r}_1 - \mathbf{r}_2)}{m_1 + m_2} = \frac{m_2\mathbf{R}}{m_1 + m_2},$$
$$\mathbf{R}_2 = \frac{-m_1\mathbf{R}}{m_1 + m_2}.$$

The total kinetic energy is $T = m_1(\mathbf{v}_1 \cdot \mathbf{v}_1)/2 + m_2(\mathbf{v}_2 \cdot \mathbf{v}_2)/2$. Since $\mathbf{v}_i = \mathbf{v} + \mathbf{V}_i$, we have

$$2T = (m_1 + m_2)(\mathbf{v} \cdot \mathbf{v}) + m_1(\mathbf{V}_1 \cdot \mathbf{V}_1) + m_2(\mathbf{V}_2 \cdot \mathbf{V}_2) + 2\mathbf{v} \cdot (m_1\mathbf{V}_1 + m_2\mathbf{V}_2).$$

The last term vanishes because $m_1\mathbf{R}_1 + m_2\mathbf{R}_2 = 0$. Consider the second term. Differentiating $\mathbf{R}_1 = m_2\mathbf{R}/(m_1 + m_2)$ shows that

$$\mathbf{V}_1 \cdot \mathbf{V}_1 = \left(\frac{m_2}{m_1 + m_2}\right)^2 V^2,$$
$$\mathbf{V}_2 \cdot \mathbf{V}_2 = \left(\frac{m_1}{m_1 + m_2}\right)^2 V^2.$$

Therefore,

$$T = \frac{(m_1 + m_2)v^2}{2} + \frac{1}{2}\frac{m_1 m_2}{m_1 + m_2}V^2.$$

The first term is the kinetic energy of a point mass $m_1 + m_2$ traveling at the speed of the center of mass. The second is the kinetic energy of a particle having the *reduced mass* $m_1m_2/(m_1 + m_2)$ and the speed of relative motion of the two particles, $V = |\mathbf{V}| = |d\mathbf{R}/dt|$. If $\mathbf{R}$ changes magnitude, the particles are *vibrating*. If $\mathbf{R}$ has a fixed magnitude, the molecule can *rotate*. If the molecule is rotating in some plane with angular velocity ω, then

$$\frac{1}{2}\frac{m_1 m_2}{m_1 + m_2}V^2 = \frac{1}{2}\frac{m_1 m_2}{m_1 + m_2}R^2\omega^2 = \frac{1}{2}I\omega^2.$$

The quantity $I = \left[m_1m_2/(m_1 + m_2)\right]R^2 = m_1R_1^2 + m_2R_2^2$ is the *moment of inertia* of the two objects (Serway and Jewett 2013, p. 312 and 328). In this case the angular momentum about the center of mass is

$$L = R_1(m_1v_1) + R_2(m_2v_2) = m_1R_1^2\omega + m_2R_2^2\omega = I\omega.$$

These two equations can be combined to give the rotational kinetic energy in terms of the angular momentum about the center of mass:

$$T = \frac{L^2}{2I}.$$

Quantum mechanically, the angular momentum cannot take on any arbitrary value. The square of the angular momentum is restricted to the values

$$L^2 = r(r + 1)\hbar^2, \qquad r = 0, 1, 2, \ldots.$$

Since there is no potential energy, the total energy of rotation of the molecule is

$$E_r = \frac{r(r + 1)\hbar^2}{2I}, \qquad r = 0, 1, 2, \ldots. \tag{14.11}$$

The spacing of the rotational levels is shown in Fig. 14.5. A detailed calculation using quantum mechanics shows that when a photon is emitted or absorbed, r must change by ± 1. Therefore the photon energy is

$$\Delta E_r = E_r - E_{r-1} = \frac{\hbar^2}{I}r, \qquad r = 1, 2, \ldots. \tag{14.12}$$

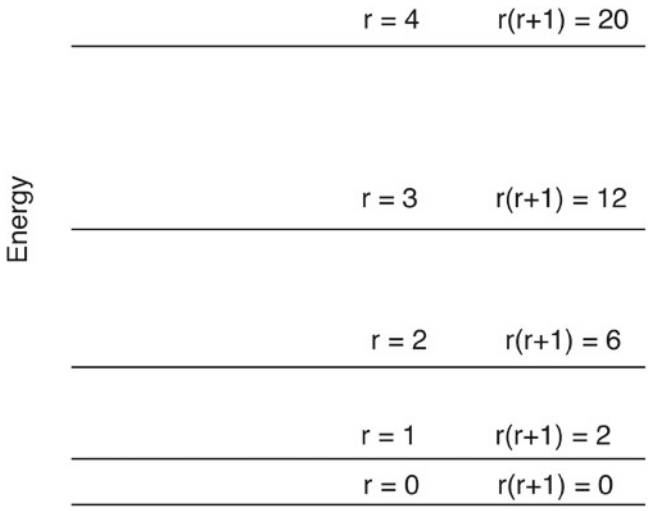


Fig. 14.5 Energy levels of a rotating molecule

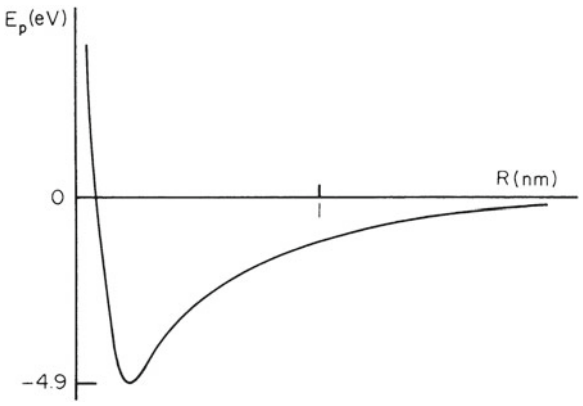


Fig. 14.6 The potential energy of a sodium ion and a chloride ion as a function of their nuclear separation

The problems at the end of the chapter show that these photons have low energies, so that rotational spectra lie in the far-infrared region (far meaning far from the visible region, i.e., very long wavelengths).

The other possibility is that the atoms in the molecule vibrate back and forth along the line joining their centers. If two masses have an equilibrium position a certain distance apart, work must done either to push them closer together or to pull them farther apart. In either case, the potential energy is increased. At the equilibrium separation the potential energy is a minimum. Figure 14.6 shows the potential energy E_p of a sodium ion and a chloride ion as a function of their separation. The potential has a minimum at a separation R_0 of about 0.2 nm. The simplest function that has a minimum is a parabola. A parabola can be used to approximate the minimum in Fig. 14.6: $E_p(R) = \frac{1}{2}k(R - R_0)^2$. Since (see Sect. 6.4) $dE_p = -F dr$, the force is $F = -dE_p/dR = -k(R - R_0)$, which is the linear approximation to the force between the two ions. The force is attractive if $R > R_0$ and repulsive if $R < R_0$.

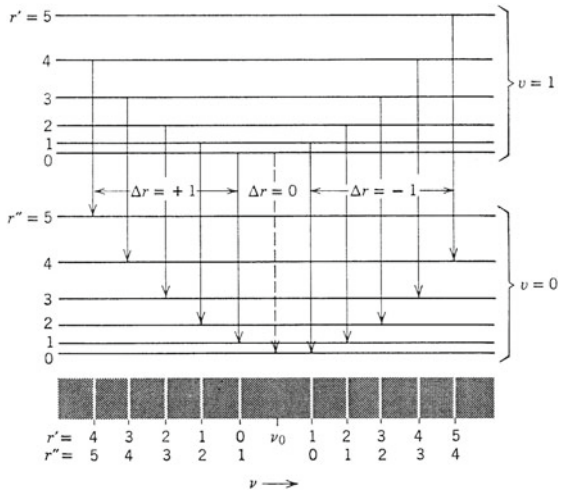


Fig. 14.7 Transitions for vibrational–rotational spectra. (Source: Eisberg and Resnick 1985. Copyright ©1985 John Wiley & Sons. Reproduced by permission of John Wiley & Sons, Inc.)

A mass subject to a linear restoring force is called a *harmonic oscillator* (Appendix F). A mass m subject to a linear restoring force $-kx$ oscillates with an angular frequency $\omega^2 = k/m$. Classically, the energy of the oscillating mass depends on the amplitude of the motion and can have any value. Quantum mechanically, it is restricted to values

$$E_v = \hbar\omega\left(v + \tfrac{1}{2}\right), \qquad v = 0, 1, 2, \ldots. \tag{14.13}$$

This is the total energy, including both kinetic and potential energy. The levels are spaced equally by an amount $\hbar\omega$. The spacing is usually greater than that for rotational levels, often in the infrared. The transitions that give rise to the emission or absorption of photons require a change in the rotational quantum numbers as well as the vibrational ones. The selection rules are

$$\Delta r = \pm 1, \qquad \Delta v = \pm 1. \tag{14.14}$$

Some of these vibrational–rotational transitions are shown in Fig. 14.7.

Finally, there can be transitions involving v, r, and the electronic quantum numbers as well. When the electronic quantum numbers change, the shape of the interatomic potential changes, as shown in Fig. 14.8. The details of molecular spectra are fairly involved and are summarized in many texts. Transitions of biological importance are discussed in Grossweiner (1994, pp. 33–38). If the electron selection rules are satisfied, the transition is fairly rapid (typically 10^{-8}s), a process called *fluorescence*. Sometimes the electron becomes

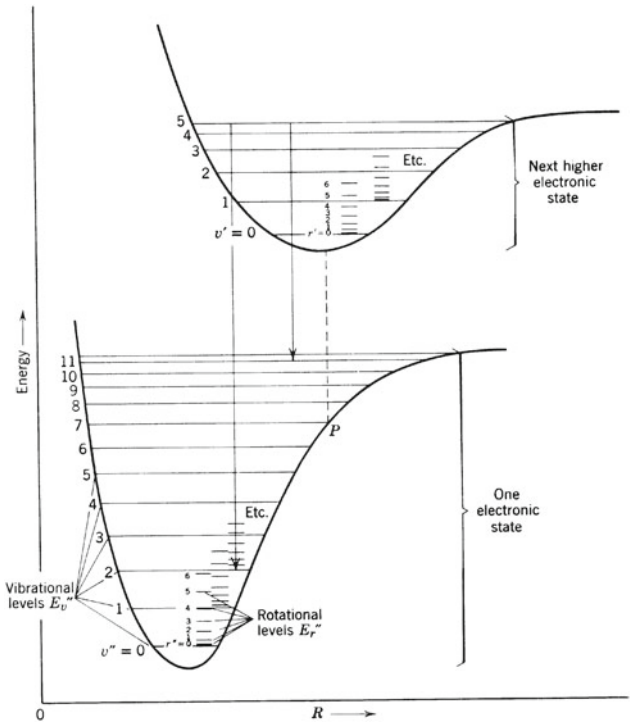


Fig. 14.8 A combination of changes in electronic quantum numbers within an atom and of vibrational and rotational quantum numbers within the molecule. (Source: Eisberg and Resnick 1985. Copyright ©1985 John Wiley & Sons. Reproduced by permission of John Wiley & Sons, Inc)

trapped in a state where it cannot decay according to the electronic selection rules of Eq. 14.10. It may then have a lifetime up to several seconds before decaying, a phenomenon called *phosphorescence*.

14.5 Scattering and Absorption of Radiation; Cross Section

In the absence of interference and diffraction effects, photons in a vacuum travel in a straight line. When they travel through matter they are apparently slowed down, leading to an index of refraction greater than unity; they may also be scattered or absorbed. Visible light does not pass through a building wall, but it does pass through a glass window. The absorption may depend on the frequency or wavelength of the light. The window can be made of colored glass. The light can also be scattered. This leads to the blue of the sky or to the white of clouds. If there is absorption as well as scattering, the clouds may appear gray instead of white. How light is scattered or absorbed in tissue has become very important in biophysics. Infrared light absorption can be used to measure chemical

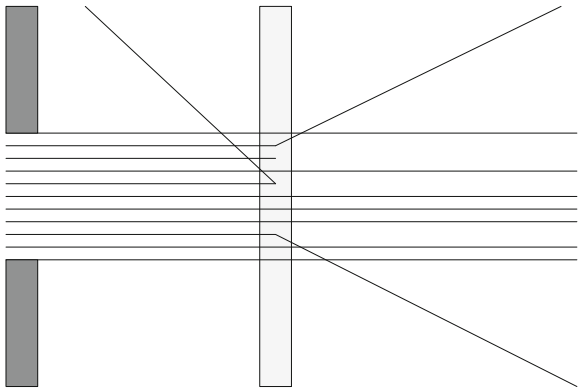


Fig. 14.9 A collimated beam of photons passes from left to right through a thin slice of material. Some photons pass through, some are scattered, and some are absorbed

composition of the body. Light is also used for therapy and for laser surgery.

This section shows how to describe a single interaction of a photon with some substance. The photon can be scattered or absorbed. Section 14.6 develops one technique for calculating what happens when the photon undergoes many scattering events before being absorbed or emerging from the material.

Imagine that we have a distant source of photons that travel in straight lines, and that we collimate the beam (send it through an aperture) so that a nearly parallel beam of photons is available to us. Imagine also that we can see the tracks of the N photons in the beam, as in Fig. 14.9. When a thin sample of material of thickness dz is placed in the beam, a certain number of photons are scattered and a certain number are absorbed. If we repeat the experiment many times, we find that the number of photons scattered fluctuates about an average value that we call dN_s and the number absorbed fluctuates about an average value dN_a. When we vary the thickness of the absorber, we find that if it is sufficiently thin, the average number of photons scattered and absorbed is proportional to the thickness as well as the number of incident photons:

$$dN_s = \mu_s N dz, \qquad dN_a = \mu_a N dz. \tag{14.15}$$

The total number of unscattered photons N changes according to

$$dN = -(dN_s + dN_a) = -N(\mu_s + \mu_a)dz$$

with solution

$$N(z) = N_0 e^{-\mu z} = N_0 e^{-(\mu_s + \mu_a)z}. \tag{14.16}$$

The quantity μ is the *total linear attenuation coefficient*. Quantities μ_s and μ_a are the linear scattering and absorption coefficients. Both depend on the material and the energy

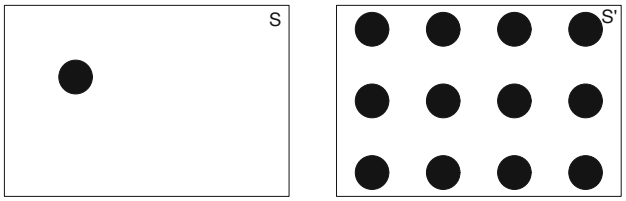


Fig. 14.10 Each circle represents the cross section σ associated with a target entity such as an atom. **a** There is one atom in area S. **b** There are N_T target atoms per unit area in area S'

of the photons. This kind of exponential absorption is known as *Beer's law* or the *Beer–Lambert* law.

The interaction of photons with matter is statistical. The *cross section* σ is an effective area proportional to the probability that an interaction takes place. The interaction takes place with a *target entity*. It is sometimes convenient to define the target to be a single molecule, at other times an atom, and still other times one of the electrons within an atom. We can visualize the meaning of the cross section by considering either a single target entity interacting with a beam of photons or a single photon interacting with a thin foil of targets. Both are shown in Fig. 14.10. For the single target in Fig. 14.10a, consider a beam of N photons passing through the area S with a uniform number per unit area N/S. Let the average number of interactions be $\overline{n}$. The cross section per target entity is defined by saying that the fraction of photons that interact is equal to the fraction of the area occupied by the cross section:

$$\frac{\overline{n}}{N} = \frac{\sigma}{S}. \tag{14.17}$$

We denote the number of photons per unit area by Φ and write Eq. 14.17 as $\overline{n} = \sigma\Phi$. This is the average number of scatterings per target entity or the *probability of interaction per target entity when the beam has Φ photons per unit area*:

$$p = \sigma\Phi. \tag{14.18}$$

Strictly speaking, $\overline{n}$ is dimensionless, σ has the dimensions m^2, and Φ has dimensions m^{-2}. However, it is often helpful to think of $\overline{n}$ as being interactions per target entity and σ as being m^2 per target entity.

Alternatively, imagine sending a beam of photons at the target of area S' shown in Fig. 14.10b. There are N_T target entities per unit area in the path of the beam, each having an associated area σ. The fraction of the photons that interact is again the fraction of the area that is covered:

$$\frac{\overline{n}}{N} = \frac{\sigma S' N_T}{S'} = \sigma N_T. \tag{14.19}$$

This is the *probability that a single photon interacts when there are N_T target entities per unit area*. Note the symmetry with Eq. 14.18. In the first case, there is one target entity and a certain number of photons per unit area. In the second case, there is one photon and a certain number of target entities per unit area.

If a number of mutually exclusive interactions can take place (such as absorption and scattering), we can define a cross section for each kind of interaction. The probabilities and the cross sections add:

$$\sigma_{\mathrm{tot}} = \sum_i \sigma_i. \tag{14.20}$$

The second interpretation we had above can be used to relate the cross section to the attenuation coefficient. The number of target entities per unit area is equal to the number per unit volume times the thickness of the target along the beam. To obtain the number of target atoms per unit volume, recall that 1 mol of atoms contains Avogadro's number N_A atoms. If A is the mass of a target containing 1 mol of atoms and the target has mass density ρ, then volume V has mass ρV and contains $\rho V/A$ mol and $N_A\rho V/A$ atoms. Therefore the number of atoms per unit volume is $N_A\rho/A$, and the number of atoms per unit area is

$$N_T = \frac{N_A\rho}{A}dz. \tag{14.21}$$

The linear coefficients are related to their corresponding cross sections by

$$\begin{aligned} \mu_s &= \frac{N_A\rho}{A}\sigma_s, \\ \mu_a &= \frac{N_A\rho}{A}\sigma_a, \\ \mu &= \frac{N_A\rho}{A}(\sigma_s + \sigma_a) = \frac{N_A\rho}{A}\sigma_{\mathrm{tot}}, \end{aligned} \tag{14.22}$$

where σ_{tot} is the sum of all the interaction cross sections.

Be careful with units! Avogadro's number is 6.022141×10^{23} entities per mole, which is the number in a **gram** atomic weight. For carbon, $A = 12.01 \times 10^{-3}$ kg mol^{-1} and $\rho = 2.0 \times 10^3$ kg m^{-3}. This is discussed further on page 433.

We may wish to know the probability that particles (in this case photons) are scattered in a certain direction. We have to consider the probability that they are scattered into a small solid angle $d\Omega$. In this case, σ is called the *differential scattering cross section* and is often written as

$$\frac{d\sigma}{d\Omega}d\Omega \qquad \text{or} \qquad \sigma(\theta)d\Omega. \tag{14.23}$$

The units of the differential scattering cross section are m^2 sr^{-1}. The differential cross section depends on θ, the angle between the directions of travel of the incident and scattered particles. In a spherical coordinate system in which

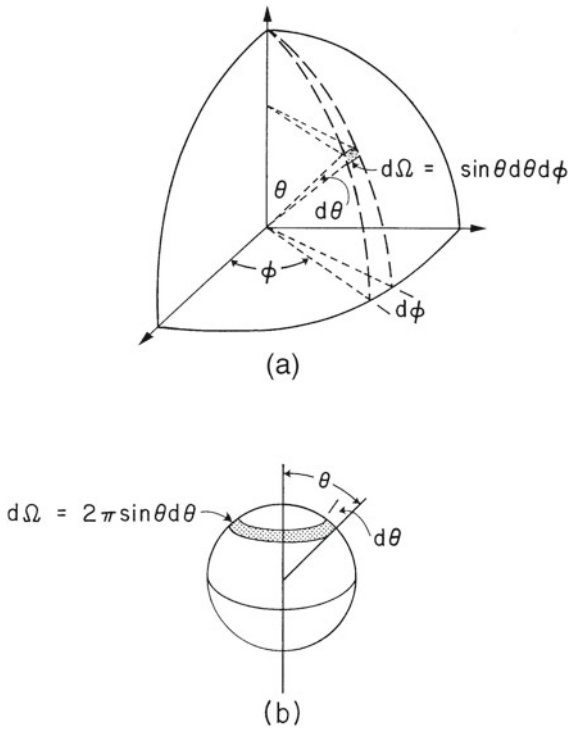


Fig. 14.11 **a** A small solid angle $d\Omega = \sin\theta\, d\theta\, d\phi$ surrounds the direction defined by angles θ and ϕ. **b** The solid angle $d\Omega = 2\pi \sin\theta\, d\theta$ results from integrating over ϕ

the incident particle moves along the z axis, the solid angle is $d\Omega = \sin\theta\, d\theta\, d\phi$ (Appendix L). If the cross section has no ϕ dependence, then the integration over ϕ can be carried out and $d\Omega = 2\pi \sin\theta\, d\theta$. These solid angles are shown in Fig. 14.11.

There are three ways to interpret the exponential decay of the primary beam. First, the number of particles remaining in the beam that have undergone no interaction decreases as the target becomes thicker, so that the number of particles available to interact in the deeper layers is less. Second, the exponential can be regarded as taking into account the fact that in a thicker sample some of the target atoms are hidden behind others and are therefore less effective in causing new interactions. The third interpretation is in terms of the Poisson probability distribution (Appendix J). Each layer of thickness dz provides a separate chance for the beam particles to interact. The probability of interacting in any one layer dz is small, $p = \sigma_{\text{tot}} N_A \rho\, dz/A$, while the total number of "tries" is z/dz. The average number of interactions is $m = p \times$ number of tries. The probability of no interaction is $e^{-m} = \exp(-\sigma_{\text{tot}} N_A \rho\, z/A) = e^{-\mu z}$.

When the cross section for scattering is large, things can become quite complicated. For example, photons may scatter many times and be traveling through the material in all directions. Various approximations have been used to model photon transport in such a case. We will examine some of them shortly. One simple correction that is often made is to consider the average direction a scattered photon travels, for example, the average value of the cosine of the scattering angle, $g = \overline{\cos\theta}$, where θ is the angle of a single scattering. If the average angle of scattering is very small, g is nearly 1. If the photon is scattered backward, $g = -1$, and if the scattering is isotropic, $g = 0$. Formally,

$$g = \frac{\int_0^\pi \sigma(\theta) \cos\theta\, 2\pi \sin\theta\, d\theta}{\int_0^\pi \sigma(\theta)\, 2\pi \sin\theta\, d\theta}. \tag{14.24}$$

The *reduced scattering coefficient*

$$\mu_s' = (1 - g)\mu_s \tag{14.25}$$

is what is usually measured.

The values of the scattering and absorption coefficients vary widely. For infrared light at 780 nm, values are roughly[3]

$$\mu_s' = 1500\,\text{m}^{-1}, \qquad \mu_a = 5\,\text{m}^{-1}.$$

14.6 The Diffusion Approximation to Photon Transport

When photons enter a substance, they may scatter many times before being absorbed or emerging from the substance. This leads to *turbidity*, which we see, for example, in milk or clouds. The most accurate studies of multiple scattering are done with *Monte Carlo* computer simulations, in which probabilistic calculations are used to follow a large number of photons as they repeatedly interact in the tissue being simulated. However, Monte-Carlo techniques use lots of computer time. Various approximate analytic solutions also exist. The field is reviewed in Chap. 5 of Grossweiner (1994).

14.6.1 Diffusion Approximation

One of the approximations, the diffusion approximation, is described here. It is useful when many scattering events occur for each photon absorption. This is a valid approximation for most tissue, but not for cerebrospinal fluid or synovial (joint) fluid.

[3] These are eyeballed from data for various tissues reported in the article by Yodh and Chance (1995). Values are up to ten times larger at other wavelengths. See Table 5.2 in Grossweiner (1994). Nickell et al. (2000) report values for skin that depend on both the direction of propagation and the degree of stretching of the skin. They are similar to the values reported here.

If the photons have undergone enough scattering in a medium, all memory of their original direction is lost. In that case, the movement of the photons can be modeled by the diffusion equation. In Chap. 4, we wrote Fick's second law as

$$\frac{\partial C}{\partial t} = D\nabla^2 C + Q.$$

The left-hand side of the equation is the rate at which the concentration, the number of particles per unit volume, is increasing. The term $D\nabla^2 C$ is the net diffusive flow into the small volume, the particle current being given by $\mathbf{j} = -D\nabla C$. The last term is the rate of production or loss of particles within the volume by other processes, depending on whether Q is positive or negative.

Let us suppose that we can apply this to photons. We will consider two contributions to Q. The concentration must be the number of *diffusing* photons per unit volume. Many in the incident beam are still traveling in the original direction and are not diffusing, but if they are scattered they become part of the diffusing photon pool. Therefore there may be a source term, which we will call s, due to the incident photons. But photons are also being absorbed. They are traveling with a speed $c_n = c/n$, where n is the index of refraction of the medium. In time dt, they travel a distance $dx = c_n dt$, and the probability that they are absorbed is $\mu_a dx = \mu_a c_n dt$. Therefore the diffusion equation for photons is

$$\frac{\partial C}{\partial t} = D\nabla^2 C - \mu_a c_n C + s. \tag{14.26}$$

Each term has the units of photons $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$.

In photon transfer, it is customary to make two changes in this equation. The first is to divide all terms by the speed of the photons in the medium,[4] c_n. The result is

$$\frac{1}{c_n}\frac{\partial C}{\partial t} = D'\nabla^2 C - \mu_a C + \frac{s}{c_n},$$

where $D' = D/c_n$ is referred to in the photon transfer literature as the *photon diffusion constant*. It has dimensions of length.

Two important quantities in radiation transfer are the *photon* or *particle fluence* and the *photon fluence rate*. The International Commission on Radiation Units and Measurements (ICRU) defines the particle fluence for any kind of particle, including photons as follows: At the point of interest construct a small sphere of radius a. Let the number of particles striking the surface of the sphere during some time interval have an *expectation value* N. (The expectation value is the mean of a set of measurements in the limit as the number of measurements becomes infinite.) The particle fluence Φ is the ratio $N/\pi a^2$, where πa^2 is the area of a great circle of the sphere, that is, the area of a circle having the same radius as the sphere. This is shown in Fig. 14.12 and is a generalization of our earlier use of Φ as the number of particles per unit area. It neatly avoids having to introduce obliquity factors, since for any direction the particles travel, one can construct a great circle on the sphere that is perpendicular to their path. The *particle fluence rate* is

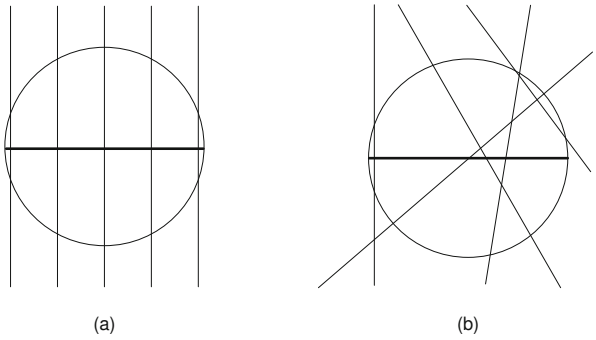


Fig. 14.12 The particle fluence is the ratio of the expectation or average value of the number of particles passing through the sphere to the area of a great circle of the sphere, πa^2. It depends on the total number of particles passing through the sphere, regardless of the direction they travel. The fluence is the same in each case shown: five particles traverse each sphere

$$\varphi = \frac{d\Phi}{dt}.$$

We saw in Chap. 4 that for a group of particles all traveling with the same speed, the number transported across a plane per unit area per unit time is equal to their concentration times their speed. The photon concentration is related to the photon fluence rate by $C = \varphi/c_n$, and the photon diffusion equation becomes

$$\frac{1}{c_n}\frac{\partial \varphi}{\partial t} = D'\nabla^2\varphi - \mu_a\varphi + s. \tag{14.27}$$

This is the form that is usually found in the literature. The units of each term are photons $\mathrm{m}^{-3}\ \mathrm{s}^{-1}$. One can show that[5]

$$D' = \frac{1}{3\left[\mu_a + (1-g)\mu_s\right]} = \frac{1}{3(\mu_a + \mu_s')}. \tag{14.28}$$

14.6.2 Continuous Measurements

If the tissue is continuously irradiated with photons at a constant rate, the term containing the time derivative vanishes. If

[4] Most papers in this field use c as the velocity of light in the medium. We prefer to reserve c for the fundamental constant, the velocity of light in vacuum.

[5] See, for example, Duderstadt and Hamilton (1976, pp. 133–136).

in addition we use a broad beam of photons so that we have a one-dimensional problem and we are far enough into the tissue so that the source term can be ignored, the model is

$$D' \frac{d^2\varphi}{dx^2} = \mu_a \varphi. \tag{14.29}$$

This has an exponential solution $\varphi = \varphi_0 e^{-\mu_{\text{eff}} x}$, where $\mu_{\text{eff}} = \left\{3\mu_a \left[\mu_a + (1-g)\mu_s\right]\right\}^{1/2}$. It is interesting to see what these numbers mean. Using the "typical" values from Sect. 14.5, the number of photons that have not interacted (are not yet attenuated) falls exponentially with a characteristic length or mean depth

$$\lambda_{\text{unatten}} = \frac{1}{\mu} = \frac{1}{\mu_a + \mu_s'} = \frac{1}{1505} = 0.66 \text{ mm}.$$

For the diffuse beam, the mean depth is about ten times this:

$$\lambda_{\text{diffuse}} = \frac{1}{\mu_{\text{eff}}} = \frac{1}{\sqrt{(3)(5)(1505)}} = 6.7 \text{ mm}.$$

These values are for a wavelength where the tissue is fairly transparent. The diffusion equation can be solved for other geometries that model the light source being used.[6] One problem with these measurements is that they give only μ_{eff}, which is a combination of μ_a and μ_s. Also, the path length may be ambiguous because the geometry cannot be modeled accurately.

14.6.3 Pulsed Measurements

A technique made possible by ultrashort light pulses from a laser is *time-dependent diffusion.* It allows determination of both μ_s and μ_a. A very short (150 ps) pulse of light strikes a small region on the surface of the tissue. A detector placed on the surface about 4 cm away records the multiply-scattered photons. A typical plot of the detected photon fluence rate is shown in Fig. 14.13. Patterson et al. (1989) have shown that the reflected fluence rate after a pulse is approximately

$$R(r,t) = \frac{z_0}{(4\pi D' c_n t)^{3/2} t} e^{-\mu_a c_n t} e^{-(r^2+z_0^2)/4D'c_n t}. \tag{14.30}$$

Here r is the distance of the detector from the source along the surface of the skin, $c_n t$ is the total distance the photon has traveled before detection, and $z_0 = 1/\left[(1-g)\mu_s\right]$ is the depth at which all the incident photons are assumed to scatter and become part of the diffuse photon pool. This curve fits Fig. 14.13 well and can be used to determine μ_a and $(1-g)\mu_s$. We can understand the various factors in Eq. 14.30. The last factor is a Gaussian spreading in the

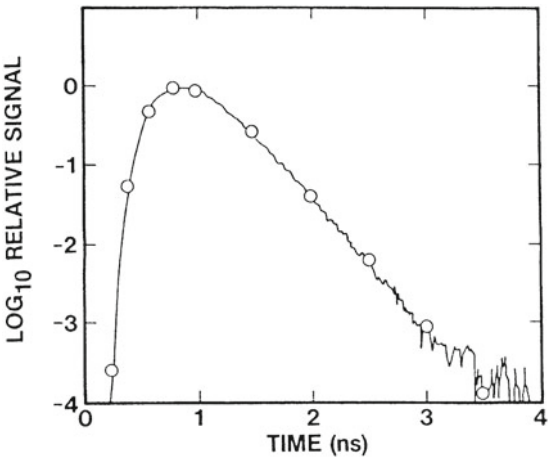


Fig. 14.13 Time-resolved infrared spectroscopy. The line is a measurement of the reflected photons from the calf of a human volunteer at a distance of 4 cm from the pulsed source. The wavelength is 760 nm. The circles are calculated using Eq. 14.30 and normalized to the peak value. (Source: Patterson et al. 1989. Copyright by the Optical Society of America.)

r direction away from the z axis where the photons were injected. This is a two-dimensional problem. Compare this with Eq. 4.77, which shows that in two dimensions $\sigma_r^2 = 4Dt$, and recall that $D = D'c_n$. The middle factor is the fraction of the photons in the pulse that have not been absorbed, $\exp(-\mu_a x)$, where x is the total distance the photons have traveled. The first factor is the normalization that reduces the amplitude of the Gaussian as it spreads.

A related technique is to apply a continuous laser beam whose amplitude is modulated at various frequencies between 50 and 800 MHz. The Fourier transform of Eq. 14.30 gives the change in amplitude and phase of the detected signal. Their variation with frequency can also be used to determine μ_a and μ_s.[7]

14.6.4 Refinements to the Model

The diffusion equation, Eq. 14.27, is an approximation, and the solution given, Eq. 14.30, requires some unrealistic assumptions about the boundary conditions at the surface of the medium ($z = 0$). Hielscher et al. (1995) compared experiment, Monte Carlo calculations, and solutions to the diffusion equation with three different boundary conditions. They found that Eq. 14.30 was the easiest to use but leads to errors in the estimates of the coefficients that become worse when the detector and source are close together. Their Monte Carlo calculations fit the data quite well. They also discuss

[6] See, for example, Grossweiner (1994, p. 98).

[7] See, for example, Sevick et al. (1991) or Pogue and Patterson (1994).

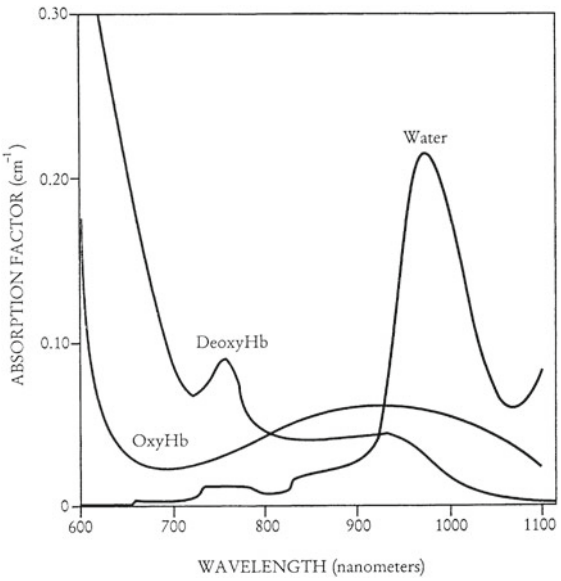


Fig. 14.14 The absorption coefficient μ_a for water, oxyhemoglobin, and deoxyhemoglobin. (Reprinted with permission from Yodh and Chance (1995). Copyright 1995, American Institute of Physics)

the reflections that occur when light goes from one medium into another with a different index of refraction.

14.7 Biological Applications of Infrared Scattering

There are a number of uses of light in the near infrared: some clinical and some in the laboratory.

14.7.1 Near Infrared (NIR)

Near-infrared light in the range 600–1000 nm is used to measure the oxygenation of the blood as a function of time by determining the absorption at two different wavelengths. Figure 14.14 shows the absorption coefficients for oxygenated and deoxygenated hemoglobin and water. The greater absorption of blue light in oxygenated hemoglobin makes oxygenated blood red. (The graph shows only wavelengths longer than 600 nm—red and infrared.) The wavelength 800 nm at which both forms of hemoglobin have the same absorption is called the *isosbestic point*. Measurements of oxygenation are made by comparing the absorption at two wavelengths on either side of this point.

One of the difficulties with these measurements is knowing the path length, since photons undergo many scatterings before being absorbed or reaching the detector. Scattering from many tissues besides hemoglobin distorts the signal. Nonetheless, *pulse oximeters* that fit over a finger are widely used. Webster (1997) provides a comprehensive discussion of the underlying physics, design, calibration, and use of pulse oximeters. The basic feature is that arterial blood flow is pulsatile, not continuous. Therefore, measuring the time-varying (AC) signal selectively monitors arterial blood and eliminates the contribution from venous blood and tissue. Scattering corrections must still be made (Farmer 1997; Wieben 1997).

Development of new applications for infrared scattering measurements continues as new detectors with different spectral sensitivities become available (Yamashita et al. 2001). Continuous sources are also used to determine blood oxygenation of tissue (Liu et al. 1995).

14.7.2 Optical Coherence Tomography (OCT)

Optical range measurements using the time delay of reflected or backscattered light from pulses of a few femtosecond (10^{-15} s) duration can be used to produce images similar to those of ultrasound A- and B-mode scans. The spatial extent of a 30 fs pulse in water is about 7 μm. Since it is difficult to measure time intervals that short, most measurements are done using interference properties of the light. *Optical coherence tomography* is conceptually similar to range measurements but uses interference measurements. It was first demonstrated by Huang et al. (1991) and has been developed extensively since then (see Schmitt 1999; Brezinski 2006 or Fercher et al. 2003). It is widely used in ophthalmology.

This is one topic for which we must use the electromagnetic wave model for light, since it depends on interference effects. Light waves differ from sound waves because the electric field in the wave is a vector perpendicular to the direction of propagation of the wave. This gives rise to an important effect—polarization—that we ignore.

Suppose that a wave $A\sin\frac{2\pi}{\lambda}(x-c_nt) = A\sin\omega(x/c_n - t)$ travels in a medium with index of refraction n. A detector responds to the energy fluence in the wave, which is proportional to the square of the amplitude averaged over time. The signal is $y \propto A^2\overline{\sin^2\omega(x/c_n - t)} = A^2/2$. The wave is split, travels two paths of different lengths, and is recombined at a detector. The signal is proportional to the power averaged over many cycles of the wave. The power is proportional to the square of the electric field:

$$y \propto (A/2)^2\overline{\left[\sin\omega(x_1/c_n - t) + \sin\omega(x_2/c_n - t)\right]^2}$$
$$= \frac{A^2}{4}\left(1 + \cos\frac{\omega}{c_n}(x_2 - x_1)\right). \qquad (14.31)$$

The signal oscillates between 0 and $A^2/2$ as the difference in path length is changed. When the path difference is zero,

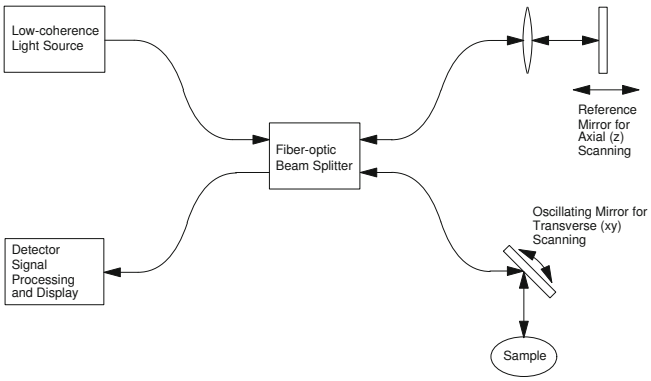


Fig. 14.15 The basic apparatus for optical coherence tomography. The features are described in the text

$y \propto A^2/2$, our original result. This dependence of the signal on path length forms the basis for *interferometry*, which can be used to measure changes in distance with high accuracy—counting maxima (fringes) as one path length is varied.

An important consideration is the *coherence* of the light beam: the number of cycles over which the phase of the wave does not change. When an atom emits light, the classical electromagnetic wave lasts for a finite time, τ_{coh} (often around 10^{-8} s). When another atom emits light, the phase is unrelated to the phase of light already emitted. This means that if $(x_2 - x_1)/c_n > \tau_{\text{coh}}$, the time average will go to zero.

Note that as long as light from a single source has been split and then recombined, the paths can be quite long. The interference fringes will be seen when the light is recombined and the path difference satisfies

$$x_2 - x_1 < c_n \tau_{\text{coh}}. \tag{14.32}$$

This provides a technique for determining the distance of a reflecting object from the light source, forming the basis for optical coherence tomography. A light source with a *short* coherence time is used for high resolution. The basic apparatus is shown in Fig. 14.15. Various light sources are used. The light pulse travels over an optical fiber to a 50/50 beam splitter. Part travels to the sample, where it is reflected back to the 50/50 coupler and then to the detector. The other half of the light goes to the reference mirror, where it is also reflected back to the detector. Changing the position of the reference mirror changes the depth of the image plane in the sample. The lateral beam position is changed to scan the sample, as in an ultrasound B-mode scan. Fig. 14.16 shows an image of the retina.

It is possible to make many kinds of images. Fig. 14.17 shows the parabolic velocity profile of blood flowing in a retinal blood vessel 176 μm diameter. It was obtained by measuring the Doppler shift in light scattered from moving blood cells. It is also possible to image glucose concentration, because glucose modifies the index of refraction and thereby the scattering coefficient (Esenaliev et al. 2001). Images are made of the surface layers of the skin, the eye, the walls of the mouth, teeth, larynx, esophagus, stomach, and intestine.

A number of tissues exhibit *birefringence*—the speed of light in the skin depends on the orientation of the electric field vector of the light wave with the cells in the tissue (de Boer et al. 2002). It is possible to make images with different orientations of the electric field vector to improve the resolution (Yasuno et al. 2002).

There are a number of offshoots to OCT, such as optical coherence microscopy and full-field OCM (Saint-Jalmes et al. 2002).

14.7.3 Raman Spectroscopy

Infrared and microwave probes are used extensively in the laboratory. Since the vibrational and rotational levels depend on the masses, separations, and forces between the various atoms bound in a molecule, it is not surprising that spectroscopy can be used to identify specific bonds. This is a useful technique in chemistry. Biological applications are difficult because the absorption coefficients are large; thin samples must be used, particularly in an aqueous environment.

One way around this is *Raman scattering*: the scattering of light in which the scattered photon does not have its original energy, but has lost or gained energy corresponding to a rotational or vibrational transition. The effect was discovered by C. V. Raman in 1928. Raman scattering can be done with light of any wavelength, from infrared to ultraviolet. An idealized example is shown in Fig. 14.18. If the scattering molecule was originally in the vibrational ground state and returns to a vibrational excited state, the Raman-scattered photon has less energy than the original photon. This is called *Stokes-Raman* scattering. If the scattering molecule was originally in a higher vibrational state and returns to the vibrational ground state, the Raman-scattered photon has higher energy than the original. The intensity of this *Anti-Stokes Raman* line will be less than the intensity of the Stokes-Raman line because populations of the original vibrational levels are governed by a Boltzmann factor. Figure 14.19 shows the Stokes-Raman shift spectrum for cholesterol. Many discussions of Raman spectroscopy are available. A fairly theoretical one by Berne and Pecora (1976) relies heavily on autocorrelation functions and spectral analysis that we saw in Chap. 11. Diem (1993) is a detailed text on vibrational spectroscopy, including Raman spectroscopy.

Raman spectroscopy has been used extensively for laboratory studies; many groups are exploring its utility for in vivo

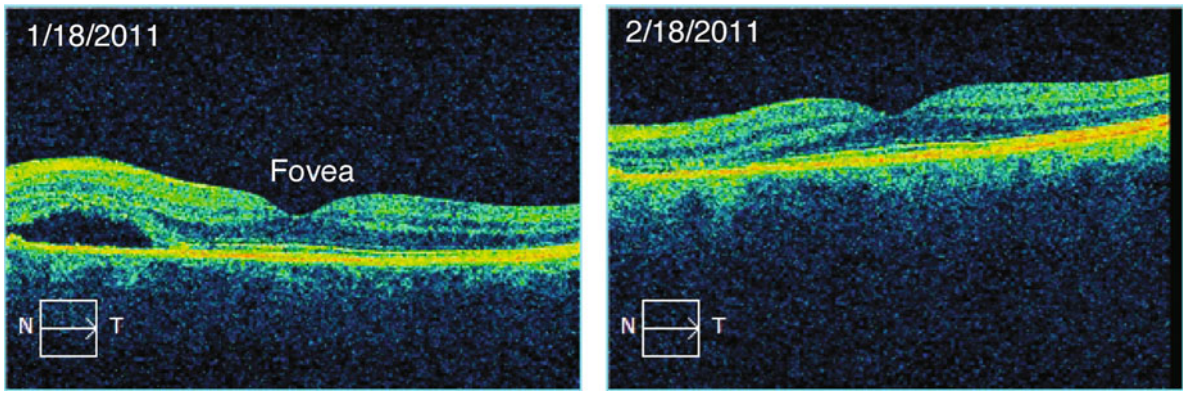


Fig. 14.16 Optical coherence tomograms of the retina. The square box in the lower left corner is about 500 μm on a side. *N* and *T* indicate nasal and temporal. The patient has a subchoroidal neovascular membrane (a collection of new and fragile blood vessels that leak) seen as a pocket of fluid on the left. The lesion was treated with a laser (photocoagulation). The tomogram 1 month later shows resolution of the fluid pocket. (Scans courtesy of Kirk Morgan, MD)

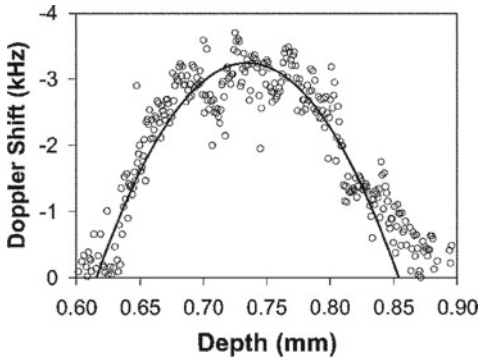


Fig. 14.17 The parabolic velocity profile of blood flowing in a single retinal vessel of diameter 176 μm. (Source: Yazdanfar et al. 2000. Used by permission)

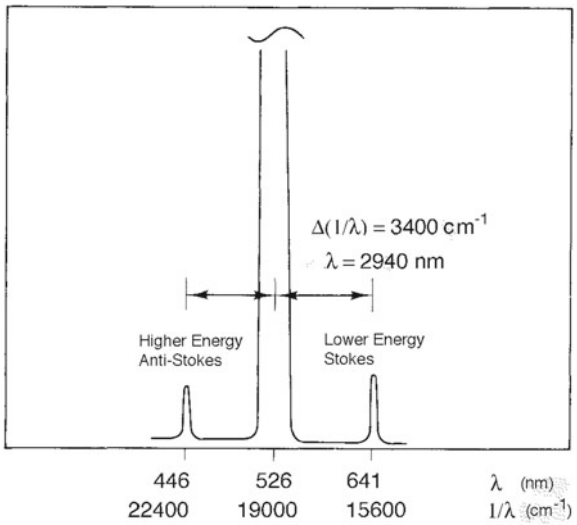


Fig. 14.18 In Raman scattering, a photon gains or loses energy due to a change in the energy of the scattering molecule. An idealized example for water is shown. The very intense line (the tall peak) has no energy change; the weak lines are Raman scattering. The abscissa is shown as wavelength λ and as reduced wave number $k/2\pi = 1/\lambda$. The Raman shift corresponds to $\Delta(1/\lambda) = 3400\ \text{cm}^{-1}$. The wavelength of this infrared transition is $\lambda = 2940$ nm, but the measurement is made near 500 nm

measurements (Hanlon et al. 2000). Infrared light between 800 and 1000 nm is usually used.

14.7.4 Far Infrared or Terahertz Radiation

For many years, there were no good sources or sensitive detectors for radiation between microwaves and the near infrared (0.1–100 THz). Developments in optoelectronics have solved both problems, and many investigators are exploring possible medical uses of THz radiation ("T rays"). Classical electromagnetic wave theory is needed to describe the interactions, and polarization (the orientation of the **E** vector of the propagating wave) is often important. The high attenuation of water in this frequency range means that studies are restricted to the skin or surface of organs such as the esophagus that can be examined endoscopically. Reviews are provided by Smye et al. (2001), Fitzgerald et al. (2002), and Zhang (2002). See the article by Armstrong (2012) for a survey of the challenges of using terahertz radiation.

14.8 Thermal Radiation

Any atomic gas emits light if it is heated to a few thousand kelvin. The light consists of a line spectrum. The famous yellow line of sodium has

$$\lambda = 589.2\ \text{nm},$$

$$\nu = c/\lambda = 509.2\ \text{THz},$$

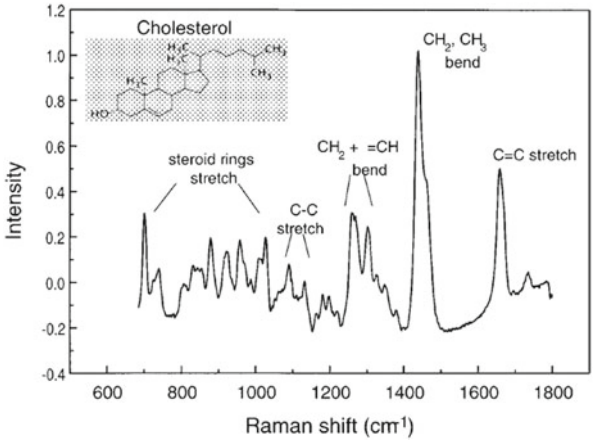


Fig. 14.19 Vibrational Raman lines for cholesterol. A continuous background has been subtracted. The abscissa is $1/\lambda = E/hc$. (Source: Hanlon et al. 2000. Used by permission)

Table 14.3 Approximate color temperatures. The range of values reflects differences between scales established by different observers

Color	T (K)
Red, just visible in daylight	750–800
"Cherry" red	975–1175
Yellow	3000–4000
White	5000–6000
Dazzling (bluish white)	> 10 000

$$E = h\nu = hc/\lambda = 3.38 \times 10^{-19}\ \text{J} = 2.11\ \text{eV}.$$

These photons are emitted when sodium atoms lose 2.11 eV and return to their ground state. If the sodium atoms are excited by thermal collisions, the probability that a sodium atom is in the excited state, relative to the probability that it is in the ground state, is given by the Boltzmann factor

$$\frac{P_{\text{excited}}}{P_{\text{ground}}} = e^{-E/k_B T}.$$

At room temperature $k_B T = 4.14 \times 10^{-21}$J, so $e^{-E/k_B T} = e^{-81.5} = 3.8 \times 10^{-36}$. The number of atoms in the excited state is utterly negligible. If the temperature is raised to 1500 K, $e^{-E/k_B T}$ is 8×10^{-8}, and enough atoms are excited to give off light as they fall back to the ground state.

If a piece of iron is heated to 1500 K, it glows with a red–orange color. Table 14.3 relates apparent color to temperature for a glowing metal. If the light is analyzed with a spectroscope, it is found to consist of a continuous range of colors rather than discrete lines.

The difference between the spectra of single atoms and the spectra of solids and liquids can be understood from the following argument. If we have N isolated identical atoms, each atom has an energy level at the energy shown in Fig. 14.20a. There are a total of N levels, one for each atom. When two of these atoms are brought close together,

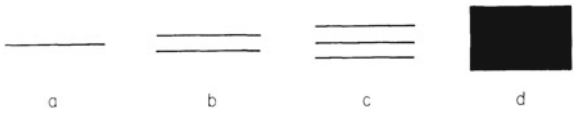


Fig. 14.20 The splitting of energy levels as many atoms are brought together. **a** A single atom. **b** Two atoms. **c** Three atoms. **d** Many atoms

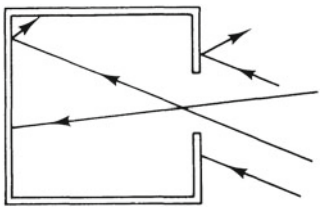


Fig. 14.21 A small hole in the wall of a cavity is a better blackbody than the walls of the cavity are. Any light that enters the hole must be reflected several times before emerging. It can be absorbed by the wall at any reflection. If the walls appear black, the hole appears even blacker. (The walls are highly absorbing diffuse reflectors)

the levels shift slightly and split into two closely spaced levels because of interaction between the atoms. The two levels for a pair of atoms are shown in Fig. 14.20b. If three atoms are brought close together, the level splits into three levels as shown in Fig. 14.20c. If a large number of atoms are brought close together, the N levels spread out into a *band*, Fig. 14.20d. Transitions from one band to another can have many different energies, and photons with a continuous range of energies can be emitted or absorbed.

The relative number of photons of different energies that will be emitted or absorbed depends on the nature of the substance. Glass and sodium chloride crystals are nearly transparent in the visible spectrum because the spacing of the levels is such that no photons of these energies are absorbed. When such substances are heated enough to populate the higher energy levels, no photons of these energies are emitted.

A substance that has so many closely spaced levels that it can absorb every photon that strikes it appears black. It is called a *blackbody*. It is difficult if not impossible to make a surface that is completely absorbing; the absorption can be improved by making a cavity, as in Fig. 14.21. Photons entering the hole in the cavity bounce from the walls many times before chancing to pass out through the hole again, and they therefore have a greater chance of being absorbed. Such a hole in a cavity is a better approximation to a blackbody than is the absorbing material lining the cavity.

If the surface is not completely absorbing, we define the *emissivity* $\epsilon(\lambda)$, which is the fraction of light absorbed at wavelength λ. (Why emission and absorption are closely related is discussed below.) If the light all passes through some transparent material or is completely reflected, then $\epsilon = 0$; if

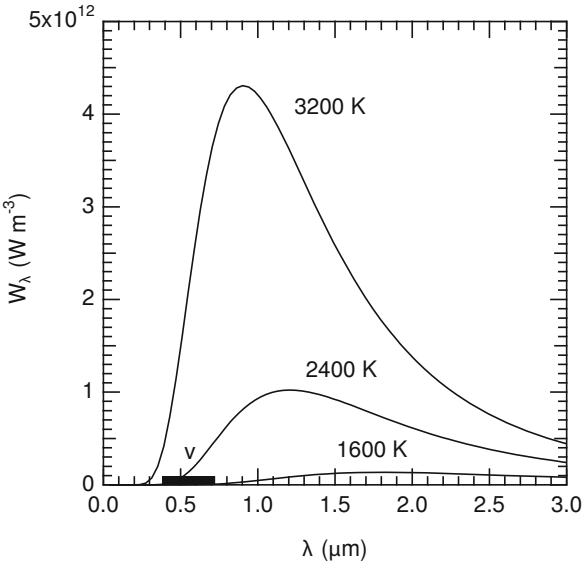


Fig. 14.22 The blackbody radiation function for several temperatures. The visible spectrum is marked by v

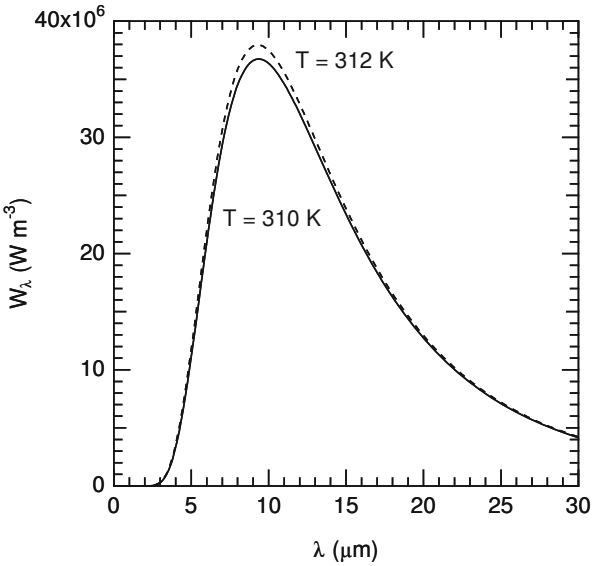


Fig. 14.23 The blackbody radiation function $W_\lambda(\lambda, T)$ for $T = 310$ K and $T = 312$ K

it is all absorbed, $\epsilon = 1$. A blackbody has $\epsilon(\lambda) = 1$ for all wavelengths. An object for which $\epsilon(\lambda)$ is constant but less than 1 is called a gray body.

When a blackbody is heated, the light given off has a continuous spectrum because the energy levels are so closely spaced. By imagining two interacting black bodies in equilibrium, one can argue[8] that the amount of energy emitted by a blackbody depends only on its temperature and not on the nature of the surfaces.

The spectrum of power per unit area emitted by a completely black surface in the wavelength interval between λ and $\lambda + d\lambda$ is

$$W_\lambda(\lambda, T)d\lambda,$$

a universal function called the *blackbody radiation function*. It has units of W m^{-3}, although it is often expressed as W cm^{-2} μm^{-1}. The value of W_λ is plotted for several different temperatures in Fig. 14.22. As the black surface or cavity walls become hotter, the spectrum shifts toward shorter wavelengths, which is consistent with the observations in Table 14.3. The visible region of the spectrum is marked on the abscissa in Fig. 14.22; even at 3200 K, most of the energy is radiated in the infrared.

Figure 14.23 plots $W_\lambda(\lambda, T)$ for two temperatures near body temperature (37° C=310 K). Compare the scales of Figs. 14.22 and 14.23, and note how much more energy is emitted by a blackbody at the higher temperature and how it is shifted to shorter wavelengths.

Much work was done on the properties of blackbody or thermal or cavity radiation in the late 1800s and early 1900s. While some properties could be explained by classical physics, others could not. The description of the function $W_\lambda(\lambda, T)$ by Max Planck is one of the foundations of quantum mechanics. We will not discuss the history of these developments, but will simply summarize the properties of the blackbody radiation function that are important to us.

The value of $W_\lambda(\lambda, T)$ is given by

$$W_\lambda(\lambda, T) = \frac{2\pi c^2 h}{\lambda^5 \left(e^{hc/\lambda k_B T} - 1\right)}. \tag{14.33}$$

Consider the expression $e^{hc/\lambda k_B T}$ in the denominator. Since light consists of photons of energy $E = h\nu = hc/\lambda$, the expression in parentheses in the denominator is $e^{E/k_B T} - 1$. For very large energies (short wavelengths) the 1 can be neglected and this expression is the Boltzmann factor.

We can find the total amount of power emitted per unit surface area by integrating[9] Eq. 14.33:

$$\begin{aligned} W_{\text{tot}}(T) &= \int_0^\infty W_\lambda(\lambda, T)d\lambda \\ &= \frac{2\pi^5 k_B^4}{15c^2h^3} T^4 = \sigma_{SB} T^4. \end{aligned} \tag{14.34}$$

This is the *Stefan–Boltzmann law*. The Stefan–Boltzmann constant, which is traditionally denoted by σ_{SB} but which

[8] For a brief discussion, see Schroeder (2000).

[9] This is not a simple integration. See Gasiorowicz (2003, p. 3).

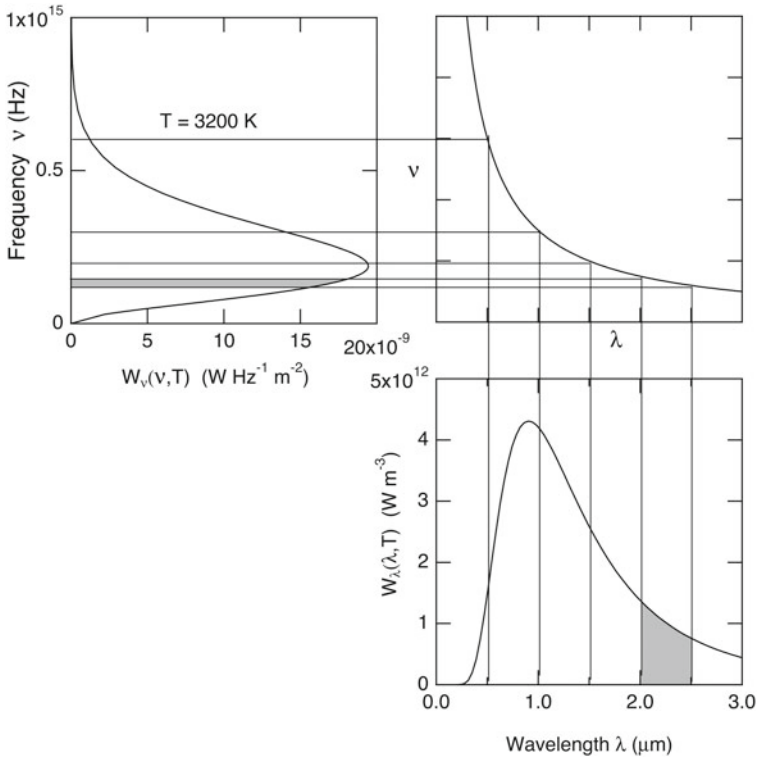


Fig. 14.24 The transformation from $W_\lambda(\lambda, T)$ to $W_\nu(\nu, T)$ is such that the same amount of power per unit area is emitted in wavelength interval $(\lambda, d\lambda)$ and the corresponding frequency interval $(\nu, d\nu)$. (For example, the two shaded areas are the same.) The spectrum shown is for a blackbody at 3200 K

has no relationship to cross section, was known empirically before Planck's work. It has the numerical value

$$\sigma_{SB} = 5.67 \times 10^{-8} \text{W m}^{-2} \text{ K}^{-4}. \tag{14.35}$$

Early experiments were performed with equipment that measured the radiation function versus wavelength. It is also possible to measure versus frequency. To rewrite the radiation function in terms of frequency, let λ_1 and $\lambda_2 = \lambda_1 + d\lambda$ be two slightly different wavelengths, with power $W_\lambda(\lambda, T)d\lambda$ emitted per unit surface area at wavelengths between λ_1 and λ_2. The same power must be emitted[10] between frequencies $\nu_1 = c/\lambda_1$ and $\nu_2 = c/\lambda_2$:

$$W_\nu(\nu, T)d\nu = W_\lambda(\lambda, T)d\lambda. \tag{14.36}$$

Since $\nu = c/\lambda$, $d\nu/d\lambda = -c/\lambda^2$, and

$$|d\nu| = +\frac{c}{\lambda^2}\,|d\lambda|\,. \tag{14.37}$$

[10] $W_\lambda(\lambda, T)$ and $W_\nu(\nu, T)$ do not have the same functional form. In fact, they have different units. The units of $W_\lambda(\lambda, T)$ are W m^{-3}, while those of $W_\nu(\nu, T)$ are W s m^{-2}.

Equations 14.33–14.37 can be combined to give

$$W_\nu(\nu, T) = \frac{2\pi\nu^2(h\nu)}{c^2(e^{h\nu/k_BT} - 1)}. \tag{14.38}$$

This transformation is shown in Fig. 14.24. The amount of power per unit area radiated in the 0.5-μminterval between two of the vertical lines in the graph on the lower right is the area under the curve of W_λ between these lines. The graph on the upper right transforms to the corresponding frequency interval. The radiated power, which is the area under the W_ν curve between the corresponding frequency lines on the upper left, is the same. Note that the peaks of the two curves are at different frequencies or wavelengths. We will see this same transformation again when we deal with x rays. We see in the figures above that at higher temperatures the peak occurs at shorter wavelengths. Equation 14.33 can be differentiated to show that at temperature T, the peak in W_λ occurs at wavelength

$$\lambda_{\max} T = \frac{hc}{4.9651 k_B} = 2.90 \times 10^{-3} \text{ m K}. \tag{14.39}$$

Similarly, we can differentiate Eq. 14.38 to show that

$$\frac{\nu_{\max}}{T} = \frac{2.82144 k_B}{h} = 5.88 \times 10^{10} \text{ K}^{-1}\text{s}^{-1}.$$

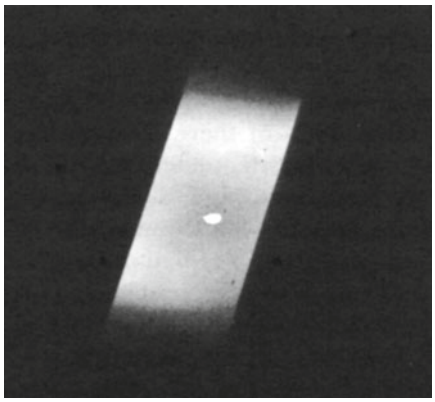


Fig. 14.25 A photograph of an incandescent tungsten tube with a small hole drilled in it. The radiation emerging from the hole is brighter than that from the tungsten surface. (Source: Halliday et al. (1992). Copyright ©1992 John Wiley & Sons. Reproduced by permission of John Wiley & Sons)

The product $\lambda_{\max}\nu_{\max} = 1.705 \times 10^8 \text{m s}^{-1} = 0.57c$.

All this is true for a blackbody. Thermodynamic arguments can be made to show that if a body does not completely absorb light at some wavelength, that is $\epsilon(\lambda) < 1$, then the power emitted at that wavelength is

$$\epsilon(\lambda) W_\lambda(\lambda, T). \tag{14.40}$$

This is the same $\epsilon(\lambda)$ that was introduced earlier in this section. It is called the *emissivity* of the surface. This implies that a surface that appears blackest when it is absorbing radiation will be brightest when it is heated. Figure 14.25 shows a small hole in a piece of tungsten that has been heated. The hole forms the opening to a cavity and is therefore more absorbing than is the tungsten surface. When heated, the hole emits more light than the tungsten surface.

14.9 Infrared Radiation from the Body

The body radiates energy in the infrared, and this is a significant source of energy loss. Infrared radiation has been used for over 40 years to image the body, but the value of the technique is still a matter of debate. We saw earlier how the *scattering* of infrared radiation by the body can be used to learn information about tissue beneath the surface.

Measurements of the emissivity of human skin have shown that for 1 μm $< \lambda \leq$ 14 μm, $\epsilon(\lambda) = 0.98 \pm 0.01$. This value was found for white, black, and burned skin (Steketee 1973). In the infrared region in which the human body radiates, the skin is very nearly a blackbody. Let us apply Eq. 14.34 to see what the blackbody radiation from the human body is. The total surface area of a typical adult male is about 1.73 m^2. The surface temperature is about

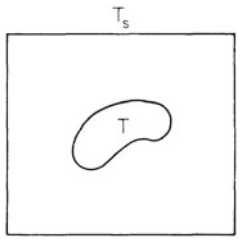


Fig. 14.26 A blackbody at temperature T within a container with wall temperature T_s

33 °C = 306 K (this is less than the core temperature of 310 K). Therefore the total power radiated is $w_{\text{tot}} = SW_{\text{tot}} = S\sigma_{SB}T^4 = 860$ W. This is a large number, nearly nine times the basal metabolic rate of 100 W. The reason it is so large is that it assumes the surroundings are at absolute zero, or that the subject is radiating in empty space with no surroundings. When there are nearby surfaces, radiation from them is received by the subject, and the net radiation is considerably less than 860 W. The easiest arrangement for which to calculate the net heat loss is a blackbody at temperature T surrounded by a similar surface at temperature T_s (Fig. 14.26). At equilibrium the temperature of both objects is the same, $T = T_s$ and the power emitted by the body is equal to the power absorbed. Increasing T increases the power emitted according to $W_{\text{tot}} = \sigma_{SB}T^4$. The body then emits more power than it absorbs. Equilibrium is restored when the body has cooled or the surroundings have warmed so that the temperatures are again the same. Thermodynamic arguments can be made to show that the net power radiated by the body is

$$w_{\text{tot}} = S\sigma_{SB}(T^4 - T_s^4). \tag{14.41}$$

If the object is not a blackbody or the wall temperature is not uniform, the net power loss is more complicated. However, this model represents a considerable improvement over our previous calculation. Suppose that the surroundings are at a temperature $T_s = 293$ K (20 °C). The net loss is

$$w_{\text{tot}} = (1.73)(5.67 \times 10^{-8})(306^4 - 293^4) = 137 \text{ W}.$$

This says that a nude subject surrounded by walls at 20 °C would have to exercise to maintain body temperature, even if the air temperature were warm enough so that heat conduction and convection losses were zero.

If you have lived in a cold climate, you have probably felt cold in a room at night when the drapes are open, even though a thermometer records air temperatures that should be comfortable. This is because of radiation from you to the cold window. The glass is transparent only in the visible range; for infrared radiation it is opaque and has a high emissivity. The

radiation of the cold window back to you is much less than your radiation to it, and you feel cold.

This same problem can occur with a premature infant in an incubator. If the incubator is placed near a window, one wall of the incubator can be cooled by radiation to the window. The infant can be cooled by radiation to the wall of the incubator, even though a shiny (low-emissivity) thermometer in the incubator records a reasonable air temperature. One solution is to be careful where an incubator is placed and insulate its walls; another is to redesign incubators with a feedback loop controlling the infant's temperature.

Infrared radiation can be used to image the body. Two types of infrared imaging are used. In infrared photography the subject is illuminated by an external source with wavelengths from 700 to 900 nm. The difference in absorption between oxygenated and deoxygenated hemoglobin allows one to view veins lying within 2 or 3 mm of the skin. Either infrared film or a solid-state camera can detect the reflected radiation.

Thermal imaging detects thermal radiation from the skin surface. Significant emission from human skin occurs in the range 4–30 μm, with a peak at 9 μm (Fig. 14.23). The detectors typically respond to wavelengths below 6–12 μm. *Thermography* began about 1957 with a report that skin temperature over a breast cancer was slightly elevated. There was great hope that thermography would provide an inexpensive way to screen for breast cancer, but there have been too many technical problems. Normal breasts have more variability in vascular patterns than was first realized, so that differences of temperature at corresponding points in each breast are not an accurate diagnostic criterion. The thermal environment in which the examination is done is extremely important. The sensitivity (ability to detect breast cancer) is too low to use it as a screening device. Thermography has also been proposed to detect and to diagnose various circulatory problems. Thermography is not widely accepted (Blume 1993; Vreugdenburg et al. 2013), though it still has its proponents (Lahiri et al. 2012).

Infrared radiation from the tympanic membrane (eardrum) and ear canal is used to measure body temperature. One instrument is based on a *pyroelectric* crystal, which generates a voltage when heated (Fraden 1991). The sensors have a permanent electric dipole moment whose magnitude changes with temperature.

14.9.1 Atherosclerotic Coronary Heart Disease

Atherosclerotic coronary heart disease (ACHD) has been or is being studied with every imaging technique described in this book. All of the techniques are invasive: a *catheter* is inserted into the artery in question. In ACHD, a fatty *plaque* forms in the *lumen* (interior passageway) of the artery.

The standard technique is coronary artery *angiography:* the heart is imaged by x-ray fluoroscopy (see Chap. 16) while a dye opaque to x rays is introduced in the vessel. This allows accurate determination of the degree of *stenosis* (blocking) of the vessel. It has been thought that when the artery is nearly blocked, the restricted blood flow leads to a *myocardial infarct* (heart attack). It has recently been realized that smaller plaques may become disrupted and lead to a myocardial infarct. Current research seeks to learn what makes these particular plaques *vulnerable*. There is an extensive literature, reviewed by MacNeill et al. (2003) and Verheye et al. (2002).

In *intravascular ultrasound* (IVUS), a 20–40 MHz transducer at the end of the catheter can detect calcium (which deposits in areas of tissue injury). IVUS *elastography* measures how the arterial wall changes during the pressure variations of the cardiac cycle, in the hope that changes in elasticity will indicate vulnerable plaques.

Coronary *angioscopy* attempts to directly view the arterial wall using a tiny fiber-optic *endoscope*. A serious problem here is blood getting between the tip of the endoscope and the arterial wall. This has been solved by temporarily *occluding* (blocking) the artery "upstream" with a balloon catheter or by flushing the area with saline solution.

Thermography has also been explored, first with a temperature-sensitive thermistor, and also with an infrared imaging mirror. Areas of inflammation have a somewhat higher temperature than surrounding areas.

Both Raman spectroscopy and near-infrared spectroscopy have been used.

In intravascular magnetic resonance imaging (MRI is described in Chap. 18), the detector coil is made small enough to fit at the tip of the catheter.

14.9.2 Photodynamic Therapy

Photodynamic therapy (PDT) uses a drug called a photosensitizer that is activated by light (Zhu and Finlay 2008; Wilson and Patterson 2008). PDT can treat accessible solid tumors such as basal cell carcinoma, a type of skin cancer (see Sect. 14.10.4). An example of PDT is the surface application of 5-aminolevulinic acid, which is absorbed by the tumor cells and is transformed metabolically into the photosensitizer protoporphyrin IX. When this molecule interacts with light in the 600–800-nm range (red and near infrared), often delivered with a diode laser, it converts molecular oxygen into a highly reactive singlet state that causes necrosis, apoptosis (programmed cell death), or damage to the vasculature that can make the tumor ischemic. Some internal tumors can be treated using light carried by optical fibers introduced through an endoscope.

14.10 Blue and Ultraviolet Radiation

The energy of individual photons of blue and ultraviolet light is high enough to trigger chemical reactions in the body. These can be both harmful and beneficial. A beneficial effect is the use of blue light to treat neonatal jaundice. The most common harmful effect is the development of sunburn, skin cancer, and premature aging of the skin.

14.10.1 Treatment of Neonatal Jaundice

Neonatal jaundice occurs when *bilirubin* builds up in the blood. Bilirubin is a toxic waste product of the decomposition of the hemoglobin that is released when red blood cells die (*hemolysis*). Bilirubin is insoluble in water and cannot be excreted through either the kidney or the gut. It is excreted only after being conjugated with glucuronic acid in the liver. Bilirubin monoglucuronate and bilirubin diglucuronate are both water soluble. They are excreted in the bile and leave via the gut. Some newborns have immature livers that cannot carry out the conjugation. In other cases there is an increased rate of hemolysis, and the liver cannot keep up. The serum bilirubin level can become quite high, leading to a series of neurological symptoms known as *kernicterus*. The abnormal yellow color of the skin called *jaundice* is due to bilirubin in the capillaries under the skin.

When the skin of a newborn with jaundice is exposed to bright light, the jaundice color goes away. Photons of blue light have sufficient energy to convert the bilirubin molecule into more soluble and apparently less harmful forms (McDonagh 1985). Photons of longer wavelength have less energy and are completely ineffective. The standard form of phototherapy used to be to place the baby "under the lights." Since the lights were bright and also emitted some ultraviolet, it was necessary to place patches over the baby's eyes. Also, since the baby's skin had to be exposed to the lights, it had to be placed in an incubator to keep it warm. A fiberoptic blanket has been developed that can be wrapped around the baby's torso under clothing or other blankets. The optical fibers conduct the light from the source directly to the skin. Eye patches are not needed, and the baby can be fed and handled. Typical energy fluence rates are $(4–6) \times 10^{-2}$ W m^{-2} nm^{-1} in the range 425–475 nm. Acceptance by nursing staff and parents is very high (Murphy and Oellrich 1990). The blanket can be used for home treatment.

14.10.2 The Ultraviolet Spectrum

Ultraviolet light can come from the sun or from lamps. The maximum intensity of solar radiation is in the green, at about 500 nm. The sun emits approximately like a thermal radiator at a temperature of 5800 K. Figure 14.27 shows a 5800 K thermal radiation curve. The power per unit area from the sun at all wavelengths striking the earth's outer atmosphere, the *solar constant*, calculated by regarding the sun as a thermal radiator, is 1390 W m^{-2}. Satellite measurements give 1372 W m^{-2} (Madronich 1993). Because of reflection, scattering, atmospheric absorption, and so forth, the amount actually striking the earth's surface is about 1000 W m^{-2}. Figure 14.27 also shows the effect of absorption of sunlight in the atmosphere. The sharp cut off at 320 nm is due to atmospheric ozone (O_3), which absorbs strongly from 200 to 320 nm. Molecular oxygen absorbs strongly below 180 nm.

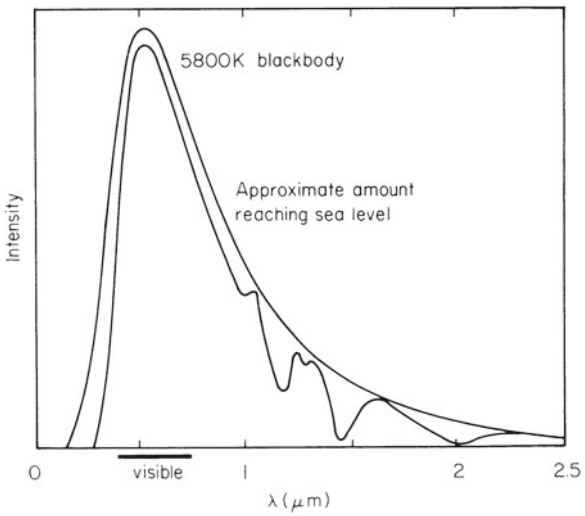


Fig. 14.27 The solar spectrum and the approximate spectrum reaching the earth after atmospheric attenuation

The ultraviolet spectrum is qualitatively divided into the following regions:

UVA	315–400 nm
UVB	280–315 nm[11]
UVC or middle UV	200–280 nm
Far UV	120–200 nm
Extreme UV	10–120 nm

Only the first three are of biological significance, because the others are strongly absorbed in the atmosphere.

Madronich (1993) gives a detailed discussion of the various factors that reduce the ultraviolet energy reaching the earth's surface. The sensitivity of DNA decreases as the wavelength increases. Figure 14.28 shows the solar radiation reaching the ground when the sun is at different angles from the zenith (directly overhead), weighted for DNA sensitivity.

[11] In Europe the range of UVB radiation is 290–300 nm.

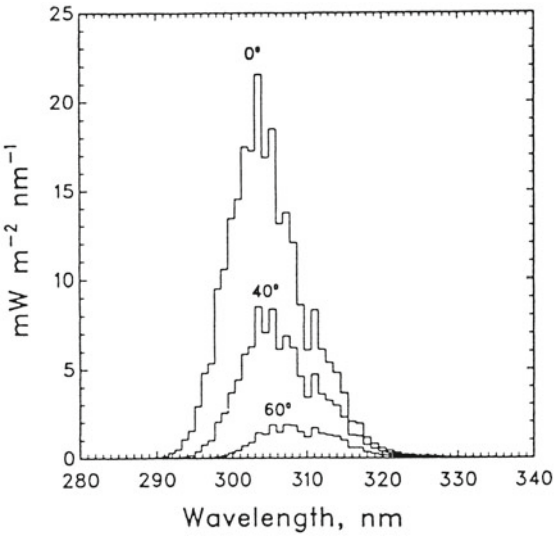


Fig. 14.28 Spectral dose rates weighted for ability to damage DNA for three different angles of the sun from overhead. The calculation assumes clear skies and an ozone layer of 300 Dobson units (1 DU = 2.69×10^{20} molecule m^{-2}). (Source: Madronich (1993). With kind permission of Springer Science and Business Media.)

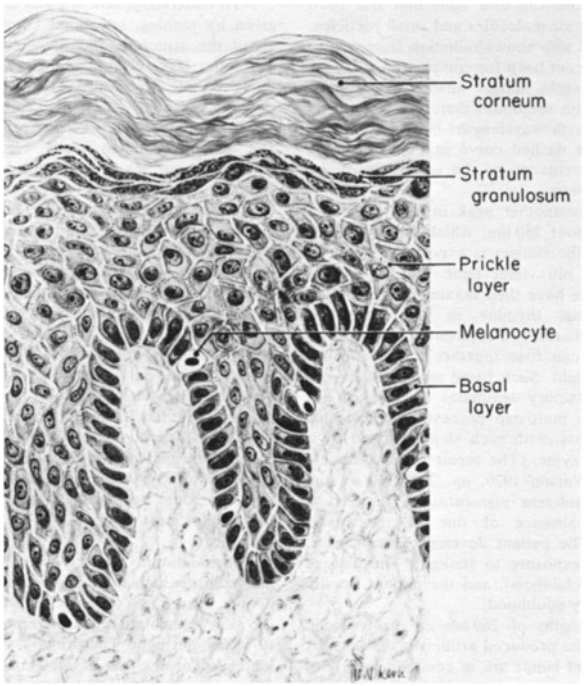


Fig. 14.29 The epidermis. The basal layer contains the cells from which the other layers are derived. As the cells move toward the surface they become the prickle layer and the stratum granulosum. The stratum corneum is dead cellular debris. The melanocytes, which produce melanin granules, are in the basal layer. (Reprinted from Pillsbury and Heaton 1980 with permission from Elsevier.)

Biological effects of ultraviolet light are reviewed by Diffey (1991).

14.10.3 Response of the Skin to Ultraviolet Light

There are several responses of the skin to ultraviolet light. In order to understand them one must know something about the anatomy and physiology of skin. The outer layer of the skin, the *epidermis*, consists of three sublayers (Fig. 14.29). A single layer of *basal cells* is on the inside. Most of these cells produce keratin, a protein that gives the outer layers of skin its strength. About 10 % of the cells are *melanocytes* that produce the pigment *melanin*. Next comes a sublayer of about seven cells, called the *prickle layer*. On top of this is a two- or three-cell layer called the *stratum granulosum* or granular layer. The surface is a layer of dead cells, primarily *keratin* and cellular debris, called the *stratum corneum* or horny layer. Basal cells are constantly produced in the basal layer, migrate outward, become the stratum corneum, and are sloughed off.

In order to discuss injury to tissue, both here by ultraviolet light and in later chapters by x rays, we need to introduce some specialized terms. The body's immediate (*acute*) response to an injury, whether it is an infection, a bump, a cut or a burn, is the *inflammatory response* described on page 122. Prolonged (*chronic*) irritation may result in abnormal cell growth. The abnormalities that result in organs or tissues that are larger than normal are *hypertrophy*, an enlargement of existing cells, and *hyperplasia*, an enlargement due to the formation of new cells. The aberrations in cell growth patterns are shown in Table 14.4. They are *metaplasia*, *dysplasia*, and *anaplasia*. Metaplasia is reversible and goes away if the stimulus or irritant is removed. Dysplasia is sometimes reversible and sometimes progresses to become cancerous. Anaplastic changes are present in nearly all forms of cancer. Anaplasia may result from dysplasia, or it may arise directly from normal cells.

The acute effect of ultraviolet radiation is reddening of the skin or *erythema* due to increased blood flow in the *dermis*, the layer beneath the epidermis. This is part of the inflammatory reaction. The amount of energy that just produces detectable erythema is called the *minimum erythemal dose*. The 1987 *erythema action spectrum* adopted by the CIE[12] shows the relative sensitivity of the skin versus wavelength

[12] *Commission International de l'Eclairage* or International Commission on Illumination.

Table 14.4 Abnormal changes in tissue

Metaplasia	A reversible change in which one cell type is replaced by another.
Dysplasia	Variation in size, shape, and organization of the cells. Literally, "deranged development"
Anaplasia	A marked, irreversible, and regressive change from adult cells that are differentiated in form to more primitive, less differentiated cells

Differences between benign and malignant tumors

Characteristic	Benign	Malignant
Histologic differentiation (microscopic appearance)	Often typical of the tissue of origin	Not well differentiated; atypical cells
Mode of growth	Expands inside a capsule	Expansive; also infiltrative, with no capsule
Rate of growth	Progressive; usually slow; few cells undergoing mitosis (division)	May be rapid, with many cells undergoing mitosis
Metastasis (distant spread)	Absent	Frequently present

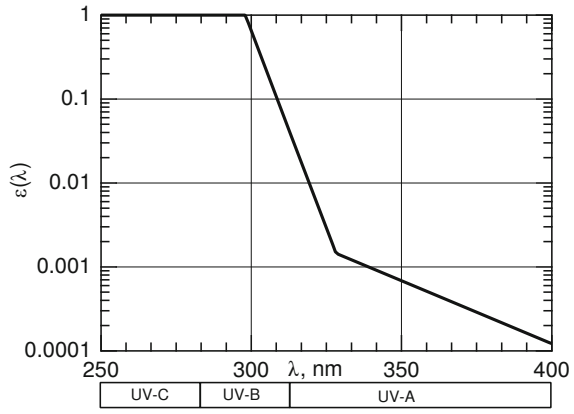


Fig. 14.30 The erythema action spectrum $\epsilon(\lambda)$ for ultraviolet light, as adopted by the CIE in 1987

for the production of erythema. It is

$$\epsilon(\lambda) = \begin{cases} 1.0, & 250 \le \lambda \le 298 \text{ nm} \\ 10^{0.094(298-\lambda)}, & 298 \le \lambda \le 328 \text{ nm} \\ 10^{0.015(139-\lambda)}, & 328 \le \lambda \le 400 \text{ nm}. \end{cases} \quad (14.42)$$

This is plotted in Fig. 14.30. The minimum erythemal dose at 254 nm is about 6×10^7 J m^{-2}. Early effects on skin include sunburn, tanning (now thought to be an injury response), and thickening. Daily exposure for 2–7 weeks causes a three- to fivefold thickening of the stratum corneum.

Some patients have an abnormally high sensitivity to ultraviolet exposure. They may exhibit abnormal photosensitivity because of various diseases or from taking drugs such as phenothiazines (a class of tranquilizers), sulfa drugs, dimethylchlortetracycline, the antidiabetic sulfonureas, thiazide diuretics, and even from drinking quinine water. Photocontact dermatitis is caused by interaction of photons with substances placed on the skin, such as perfumes containing furocoumarins, lime peel, fungi, and fluorescein dye used in lipsticks.

14.10.4 Ultraviolet Light Causes Skin Cancer

Chronic exposure to ultraviolet radiation causes premature aging of the skin. The skin becomes leathery and wrinkled and loses elasticity. The characteristics of photo-aged skin are quite different from skin with normal aging (Kligman 1989). UVA radiation was once thought to be harmless. We now understand that UVA radiation contributes substantially to both premature skin aging and skin cancer. This can be understood in the context of studies showing that both UVA and UVB suppress the body's immune system, and that this immunosuppression plays a major role in cancer caused by ultraviolet light (Kripke 2003; Moyal and Fourtanier 2002).

There are three types of skin cancer. *Basal-cell carcinoma* (BCC) is most common, followed by *squamous-cell carcinoma* (SCC). These are together called *nonmelanoma* or *nonmelanocytic skin cancer* (NMSC). Basal-cell carcinomas can be quite invasive (Fig. 16.40) but rarely metastasize or spread to distant organs. Squamous-cell carcinomas are more prone to metastasis. *Melanomas* are much more aggressive and frequently metastasize.

Armstrong and Kricker (2001) review the epidemiology of skin cancer. There are geographic variations of *incidence*, the number of newly diagnosed cases per 100,000 population per year. Incidence in New Mexico around 1980 for the three types of skin cancer is given in Table 14.5 for Anglos and Hispanics. Their review includes ambient solar radiation, ethnic origin, color of unexposed skin, personal exposure history, and personal use of skin protection.

The International Agency for Research on Cancer (IARC 2009) has classified all UV radiation (including UVA) as "Group 1, carcinogenic to humans."

Table 14.5 Estimates of skin cancer incidence rates per 100,000 in New Mexico, about 1980. (From Fig. 3 in Armstrong and Kricker 2001)

Cancer type	Population	Males	Females
Melanoma	Anglo	11.6	11.4
	Hispanic	1.2	1.5
SCC	Anglo	143	55
	Hispanic	13	12
BCC	Anglo	495	304
	Hispanic	64	35

SCC squamous-cell carcinoma, *BCC* basal-cell carcinoma

There has been an alarming increase in the use of tanning parlors by teenagers and young adults. These emit primarily UVA, which can cause melanoma. Exposure rates are two to three times greater than solar radiation at the equator at noon (Schmidt 2012). Many states now prohibit minors from using tanning parlors. Proponents of tanning parlors point out that UVB promotes the synthesis of vitamin D; however, the exposure to UVB in a tanning parlor is much higher than needed by the body for vitamin D production. Tanning as a source of Vitamin D is no longer recommended at any age level (Barysch et al. 2010). An analysis by Lazovich et al. (2010) concludes: "our results add considerable weight to the IARC report that indoor tanning is carcinogenic in humans and should be avoided to reduce the risk of melanoma." Australia has the highest incidence of skin cancer in the world and is a leader in studying and mitigating skin cancer. A recent review by O'Sullivan and Tait (2014) makes an even stronger recommendation: "The research to date supports a complete ban of indoor tanning as it has shown that less stringent regulations are ineffective due to the lack of adherence to them and enforcement of them. Australia and New Zealand are in a powerful position to lead the developed world by imposing a complete ban on indoor tanning. It is imperative to act on this evidence to reduce the risk of further avoidable morbidity and mortality."

14.10.5 Protection From Ultraviolet Light

Protection from the sun certainly reduces erythema and probably reduces skin cancer. Protection is most important in childhood years, both because children receive three times the annual sun exposure of adults and because the skin of children is more susceptible to cancer-causing changes. The simple sun protection factor (SPF) alone is not an adequate measure of effectiveness, because it is based on erythema, which is caused mainly by UVB. Some sunscreens do not adequately protect against UVA radiation. Buka (2004) reviews both sunscreens and insect repellents for children. He finds several products that adequately block both UVA and UVB. Look for a sunscreen labeled "broad spectrum" or with at least three stars in a UVA rating system. An adequate amount must be used: for children he recommends 1 fluid ounce (30 ml) *per application* of a product with SPF of 15 or more. The desired application of sunscreen is 2 mg cm^{-2}. Typical applications are about half this amount. It has been suggested that one make two applications (Teramura et al. 2012) or use a sunscreen with a very high SPF (Hao et al. 2012).

Because of the high reflectivity of sand and snow, beach umbrellas provide at most a factor of two protection. Hats need to have a brim that is at least 7.5 cm wide (Diffey and Cheeseman 1992). Automobile window glass provides protection against UVB; however untinted glass transmits enough UVA to present a significant exposure over several hours of driving (Kimlin and Parisi 1999).

14.10.6 Ultraviolet Light Damages the Eye

The effect of ultraviolet light on the eye has been reviewed by Bergmanson and Söderberg (1995). Acute effects include *keratitis* (inflammation of the cornea, the transparent portion of the eyeball) and *conjunctivitis* (inflammation of the conjunctiva, the mucous membrane covering the eye), also known as snow blindness or welder's flash. Laboratory studies show that ultraviolet-light exposure causes thickening of the cornea and disrupts corneal metabolism. UVC radiation is absorbed by the cornea. The lens absorbs UVB and, in older persons, UVA and visible light. Only a little UVA light reaches the retina. The retina is also susceptible to trauma from blue light. Low doses cause photochemical changes in tissues, while high doses also cause thermal damage.

Chronic low exposure to ultraviolet light causes permanent damage to the cornea, known as *droplet keratopathy* or *spheroid degeneration*. UVA radiation is a significant factor in the development of a *pterygium*, a hyperplasia of the conjunctiva that may grow over the cornea and impair sight. Rarely, it causes blindness.

Properly designed spectacles and contact lenses can protect the eye against ultraviolet light (Giasson et al. 2005). However, both must be designed to absorb ultraviolet. Soft contacts are larger and provide more protection than rigid gas-permeable contacts. Protection from high-intensity ultraviolet light requires sunglasses or welding goggles. Wide-brimmed hats also help protect the eye from ultraviolet light.

14.10.7 Ultraviolet Light Therapy

Ultraviolet light is used in therapy, primarily for the treatment of a skin disease called *psoriasis*, an inflammatory disorder in which the basal cells move out to the stratum corneum in much less than the normal 28 days. The skin

is red and has thick scaling. UVB radiation, often in conjunction with coal tar applied to the skin, has been used as a treatment for psoriasis since the 1920s. In the 1960s a treatment was developed that uses UVA and a chemical either applied to the skin or administered systemically (photochemotherapy or PUVA—psoralen UVA). The chemical is a psoralen derivative. It affects DNA, and when the affected DNA is irradiated with ultraviolet light, cross-links form, preventing replication. There are well-defined guidelines for the use of PUVA to treat psoriasis (Stern 2007). PUVA therapy is also useful in cutaneous T-cell lymphoma, a disease that first becomes apparent on the skin and then moves to internal organs.

Another treatment, *extracorporeal photopheresis*, involves removing the patient's blood, extracting the red blood cells, irradiating the plasma and white blood cells with UVA light outside the body, and returning the red blood cells and the irradiated white blood cells and plasma to the patient (Grossweiner 1994, pp. 167f; Knobler et al. 2009).

14.11 Heating Tissue with Light

Sometimes tissue is irradiated in order to heat it; in other cases tissue heating is an undesired side effect of irradiation. In either case, we need to understand how the temperature changes result from the irradiation. Examples of intentional heating are *hyperthermia* (heating of tissue as part of cancer therapy) or *laser surgery* (tissue ablation[13]). Tissue is ablated when sufficient energy is deposited to vaporize the tissue. Heating may be a side effect of phototherapy.

The temperature changes are often modeled by a heat-flow equation containing a source term for the deposition of photon energy and a term representing flow of energy away from the site in warmed blood. This is one form of the *bioheat equation*, which can include additional terms in more complicated models.

The linear equation for heat conduction was mentioned as one form of the transport equation in Table 4.3:

$$j_H = -K\frac{dT}{dx},$$

with the units of the thermal conductivity K being J K^{-1} m^{-1} s^{-1}. When extended to three dimensions and combined with the equation of continuity (conservation of energy), this gives a heat-conduction equation with the same form as Fick's second law for diffusion:

$$\rho_t c_t \frac{\partial T}{\partial t} = K\nabla^2 T. \tag{14.43}$$

[13] In surgery, *ablation* means the excision or amputation of tissue.

Here ρ_t is the density of the tissue (kg m^{-3}) and c_t the tissue specific heat capacity (J K^{-1} kg^{-1}). The left-hand side of the equation is the rate of energy increase in the tissue per unit volume, and the right-hand side is the net rate of heat flow into that volume by conduction—energy flowing because warmer molecules with more kinetic energy transfer energy to cooler neighbors in a collision process analogous to a random walk. This model is for solids; in liquid one must also consider convection.

We now add a term for energy carried away by flowing blood. In the linear approximation it is proportional to the temperature difference between the tissue and the blood supply and also to the rate of blood flow. Units for this term can be quite confusing and need to be examined in detail. Blood flow is usually defined by physiologists as the *perfusion* P, which is the volume flow of blood per unit mass of tissue. The SI units for P are

$$P\,\frac{m^3\ \text{(blood)}}{\left[\text{kg (tissue)}\right]\ \text{s}}.$$

Its product with the tissue density is the volume flow of blood per unit volume of tissue:

$$\rho_t P = \frac{\left[\text{kg (tissue)}\right]\left[\text{m}^3\ \text{(blood)}\right]}{\left[\text{m}^3\ \text{(tissue)}\right]\left[\text{kg (tissue)}\right]\ \text{s}} = \frac{\text{m}^3\text{(blood)}}{\text{m}^3\text{(tissue)}\ \text{s}} = \text{s}^{-1}.$$

The quantity is analogous to clearance (Chap. 2). Its inverse is the time it takes for a volume of blood equal to the tissue volume to flow through the tissue. Each term of our heat-flow equation has units of energy per unit volume of tissue per second. If we assume that the blood enters the tissue at temperature T_0 and leaves at temperature T, the energy lost by the volume is the heat capacity of blood, c_b, times its mass per unit volume times the temperature rise. The new term in the heat-flow equation is

$$c_b\frac{\text{J}}{\text{K kg (blood)}} \times \rho_b\frac{\text{kg (blood)}}{\text{m}^3\ \text{(blood)}} \times \rho_t P\frac{\text{m}^3\ \text{(blood)}}{\text{m}^3\ \text{(tissue)}\ \text{s}} \times [(T - T_0)\ \text{K}]$$

or

$$c_b\rho_b\rho_t P(T - T_0)\frac{\text{J}}{\text{m}^3\ \text{(tissue)}\ \text{s}},$$

so the heat-flow equation with blood flow added is

$$\rho_t c_t \frac{\partial T}{\partial t} = K\nabla^2 T - c_b\rho_b\rho_t P(T - T_0).$$

The last term we consider is the energy deposited by the photon beam. In Sect. 14.6 we defined the particle fluence and particle fluence rate for photons. The definition can be used for both collimated beams and diffuse radiation. In a similar way we define the *energy fluence* Ψ as the ratio of the

expectation value of the amount of photon energy traversing a small sphere of radius a divided by the area of a great circle of the sphere, πa^2. The *energy fluence rate* is

$$\psi = \frac{d\Psi}{dt}. \tag{14.44}$$

The energy per unit volume lost by a beam with energy fluence rate ψ can be determined by the following argument. Consider only the fluence rate due to photons traveling in a certain direction. Orient the z axis in that direction and consider a small volume $dSdz$. The rate at which energy flows into the volume is ψdS, and the rate at which it is absorbed is $\psi dS\mu_a dz$. Therefore, the rate of absorption per unit volume is $\mu_a\psi$, independent of the direction the photons travel. The final heat-flow equation is

$$\rho_t c_t \frac{\partial T}{\partial t} = K\nabla^2 T - c_b\rho_b\rho_t P(T - T_0) + \mu_a\psi. \tag{14.45}$$

For monoenergetic photons, the photon energy fluence rate is related to the photon fluence rate by

$$\psi = h\nu\varphi. \tag{14.46}$$

In general, one must first solve Eq. 14.27 to determine ψ and then solve Eq. 14.45. We could add other terms, such as one for the thermal energy produced by metabolism within the tissue.

Sometimes Eq. 14.45 is written with all terms divided by $\rho_t c_t$, and sometimes with all terms divided by K. If we divide by $\rho_t c_t$, the equation is similar in form to the diffusion equation in Chap. 4:

$$\frac{\partial T}{\partial t} = D\nabla^2 T - \frac{c_b}{c_t}\rho_b P(T - T_0) + \frac{\mu_a}{\rho_t c_t}\psi, \tag{14.47}$$

where

$$D = \frac{K}{\rho_t c_t}. \tag{14.48}$$

Values of D are in the range $(0.5\text{–}2.5)\times 10^{-7}$ m^2 s^{-1} depending on the tissue type (Grossweiner 1994, pp. 127–129). We saw in Chap. 4 that for a spreading Gaussian solution to the diffusion equation the variance is $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = 2Dt$. The thermal relaxation time, that is, the average time for the temperature rise to spread a distance x, is therefore $x^2/2D$ in one dimension, $x^2/4D$ in two dimensions, and $x^2/6D$ in three dimensions.

There is an interplay between the thermal conductivity term and the blood-flow term. The *thermal penetration depth* δ_{th} is the distance at which the two terms are comparable. For larger distances blood flow is more important. To estimate the penetration depth, assume that $T - T_0$ changes over this distance. Then the Laplacian is approximated by $\nabla^2 T \approx (T - T_0)/\delta_{\text{th}}^2$. Equating the diffusive and blood flow terms gives

$$D\frac{T - T_0}{\delta_{\text{th}}^2} = \frac{c_b}{c_t}\rho_b P(T - T_0)$$

so

$$\delta_{\text{th}}^2 = D\frac{c_t}{c_b}\frac{1}{\rho_b P} = \frac{K}{\rho_t c_b \rho_b P}. \tag{14.49}$$

Grossweiner (1994) discusses values for the various tissue parameters, their temperature dependence, and simple models for tissue heating and ablation.

14.12 Radiometry and Photometry

This section develops some of the concepts and vocabulary of *radiometry*, the measurement of radiant energy. We will be considering five types of radiant energy in the remaining chapters: infrared radiation, visible light, ultraviolet radiation, x rays, and charged particles. Concepts for the measurement of radiant energy were developed simultaneously in different disciplines and even in different wavelength regions, depending on the purpose and the measurement techniques that were originally available.

It is recommended that the term *photometry* be reserved for measurement of the ability of electromagnetic radiation to produce a human visual sensation, that *radiometry* be used to describe the measurement of radiant energy independent of its effect on a particular detector, and that *actinometry* be used to denote the measurement of photon flux or photon dose (total number of photons) independent of any subsequent photoactivated process (Zalewski 2009, p. 34.10). This section reviews radiometric units and introduces a few of the related units from photometry and actinometry. Nomenclature is slightly different for x rays and charged particles.

Section 14.6 described two quantities, the photon fluence and the photon fluence rate. The energy fluence and energy fluence rate were introduced in Sect. 14.11. These are reviewed and compared here so that all the definitions are in one place. The definitions are summarized in Table 14.6. Symbols are shown for quantities used in this text. The third column shows symbols that have been recommended by the American Association of Physicists in Medicine (AAPM 57 1996). They often differ from the usage in this book.

Table 14.6 A comparison of radiometric, photometric, and actinometric quantities. Symbols are given for those quantities used in this text. The column "Symbol sometimes used" gives an alternate symbol that is often found. See, for example, *AAPM57*(1996)

Radiometric quantity	Symbol used here	Symbol sometimes used	Units	Photometric quantity	Symbol	Units	Actinometric quantity	Symbol	Units
General quantities									
Radiant energy emitted, transferred, or received	R	Q	J	*Luminous energy*	R_v	lm s	*Number of photons* emitted, transferred, or received	N	
Radiant flux: radiant power emitted, transferred, or received	P	P or Φ or $\dot{R}$	W	*Luminous flux*	P_v	lm	*Photon flux*		s^{-1}
Radiance: radiant power per unit solid angle per unit area of surface projected perpendicular to the radiant energy. It can be defined on the surface of a source or detector or at any point on the path of a ray of radiation	L	r	$W\,m^{-2}\,sr^{-1}$	*Luminance*	L_v	candela m^{-2} (cd m^{-2})	*Photon flux radiance*		m^{-2} sr^{-1}
Energy fluence: ratio of the expectation value of the radiant energy striking a small sphere to the area of a great circle of the sphere	Ψ	H_0	$J\,m^{-2}$				*Photon fluence:* ratio of expectation value of the number of photons striking a small sphere to the area of a great circle of the sphere	Φ	m^{-2}
Energy fluence rate: energy fluence per unit time	ψ	E_0	$W\,m^{-2}$				*Photon fluence rate:* photon fluence per unit time	ϕ	m^{-2} s^{-1}
Quantities emitted from a surface									
Radiant intensity: radiant power or flux emitted by a point source in a given direction per unit solid angle		I	$W\,sr^{-1}$	*Luminous intensity*		lm sr^{-1} or candela (cd)	*Photon flux intensity.*		sr^{-1}
Exitance: radiant power or flux emitted or reflected per unit area		W_r	$W\,m^{-2}$	*Luminous exitance*		lm m^{-2}	*Photon exitance*		m^{-2}
Quantities incident on a surface									
Irradiance: power per unit area incident on a surface	E	E	$W\,m^{-2}$	*Illuminance*		lm m^{-2} or lux	*Photon flux irradiance*		m^{-2} s^{-1}
Radiant exposure: radiant energy arriving per unit area		H	$J\,m^{-2}$	*Luminous exposure*		lm s m^{-2}	*Photon flux exposure*		m^{-2}

14.12.1 Radiometric Definitions

14.12.1.1 Radiant Energy and Power

The total amount of energy being considered is the *radiant energy* R, measured in joules. It can be the energy emitted by a source, transferred from one region to another, or received by a detector. We use subscripts s and d to refer to the source and detector. In optics the radiant energy is electromagnetic radiation. In radiological physics we will also consider energy transported by charged particles such as electrons and protons, and by neutral particles such as photons and neutrons.

The rate at which the energy is radiated, transferred, or received is the *radiant power* P (watts).

14.12.1.2 Point Source: Radiant Intensity

The simplest source is a point that radiates uniformly in all directions. The *radiant intensity* or radiant power per unit solid angle (Appendix A) leaving a point source radiating uniformly in all directions is

$$\frac{dP}{d\Omega} = \frac{P}{4\pi} \quad (\mathrm{W\ sr^{-1}}). \tag{14.50}$$

The power per unit area falls as $1/r^2$, while the power per unit solid angle is independent of r.[14] A point source need not radiate uniformly in all directions. For example, a searchlight 1 m in diameter viewed from a point several kilometers away appears to be a point. The light might be confined to a cone with a half-angle of 1°. Then a plot of $dP/d\Omega$ might look like Fig. 14.31. The total power radiated by the point source is

$$P = \int \frac{dP}{d\Omega}\, d\Omega. \tag{14.51}$$

If the power per unit solid angle is symmetric about the axis of the beam and θ is the angle with respect to the beam axis, then (see Appendix L)

$$P = \int_0^{\pi} \frac{dP}{d\Omega} 2\pi \sin\theta\, d\theta.$$

14.12.1.3 Extended Source: Radiance

The radiant energy leaving a source can travel in many different directions. The radiation striking a surface can come from many different directions. If we consider any small area in space there will generally be radiation passing through that area traveling in many different directions. In each case, the radiant energy or the radiant power is proportional to the magnitude of the small area projected perpendicular to the direction the energy is traveling, and to the size of the solid angle—the range of directions—being considered.

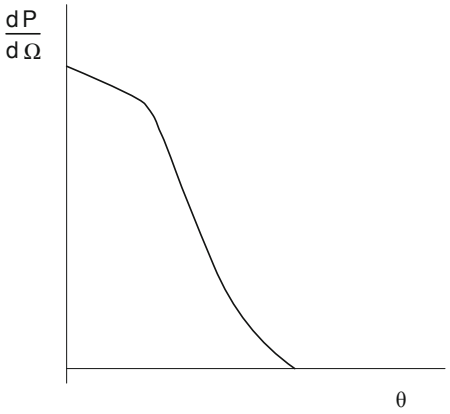


Fig. 14.31 A plot of power per unit solid angle as a function of angle from the axis of a hypothetical searchlight

The *radiance* L is the amount of radiant power per unit solid angle per unit surface area projected perpendicular to the direction of the radiant energy. The radiance of radiation traveling through a small area in space is sometimes difficult to visualize. Figures 14.32 and 14.33 may help. Figure 14.32 shows radiation leaving three points on a surface at the left. Some of it passes through the surface represented by the vertical line on the right. The energy passing through that surface has components from each point on the radiating surface. Figure 14.33 shows radiation in a very narrow cone of solid angles passing through surface dS whose normal is at an angle θ with the beam direction. The radiance is the power per unit solid angle divided by $dS\cos\theta$.

We have already seen the *energy fluence* Ψ, which is a measure of the total radiation entering or leaving a small volume of space. It is the total amount of energy striking a small sphere of radius a divided by the area of a great circle πa^2 in the limit as the radius approaches zero. Strictly speaking, if we repeat the experiment many times, the amount of energy striking the sphere fluctuates. The energy fluence is defined in terms of the expectation value of this fluctuating quantity. Figure 14.12 shows two examples. In Fig. 14.12a, a parallel beam with energy R passes through a circular area πa^2 for a time Δt. In Fig. 14.12b, a total amount of energy R strikes a sphere of radius a from many different directions. In both cases, $\Psi = R/\pi a^2$. Notice that some of the energy passing through the sphere passes outside a great circle that is not perpendicular to the direction in which the radiation is traveling, but it does pass through a great circle constructed perpendicular to its direction of travel.

[14] The lighting industry calls $dP/d\Omega$ the intensity, while in physical optics intensity is used for power per unit area. We will try to avoid using the word intensity alone.

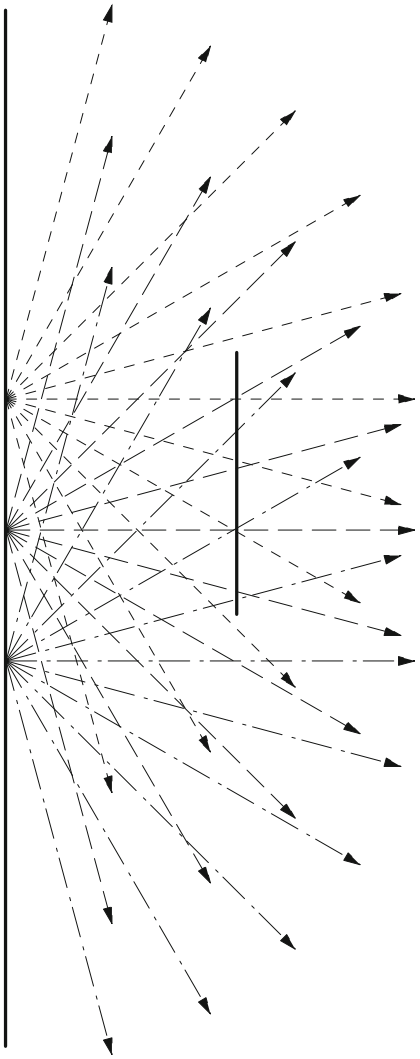


Fig. 14.32 Radiation emitted from different points of the surface on the left strikes the surface on the right

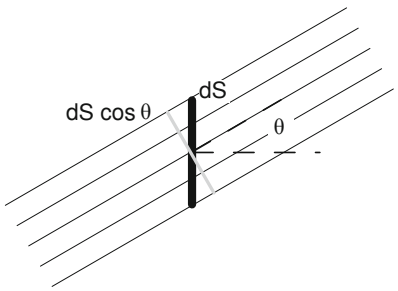


Fig. 14.33 Surface area dS, projected perpendicular to the direction of the radiation, has projected area $dS\cos\theta$

The *energy fluence rate* is the amount of energy fluence per unit time (which for the small sphere is $P/\pi a^2$):

$$\psi = \frac{d\Psi}{dt}. \tag{14.52}$$

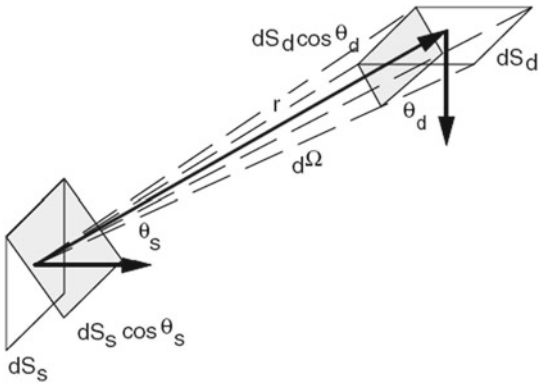


Fig. 14.34 Radiant energy is emitted from a source of surface area dS_s into a cone of solid angle $d\Omega$. The direction of emission is at an angle θ_s with the normal to the surface. A detecting surface has an element of area dS_d oriented at a direction θ_d to the direction of travel of the radiation from source to detector. The shaded rectangles show the projections of dS_s and dS_d perpendicular to the line of length r from source to detector

The *exitance* W_r is the radiant power or flux emitted per unit area of a surface.

14.12.1.4 Energy Striking a Surface: Irradiance

Now consider the energy striking a surface. The *irradiance* E is the power per unit area incident on a surface. The strict definition is the ratio of the power incident on an infinitesimal element of detector surface dS_d to the area projected perpendicular to the direction the radiant energy is traveling. If θ_d is the angle between a normal to the surface and the direction of propagation, the irradiance is

$$E = \frac{1}{\cos\theta_d}\frac{dP}{dS_d}. \tag{14.53}$$

For a point source radiating uniformly in all directions, the power at distance r is spread uniformly over a sphere of area $4\pi r^2$, so the irradiance on a detecting surface perpendicular to a line back to the source is

$$E = \frac{P}{4\pi r^2} \qquad \text{(isotropic point source)}. \tag{14.54}$$

For an extended source the power emitted by the surface is proportional to both the size of the emitting area dS_s and the solid angle of the cone $d\Omega$ into which the energy is radiated, as shown in Fig. 14.34. The solid angle subtended by a small element of area on the detector is $d\Omega$, as shown by the dashed lines. The amount of power radiated into $d\Omega$ from dS_s is

$$L dS_s d\Omega = \frac{1}{\cos\theta_s}\frac{d^2P}{dS_s\, d\Omega} dS_s\, d\Omega, \tag{14.55}$$

where the radiance L depends on the direction of emission as well as the location on the surface. This equation is valid

whether the energy is emitted directly from the source (as in a glowing object) or is scattered by the surface (as from this page). The total power emitted is

$$P = \int\int L\, dS_s\, d\Omega. \tag{14.56}$$

The distinction between angles and areas for the source and the detector is shown in Fig. 14.34. Note that the solid angle subtended at the source by dS_d is $d\Omega = dS_d \cos\theta_d / r^2$. The power into an area dS_d of the detector from area dS_s of the source is therefore

$$d^2P = \frac{L \cos\theta_s \cos\theta_d\, dS_s\, dS_d}{r^2}. \tag{14.57}$$

14.12.1.5 Plane-Wave Relationships

We can derive some useful relationships for a beam of collimated radiation all traveling in one direction (a *plane wave*). Imagine that the collimated beam comes from a distant point source radiating power P. The energy fluence rate at distance r from the source is the power through a sphere of radius a divided by πa^2:

$$\psi = \frac{\pi a^2 P}{4\pi r^2}\frac{1}{\pi a^2} = \frac{P}{4\pi r^2}.$$

This is also the power per unit area incident on a circle of radius a oriented perpendicular to the beam. Therefore, for a collimated beam,

$$\psi = E \quad \text{(collimated beam)}. \tag{14.58}$$

14.12.1.6 Isotropic Radiation: Lambert's Law

In general, L may depend on the angle of emission. In some cases, such as reflection from a "perfectly diffuse" surface, the radiation is isotropic : $L = L_0$. This is called *Lambert's law of illumination* or Lambert's cosine law.[15]

A surface described by Lambert's law will have equal power per unit area in the image regardless of the viewing angle. Look at surfaces around you. Do similar surfaces illuminated the same way appear to have the same brightness when they are oblique to your line of vision?

The power incident on a small element of surface area dS_d from angle $d\Omega$ is $L_0 dS_d \cos\theta_d\, d\Omega$, where θ_d is the angle that the incident radiation makes with the normal to the surface. The solid angle is $2\pi \sin\theta_d\, d\theta_d$ (see Fig. 14.11b). The irradiance is

$$E = \frac{dS_d 2\pi L_0 \int_0^{\pi/2} \cos\theta_d \sin\theta_d d\theta_d}{dS_d} = \pi L_0. \tag{14.59}$$

[15] Sometimes Eq. 14.57 is defined without the factor $\cos\theta_s$, in which case Lambert's law has the form $L(\theta_s) = L_0 \cos\theta_s$.

The same geometry is used with dS_s to show that for isotropic radiation, the exitance is

$$W_r = \pi L_0. \tag{14.60}$$

To determine the energy fluence rate for isotropic radiation consider a small sphere of radius a and the radiation arriving in a small solid angle $d\Omega$ about a line perpendicular to a great circle of the sphere. The power is $L_0 \pi a^2 d\Omega$. This argument applies for any direction of the radiation. Integrating over all directions gives the total power $L_0 \pi a^2 4\pi$. Therefore, for isotropic (Lambertian) radiation,

$$\psi = 4\pi L_0 = 4E \quad \text{(isotropic radiation)}. \tag{14.61}$$

14.12.1.7 The Spectrum

When the energy is not monochromatic, we define the amount of energy per unit wavelength interval as R_λ, with units J m^{-1} or J nm^{-1}. The total energy between wavelengths λ_1 and λ_2 is

$$\int_{\lambda_1}^{\lambda_2} R_\lambda(\lambda)\, d\lambda \tag{14.62a}$$

and between frequencies ν_1 and ν_2 it is

$$\int_{\nu_1}^{\nu_2} R_\nu(\nu) d\nu. \tag{14.62b}$$

The relationship between R_λ and R_ν is the same as in Eqs. 14.36 and 14.37.

14.12.2 Photometric Definitions

For the photometric units we also need to know the sensitivity of the eye. The eye contains two types of light receptors: *rods*, which have no color discrimination but are most sensitive, and *cones*, which are less sensitive and can discriminate color. *Photopic* vision is normal vision at high levels of illumination in which the eye can distinguish colors. *Scotopic* vision occurs at low light levels with a dark-adapted eye. The CIE has established the spectral efficiency function V for the eye of a standard observer for both photopic vision $[V(\lambda)]$ and scotopic vision $[V'(\lambda)]$. Both are normalized to unity at their peak (Fig. 14.35).

The *luminous flux* P_v in lumens (lm) is the analog of the energy flux P. The peak sensitivity for photopic vision is for green light, $\lambda = 555$ nm. At that wavelength the relationship between P and P_v is

$$\begin{aligned} P = 1\,\text{W} &\Longleftrightarrow P_v = 683\,\text{lm}, \\ P_v = 1\,\text{lm} &\Longleftrightarrow P = 1.464 \times 10^{-3}\,\text{W}. \end{aligned} \tag{14.63a}$$

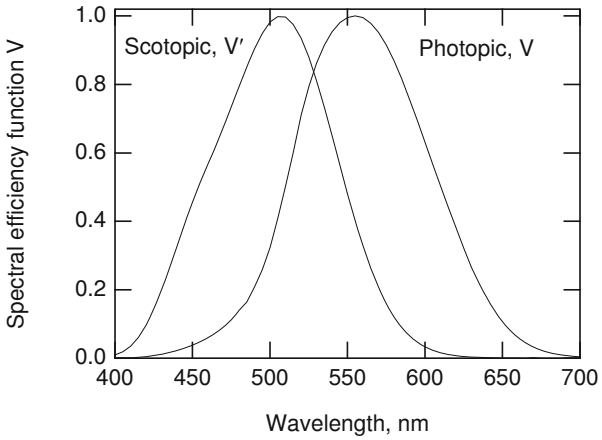


Fig. 14.35 The spectral efficiency functions for the CIE standard eye. Plotted from data in Table 2 of Zalewski (2009)

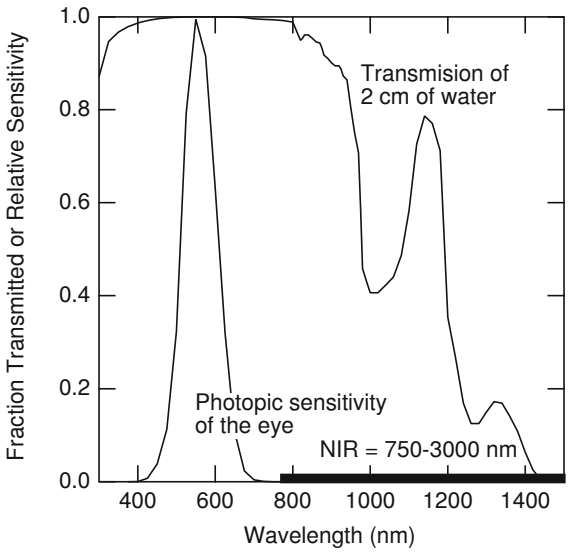


Fig. 14.36 Transmission of light through 2 cm of water, compared to the spectral efficiency of the eye

The ratio P_v/P at 555 nm is the *luminous efficacy* for photopic vision, $K_m = 683$ lm W^{-1}. For a distribution of wavelengths,

$$P_v(\text{photopic}) = K_m \int_{400\text{ nm}}^{700\text{ nm}} V(\lambda) P_\lambda(\lambda)\, d\lambda. \qquad (14.63b)$$

An analogous relationship holds for scotopic vision, with $K'_m \approx 1700$ lm W^{-1}:

$$P_v(\text{scotopic}) = K'_m \int_{400\text{ nm}}^{700\text{ nm}} V'(\lambda) P_\lambda(\lambda)\, d\lambda. \qquad (14.63c)$$

If P were spread uniformly over the visible spectrum, the overall conversion efficiency would be about 200 lm W^{-1}. A typical incandescent lamp has an efficiency of 10–20 lm W^{-1}, while a fluorescent lamp has an efficiency of 60–80 lm W^{-1}. A typical LED replacment lamp is about 75 lm W^{-1}. The number of lumens per steradian is the *luminous intensity*, in lm sr^{-1}. The lumen per steradian is also called the *candela*. Other units are shown in Table 14.6.

The peak of the eye's spectral efficiency function is at about the peak of the sun's blackbody spectrum when plotted as a function of wavelength (Eq. 14.33). Some authors have speculated that this is because we evolved in sunlight. There is a severe problem with this argument. The spectral efficiency function has the same value whether we consider a particular wavelength or its corresponding frequency. The blackbody spectrum is a distribution function—per wavelength interval (Eq. 14.33) or per frequency interval (Eq. 14.38).[16] The sun's blackbody spectrum plotted versus frequency peaks at a frequency corresponding to a wavelength of 880 nm, far from the peak of the spectral efficiency function (See Fig. 14.24). Soffer and Lynch (1999) have discussed this at length and describe several of the errors in the literature. The structures in the human eye, as in all vertebrate eyes, are mostly water. All vertebrate eyes are sensitive between 390 and 760 μm, with a peak at 500–550 μm. It is interesting to compare the spectral efficiency function with the transmission of light through 2 cm of water (Fig. 14.36). The eye's response is pretty well centered in this absorption window. Many insects, crustaceans, fish, birds, and reptiles have ultraviolet-sensitive receptors (Kevan et al. 2001).

14.12.3 Actinometric Definitions

The actinometric quantities count the number of photons. For monochromatic photons the energy is the number of photons times $h\nu$. Therefore an actinometric quantity is easily obtained when the radiometric quantity is known. The units are shown in Table 14.6.

14.13 The Eye

This section presents a simple model for the eye, sufficient for us to understand how refractive errors are corrected and to see how photons strike the retina, so that the sensitivity of the eye can be determined in the next section. For a detailed but nonmathematical introduction to the eye and vision, see Rodieck (1998).

[16] Other distribution functions are also useful, for example, per logarithmic frequency or wavelength interval. See Soffer and Lynch (1999) or Heald (2003).

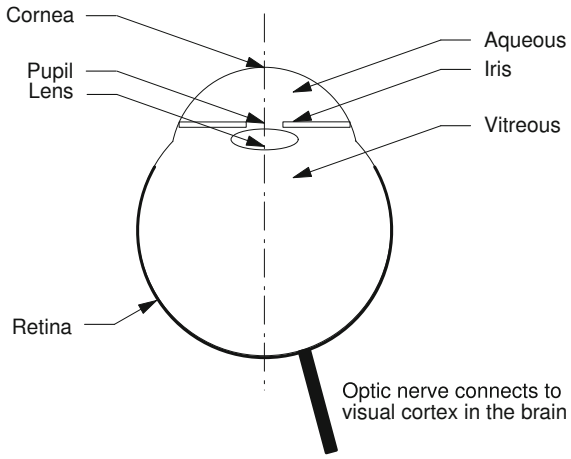
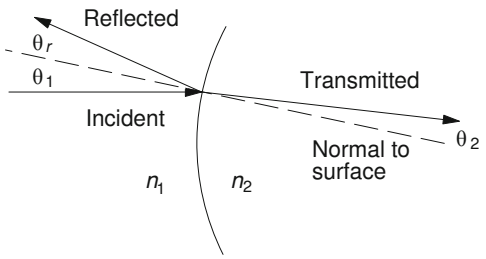


Fig. 14.37 A simplified cross section of the left eye, viewed from above



$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

$$\theta_1 = \theta_r$$

Fig. 14.38 Light passing from one medium to another with a different index of refraction. All angles are measured with respect to the normal to the surface

A simplified cross section of the eye is shown in Fig. 14.37. The principal components through which the light passes are the curved, thin, transparent *cornea*, the *aqueous*, the *lens*, the *vitreous*, and the *retina*. The *iris* defines the area of the *pupil*, the opening in front of the lens through which light passes.

When light passes through a surface from one medium to another, part is reflected and part is transmitted. The transmitted light usually changes direction, a process called *refraction*. Figure 14.38 shows the angles involved, all measured with respect to the dashed line, which is normal to the surface at the point where the light ray strikes. The angle the reflected light makes with the normal is the same as the angle of incidence, $\theta_r = \theta_1$. The direction the refracted light travels is described by *Snell's law*, $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

When light from an object strikes the eye, it must be refracted to form an image on the retina. Most of the refraction takes place at the surface between the air and the cornea.

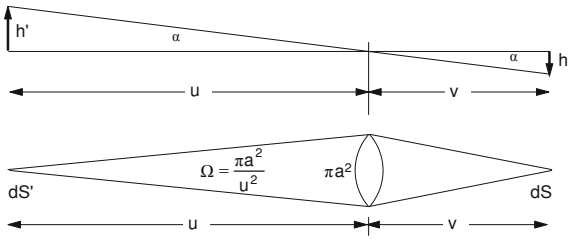


Fig. 14.39 A source of height h' emits light in all directions. Some of this light is intercepted by a lens and focused in an image. **a** Relation between object and image distances and sizes. **b** Collection of light by the lens

The cornea is very thin, and a light ray is deflected only a very small distance before it strikes the aqueous. Thus, most of the refraction occurs because of the difference between the index of refraction of the air ($n = 1.00$) and the aqueous ($n = 1.33$). The light then passes through the lens ($n = 1.42$) and the vitreous ($n = 1.33$). The lens changes shape to provide the adjustable part of the overall refraction.

A number of models at varying levels of sophistication are used to describe the formation of the image on the retina. The most detailed take into account the refraction at each surface where the index of refraction changes, including variations in different layers of the lens itself. Others treat only the refraction at the air–cornea, aqueous–lens, and lens–vitreous interfaces. The simplest model, and the one we will use, treats the eye as a thin lens of adjustable focal length f, with object distance u and fixed image distance v, as shown in Fig. 14.39. The object and image distances and focal length are related by the *thin-lens equation*:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}. \tag{14.64}$$

When the object is infinitely far away the image distance is equal to the focal length of the lens, $v = f$. A typical value for v is 1.7 cm. As the object is brought closer to the eye v cannot change, but the lens changes to decrease the focal length.

In ophthalmology and optometry it is customary to describe the refraction of the eye in terms of the *vergence*. When light rays are emanating from a point they are diverging, and the vergence is negative. When they are coming toward a point the vergence is positive and they are converging. When they are parallel, the vergence is zero. Quantitatively, the vergences for the geometry shown in Fig. 14.39 are

$$U = -\frac{1}{u} \quad \text{(diverging from the object)},$$

$$V = \frac{1}{v} \quad \text{(converging to the image)}, \tag{14.65}$$

$$F = \frac{1}{f} \quad \text{(a converging lens)}.$$

Table 14.7 Convergence power of the eye in diopters

Refracting structure	Relaxed normal eye	Most converging eye (age 25)
Air-cornea surface	45	45
Lens	14	24
Entire eye	59	69

The relationship between the vergences is

$$V = F + U. \tag{14.66}$$

When the distances are in meters, the vergences are in *diopters*.

A given eye requires a particular value of V to form the image on the retina. The converging power of all the refracting surfaces in the eye must be $F = V$ in order to focus on an object infinitely far away. Closer objects require more convergence from the eye, which is provided by the lens. Table 14.7 shows typical values for the converging power of the eye. Most of the convergence is provided by the front surface. When the eye is relaxed, $F = V = 59$ diopters, $U = 0$, and the eye is focused on an object infinitely far away. With $F = 69$ diopters, $U = 10$, and the eye is focused on an object 0.1 m away. This ability of the lens to change shape and provide additional converging power is called *accommodation*.

In the normal or *emmetropic* eye, the length of the eye is such that when the lens is relaxed, rays with no vergence (parallel rays from a source infinitely far away) are focused on the retina ($V = F$).

In farsightedness or *hyperopia*, parallel rays come to a focus behind the retina. The relaxed eye does not have enough converging power ($F < V$). The subject can focus on distant objects by providing some additional converging power from the lens, but then the lens cannot provide enough converging power to focus on nearby objects. A corrective lens, either spectacles or a contact lens, provides additional convergence.

In nearsightedness or *myopia*, parallel rays come to a focus in front of the retina. The eye is slightly too long for the shape of the cornea ($F > V$). The total converging power of the eye is too great, and the relaxed eye focuses at some closer distance, from which the rays are diverging. Accommodation can only increase the converging power of the eye, not decrease it, so the unassisted myopic eye cannot focus on distant objects. Myopia can be corrected by placing a diverging spectacle or contact lens in front of the eye, so that incoming parallel rays are diverging when they strike the cornea.

When the eye is not symmetric about an axis through the center of the lens, the images from objects oriented at different angles in the plane perpendicular to the axis form at different distances from the lens. This is called *astigmatism*, and it can be corrected with a spectacle lens that is not symmetric about the axis. The lack of symmetry usually occurs at the surface of the cornea, so a contact lens can restore the symmetry.

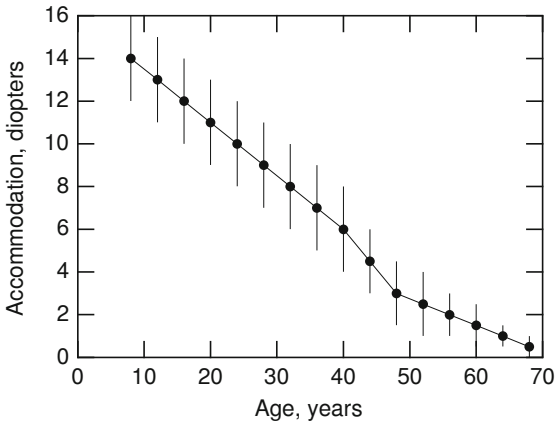


Fig. 14.40 Accommodation versus age. There is considerable variation between individuals, shown by the *error bars*

Surgery to change the radius of curvature of the cornea can also be used to correct errors of refraction.

As we age, the accommodation of the eye decreases, as shown in Fig. 14.40. A normal viewing distance of 25 cm or less requires 4 diopters or more of accommodation. The graph shows that this limit is usually reached in the early 40s. To make up for the lack of accommodation, one can place a converging lens in front of the eye when viewing nearby objects (reading glasses). Bifocals provide a different amount of convergence at the top and bottom of the lens. This can be done either by grinding the lower portion of the lens with a different radius of curvature or by fusing glass with a different index of refraction into the lens.

The sharpness of the image is reduced by two other effects: *chromatic aberration* and *spherical aberration*. Chromatic aberration occurs because the index of refraction varies with wavelength. There is nearly a 2-diopter change in overall refractive power from the red to the blue. Spherical aberration occurs because the refractive power changes with distance from the axis of the eye. This is different from astigmatism, which is a departure from symmetry at different angles about the axis.

A concept important in both vision and photography is *depth of field*. The retina has a finite spatial resolution, so the image of a point still appears sharp, even if it is slightly out of focus. Consider Fig. 14.41. The retina is behind the plane in which the image is in focus. In dim light, the pupil of the eye is fully open and light from a point object is spread out over the larger circle on the retina defined by the solid rays. In brighter light the pupil is smaller, and light from the same point object is confined to the smaller circle defined by the

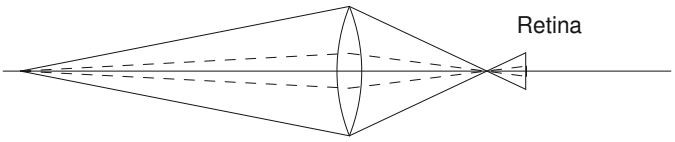


Fig. 14.41 Depth of field is illustrated by this ray diagram. The retina is slightly behind the plane of focus. In dim light, the pupil of the eye is fully open and light from a point object is spread out over the larger circle on the retina. When the light is brighter and the pupil is smaller, light from the same point object is confined to the smaller circle defined by the *dashed lines*

dashed lines. As long as this circle is smaller than the spatial resolution, the image is sharp. This is why we can see better in brighter light. An older person whose accommodation is less and who is trying to avoid bifocals often finds that bright light makes it easier to see nearby objects.

Point-spread functions and modulation transfer functions can be used to describe the image. (See, for example, Charman (2009) or Greivenkamp et al. (1995).) A simpler model describes the image by a Gaussian with a certain standard deviation, equal to the square root of the sum of the variances due to various effects. The maximum photopic (bright-light) resolution of the eye is limited by four effects: diffraction of the light passing through the circular aperture of the pupil (5–8 μm), spacing of the receptors (≈ 3 μm), chromatic and spherical aberrations (10–20 μm), and noise in eyeball aim (a few micrometers) (Stark and Theodoridis 1973). The total standard deviation is $(6^2+3^2+15^2+5^2)^{1/2} = 17$ μm in the image on the retina. Since the diameter of the eyeball is about 2 cm, this corresponds to an angular size (α in Fig. 14.39) of $(17 \times 10^{-6})/(2 \times 10^{-2}) = 8.5 \times 10^{-4}$ rad $= 0.048^\circ =$ 2.9 min of arc. (For further discussion, see Cornsweet (1970, Chap. 3).)

14.14 Quantum Effects in Dark-Adapted Vision

The visual process involves two steps. First, the eye creates an image of an external object on the retina as described above. Then the photon stimulus is transduced into neurological signals that are interpreted by the central nervous system. The discussion here is limited to a classic experiment on scotopic vision that shows the importance of quantum effects (shot noise) in human vision in dim light. For a more detailed discussion of how photoreceptors detect photons, see Rodieck (1998).

The experiment was performed by Hecht, Shlaer, and Pirenne in 1942. It has been described in many places. A detailed nonmathematical description is that by Cornsweet (1970).

The retina can be divided into two regions. The *fovea*, the area of greatest visual discrimination, is composed entirely

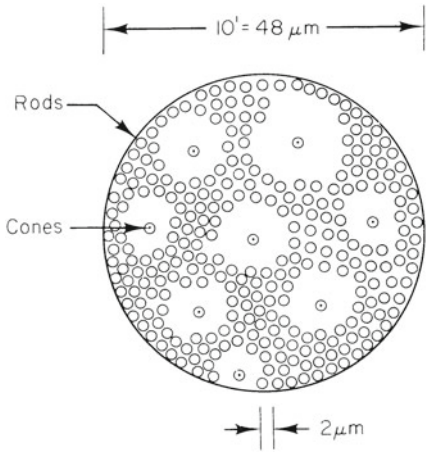


Fig. 14.42 An example of a 10-minute-of-arc field superimposed on the rods and cones in the retina in the region of greatest sensitivity

of cones. The percentage of rods is highest a few millimeters away from the fovea, and this part of the retina is most sensitive to faint light. The dark-adapted eye increases sensitivity by a factor of about 5000.

The experiment was done by having the subject look directly at a very dim red fixation point while a green light was flashed in such a place that its image fell on the most sensitive part of the retina. Experiments on the sensitivity of the dark-adapted eye to flashes of weak light have shown that if the flash duration is less than 100 ms and the light on the retina covers a *receptor field* less than 10 min of arc in size, the scotopic response of the eye depends on the total amount of energy or the total number of photons in the flash. Photons striking anywhere within the receptor field during this time have the same effect; the eye must combine the effects occurring in all receptors in the receptor field in a tenth of a second. A scotopic receptor field is shown in Fig. 14.42. This scotopic field size (10 min of arc) cannot be compared to the 2.9 min for maximum resolution, which is for photopic vision on a different part of the retina.

In the Hecht–Shlaer–Pirenne experiment, the flashes were short enough and small enough so that only the total number of photons was important. The fraction of flashes that the subject recognized was measured as a function of the total flash energy. A typical response curve is shown in Fig. 14.43. Let q be the number of photons striking the cornea in front of the pupil in each flash, which is the total energy in the flash divided by the energy of each photon. For the 510-nm green light used, the photon energy is $hc/\lambda = 3.89 \times 10^{-19}$ J. The number of photons striking the cornea can be determined as follows. Let Lt be the radiance times the duration of the flash. Consider Eq. 14.57 with both θ_s and θ_d nearly zero. Refer also to the lower half of Fig. 14.39. The energy striking

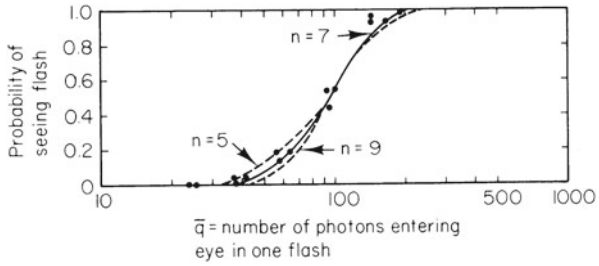


Fig. 14.43 Typical response in the experiments of Hecht, Shlaer, and Pirenne. Curves are calculated using Eq. 14.69. (Data are from Hecht et al. (1942))

the cornea over the pupil area is

$$\frac{(Lt)\,dS_s dS_d}{r^2} = \frac{(Lt)\,dS'\,(\pi a^2)}{u^2}.$$

Because $h = h'v/u$, the area on the retina where photons from dS' fall is $dS = dS'(v/u)^2$. The number of photons striking the cornea that would be in dS if there were no losses is

$$q = \frac{(Lt)(\pi a^2)dS'}{hvu^2} = \frac{(Lt)(\pi a^2)\,dS}{hvv^2}. \qquad (14.67)$$

The number of photons fluctuates from flash to flash. Therefore we should speak of $\overline{q}$, the average number of photons striking the cornea per flash. Of these, only some fraction f actually reach the retina and are absorbed by a visual pigment molecule. The average number absorbed is

$$m = f\overline{q}. \qquad (14.68)$$

Let us next postulate that some minimum number of quanta n must be absorbed during the flash in order for the subject to see it. If the average number absorbed per flash is m, there will sometimes be more and sometimes less than n photons absorbed per flash. The probability of absorbing x photons per flash is given by the Poisson distribution $P(x; m)$ (Appendix J). The probability of seeing the flash is the probability that x is greater than or equal to n :

$$P(\text{seeing}) = \sum_{x=n}^{\infty} P(x; m) = 1 - \sum_{x=0}^{n-1} P(x; m)$$
$$= 1 - e^{-m}\left(1 + m + \frac{m^2}{2!} + \cdots + \frac{m^{n-1}}{(n-1)!}\right). \qquad (14.69)$$

This function is plotted in Fig. 14.44 as a function of m for various values of n, with both a linear and a logarithmic scale for m.

Hecht, Shlaer, and Pirenne used an ingenious method to determine n. They plotted their data versus the logarithm of $\overline{q}$. Since $m = f\overline{q}$, $\log m = \log f + \log \overline{q}$; different values

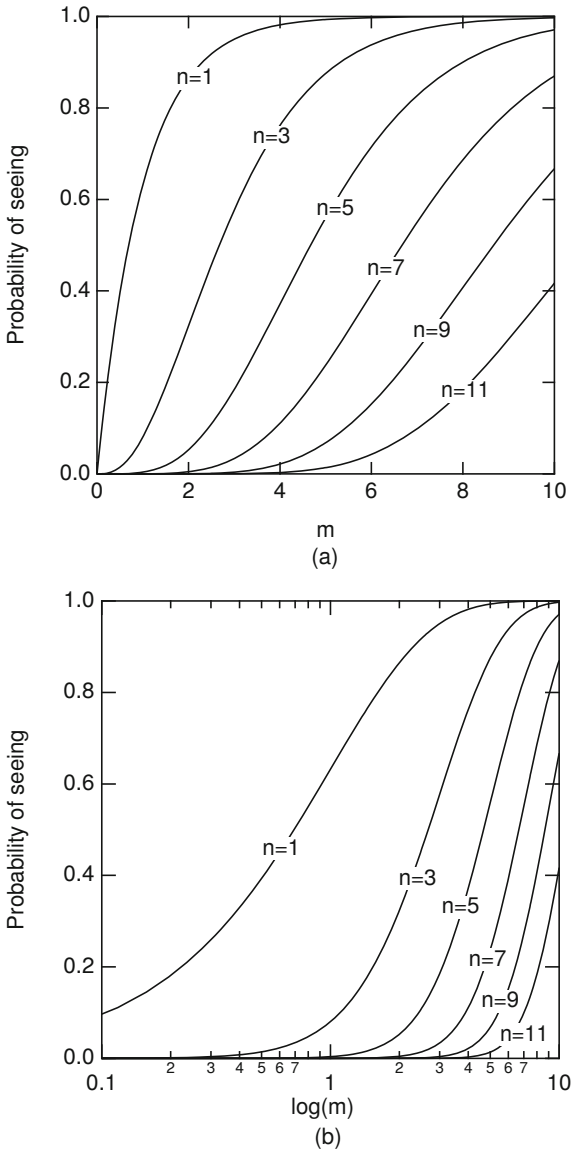


Fig. 14.44 The probability of seeing a flash, plotted versus **a** m; **b** $\log m$

of f correspond to shifting the curve along the axis. They then compared the experimental data to various theoretical curves for the probability of seeing a flash, plotted against $\log m$. Sliding the paper containing the data along the $\log m$ axis is equivalent to trying different values of f. The data in Fig. 14.43 are shown along with the curves for $n = 5$, 7, and 9. For these data, $n = 7$ gives the best fit. From Fig. 14.43, a 55% chance of detecting the flash corresponds to 100 photons for $\overline{q}$ while being consistent with $m = 7$. Therefore, $f = 0.07$.

Hecht, Shlaer, and Pirenne deduced that about seven photons must be absorbed by the rods in the area of integration shown in Fig. 14.42 within 0.1 s in order for the brain to

detect the flash of light. Their data were consistent with the hypothesis that the photons arrived at random, with the actual number in each flash obeying a Poisson distribution. Later work by Sakitt (1972) is consistent with the rods counting individual photons, with false positives produced by thermal noise within the retina (Barlow 1956).

The phototransduction mechanism is quite complicated. Rieke and Baylor (1998) reviewed the detection of photons by rod cells. When stimulated with dim light pulses, the rod cell responds to each flash consistent with the absorption of 0, 1 or, 2 photons. The rods have a dark current that is reduced when light falls on them. In other words, the light hyperpolarizes the cell. This lowers the rate of release of cyclic GMP. The review discusses what is known about the chemical transduction process.

If the light intensity is increased, m increases. There will be shot-noise fluctuations with a standard deviation equal to $m^{1/2}$, and the eye should be unable to detect brightness changes smaller than this. Measurements by Horace Barlow in 1956 showed that as long as short flashes spanning only one visual field are used, the minimum detectable intensity depends on the square root of the light intensity. This statistical limit to detecting intensity changes is a lower limit; for larger sources and longer exposure times, the minimum detectable brightness change is larger and is more nearly proportional to the intensity than to the square root of the intensity (Rose 1973).

14.15 Color Vision

The eye can detect color because there are three types of cones in the retina, each of which responds to a different wavelength of light (*trichromate vision*): red, green, and blue, the primary colors. However, the response curve for each type of cone is broad, and there is overlap between them (particularly the green and red cones). The eye responds to yellow light by activating both the red and green cones. Exactly the same response occurs if the eye sees a mixture of red and green light. Thus, we can say that red plus green equals yellow. Similarly, the color cyan corresponds to activation of both the green and blue cones, caused either by a monochromatic beam of cyan light or a mixture of green and blue light. The eye perceives the color magenta when the red and blue cones are activated but the green is not. Interestingly, no single wavelength of light can do this, so there is no such thing as a monochromatic beam of magenta light; it can only be produced my mixing red and blue. Mixing all three colors, red and green and blue, gives white light. Color printers are based on the colors yellow, cyan, and magenta, because when we view the printed page, we are looking at the reflection after some light has been absorbed by the ink. For instance, if white light is incident on a page containing ink that absorbs blue light, the reflected light will contain red and green and therefore appear yellow. Human vision is trichromate, but other animals (such as the dog) have only two types of cones (*dichromate vision*), and still others have more than three types.

Some people suffer from color blindness. The most common case is when the cones responding to green light are defective, so that red, yellow, and green light all activate only the red receptor. Such persons are said to be red–green color blind: they cannot distinguish red, yellow, and green, but they can distinguish red from blue.

As with pitch perception, the sensation of color involves both physics and physiology. For instance, one can stare at a blue screen until the cones responding to blue become fatigued, and then immediately stare at a white screen and see a yellow afterimage. Many other optical illusions with color are possible.

Symbols Used in Chapter 14

Symbol	Use	Units	First used page
a	Radius	m	390
c	Speed of light in a vacuum	$\mathrm{m\,s^{-1}}$	381
c_n	Speed of light in a medium	$\mathrm{m\,s^{-1}}$	381
c_b, c_t	Specific heat of blood, tissue	$\mathrm{J\,kg^{-1}\,K^{-1}}$	404
e	extinction coefficient	m^2	417
e	Charge on an electron	C	383
f	Focal length	m	411
f	Fraction of photons reaching retina		414
g	Scattering anisotropy factor		389
h	Planck's constant	J s	382
h, h'	Image height, object height	m	411
$\hbar$	Planck's constant divided by 2π	J s	382
i	Label of energy level		383
j	Total angular momentum quantum number		384
j_H	Energy transport in heat flow	$\mathrm{W\,m^{-2}}$	404
k	Spring constant	$\mathrm{N\,m^{-1}}$	386
k_B	Boltzmann constant	$\mathrm{J\,K^{-1}}$	395
l	Orbital angular momentum quantum number		383
m	Mass	kg	385
m	Average number		414
m_e	Mass of electron	kg	383
m_i	Mass of ith particle	kg	385
m_j, m_l, m_s	z quantum number for angular momentum		383
n	Index of refraction		381
n	Principal quantum number		383

$\overline{n}$	Average number of photons that interact		388
n	Minimum number of photons to trigger a response		414
p	Probability		389
q	Electric charge	C	382
q	Number of photons		414
$\overline{q}$	Average value of q		414
r	Rotational quantum number		385
r, $\mathbf{r}$	Coordinate	m	385
s	Spin quantum number		383
s	Source term in diffusion equation	$m^{-3}\,s^{-1}$	390
t	Time	s	382
$\mathbf{v}$	Velocity	$m\,s^{-1}$	382
v	Vibrational quantum number		386
u, v	Object and image distances	m	411
w_{tot}	Net power radiated	W	398
x, z	Distance	m	382
z_0	Depth of first scattering	m	391
A	Amplitude of wave		392
A	Molar mass	kg	388
$\mathbf{B}$	Magnetic field	T	382
C	Concentration	m^{-3}	390
D	Diffusion constant	$m^2\,s^{-1}$	390
D'	Photon diffusion constant	m	390
D	Thermal diffusion constant	$m^2\,s^{-1}$	405
$\mathbf{E}$	Electric field	$V\,m^{-1}$	382
E	Energy	J	382
E_p	Potential energy	J	386
E_r	Rotational energy	J	385
E_v	Vibrational energy	J	386
E	Irradiance	$W\,m^{-2}$	409
F, $\mathbf{F}$	Force	N	382
F	Converging power of a lens	diopter (m^{-1})	411
I	Moment of inertia	$kg\,m^2$	385
K	Thermal conductivity	$W\,K^{-1}\,m^{-1}$	404
K_m	Luminous efficiency, photopic	$lm\,W^{-1}$	410
K'_m	Luminous efficiency, scotopic	$lm\,W^{-1}$	410
L	Angular momentum	$kg\,m^2\,s^{-1}$	385
L	Radiance	$W\,m^{-2}\,sr^{-1}$	409
N	Number of photons		387
N_a	Number absorbed		387
N_s	Number scattered		387
N_A	Avogadro's number		388
N_T	Number of target entities per unit area	m^{-2}	388
P	Probability		395
P	Tissue perfusion	$m^3\,kg^{-1}\,s^{-1}$	404
P	Radiant power	W	407
P_v	Luminous flux	lm	410
Q	Rate of production	$m^{-3}\,s^{-1}$	390
R, $\mathbf{R}$	Coordinate of atom, distance	m	385
R_λ	Radiant energy per unit wavelength interval	$J\,m^{-1}$ or $J\,nm^{-1}$	409
R	Reflected fluence rate	$m^2\,s^{-1}$	391
R	Radiant energy	J	407
R	Rydberg constant	$m-1$	417
S, S'	Surface area	m^2	388
T	Period	s	382
T	Kinetic energy	J	385
T, T_s, T_o	Temperature	K	395
U	Object vergence	diopter (m^{-1})	411
V, $\mathbf{V}$	Velocity	$m\,s^{-1}$	385
V	Photopic spectral efficiency function		409
V'	Scotopic spectral efficiency function		409
V	Image vergence	diopter (m^{-1})	411
W_λ	Blackbody radiation function	$W\,m^{-3}$ or $W\,m^{-2}\,nm^{-1}$	396
W_ν	Blackbody radiation function	$W\,m^{-2}\,Hz^{-1}$	397
W_r	Exitance	$W\,m^{-2}$	408
α	Angle		411
δ_{th}	Thermal penetration depth	m	405
ϵ_0	Electrical permittivity of free space	$N^{-1}\,C^2\,m^{-2}$	382
ϵ	Emissivity		395
$\epsilon(\lambda)$	Erythema action spectrum		402
θ, ϕ	Angles		388
φ	Particle fluence rate	$m^{-2}\,s^{-1}$	390
λ	Wavelength	m	382
μ	Total linear attenuation coefficient	m^{-1}	387
μ_a	Linear absorption coefficient	m^{-1}	387
μ_s	Linear scattering coefficient	m^{-1}	387
μ'_s	Reduced linear scattering coefficient	m^{-1}	389
μ_{eff}	Effective linear attenuation coefficient	m^{-1}	391
μ_0	Magnetic permeability of free space	$\Omega\,s\,m^{-1}$	382
ρ, ρ_b, ρ_t	Density, density of blood, density of tissue	$kg\,m^{-3}$	388
σ, σ_i, σ_a, σ_s, σ_{tot}	Cross section	m^2	388
$\sigma(\theta)$, $d\sigma/d\Omega$	Differential scattering cross section	$m^2\,sr^{-1}$	388
σ_{SB}	Stefan–Boltzmann constant	$W\,m^{-2}\,K^{-4}$	397
σ_r^2, σ_x^2, σ_y^2, σ_z^2	Variance for diffusion or heat flow	m^2	405
ν	Frequency	s^{-1}	382
τ_{coh}	Coherence time	s	393
ω	Angular frequency	(radian) s^{-1}	382
ψ	Energy fluence rate	$W\,m^{-2}$	405
Ψ	Energy fluence	$J\,m^{-2}$	405
Φ	Particle fluence	m^{-2}	388
Ω	Solid angle	sr	388

Problems

Section 14.1

Problem 1. The velocity of light c depends on the parameters ϵ_0 and μ_0. Use dimensional analysis to find what the dependence must be. Insert numerical values to obtain c.

Problem 2. An einstein is 1 mol of photons. Derive an expression for the energy in an einstein as a function of wavelength. Express the answer in kilocalories and the wavelength λ in nanometers.

Section 14.3

Problem 3. Use Eq. 14.8 to derive Eq. 14.9.

Problem 4. (a) Starting with Eq. 14.8, derive a formula for the hydrogen atom spectrum in the form

$$\frac{1}{\lambda} = R\left[\frac{1}{n^2} - \frac{1}{m^2}\right],$$

where n and m are integers. R is called the *Rydberg constant*. Find an expression for R in terms of fundamental constants.

(b) Verify that the wavelengths of the spectral lines a–d at the top of Fig. 14.3 are consistent with the energy transitions shown at the bottom of the figure.

Problem 5. The Lyman series, part of the spectrum of hydrogen, is shown at the top of Fig. 14.3 as the line labeled *a* and the band of lines to the left of that line. Create a figure like Fig. 14.3, but which shows a detailed view of the Lyman series. Let the wavelength scale at the top of your figure range from 0 to 150 nm, as opposed to 0–2 μm in Fig. 14.3. Also include an energy level drawing like at the bottom of Fig. 14.3 and indicate which transitions correspond to which lines in the Lyman spectrum. Indicate the shortest possible wavelength in the Lyman spectrum, show what transition that wavelength corresponds to, and determine how this wavelength is related to the Rydberg constant.

Problem 6. The left side of Fig. 14.1 shows the emission of a photon during a transition from an initial state with energy E_i to a final one with energy E_f. Usually the Boltzmann factor ensures that the population of the initial state is less than the final state. In some cases however, when the initial state is *metastable*, one can create a *population inversion*. Photons with energy $h\nu$ corresponding to the energy difference $E_i - E_f$ can produce *stimulated emission* of other photons with the same energy, a type of positive feedback. Lasers work on this principle. Suppose a laser is made using two states having an energy difference of 1.79 eV. What is the wavelength of the output light? What color does this correspond to? Lasers have many uses in medicine (Peng et al. 2008).

Section 14.4

Problem 7. Estimate $\hbar^2/2I$ for an HCl molecule. What would the spacing of rotational levels be?

Problem 8. An inulin molecule has a molecular weight of 4000 dalton (that is, 1 mol has a mass of 4000 g). Assume that it is spherical with a radius of 1.2 nm. What is the angular frequency ω of a photon absorbed when its rotational quantum number changes from 10 to 11? The moment of inertia of a sphere rotating about an axis through its center is $I = (2/5)mR^2$.

Problem 9. The rotational spectrum of HCl contains lines at 60.4, 69.0, 80.4, 96.4, and 120.4 μm. What is the moment of inertia of an HCl molecule?

Problem 10. Consider a combined rotational–vibrational transition for which r goes from 1 to 0 while v goes from v to $(v - 1)$. Find the frequencies of the photons emitted in terms of the moment of inertia of the molecule I, the angular frequency of vibration of the atoms in the molecule ω, and the quantum number v.

Problem 11. A rotating molecule emits photons when the rotational quantum number changes by 1. Find the ratio of the angular frequency of the photons, ω_{phot}, to the angular frequency of rotation of the molecule, ω_{rot}, as a function of the rotational quantum number r.

Section 14.5

Problem 12. A beam with 200 particles per square centimeter passes by an atom. The particles are uniformly and randomly distributed in the area of the beam.

(a) Fifty particles are scattered. What is the total scattering cross section?

(b) Ten particles are scattered in a cone of 0.1 sr solid angle about a particular direction. What is the differential cross section in $\text{m}^2\ \text{sr}^{-1}$?

Problem 13. The differential scattering cross section for a beam of x-ray photons of a certain energy from carbon at an angle θ is $50 \times 10^{-30}\ \text{m}^2\ \text{sr}^{-1}$. A beam of 10^5 photons strikes a pure carbon target of thickness 0.3 cm. The density of carbon is $2\ \text{g cm}^{-3}$, and the atomic weight is 12. The detector is a circle of 1 cm radius located 20 cm from the target. How many scattered photons enter the detector?

Problem 14. Photochemists often use the *extinction coefficient* e, defined by $\mu_a = eC$, where C is the concentration in moles per liter. This assumes the substance being measured is dissolved in a completely transparent solvent.

(a) What are the units of the extinction coefficient?

(b) What is the conversion between the extinction coefficient and the absorption cross section?

Problem 15. Suppose that the absorption coefficient in some biological substance is 5 m^{-1}. Make the very crude assumption that the substance has the density of water and a molecular weight of 18. What is the absorption cross section?
Problem 16. For blue light ($\lambda = 470$ nm), the attenuation coefficient in air is about 2×10^{-5} m^{-1}, and the attenuation coefficient in pure water is about 5×10^{-3} m^{-1}. Calculate the distance that blue light must pass through air and water before the intensity is reduced to 1% of the original intensity. Compare these distances to the thickness of the atmosphere and the depth of the ocean. Do you think that aquatic plants can use photosynthesis effectively at the bottom of the ocean? For more on the differences between the optical properties of air and water, see Denny (1993).

Section 14.6

Problem 17. (a) Find the slope of $\log R$ versus t in Eq. 14.30. What is its value for large times?

(b) What can be determined from the time when R has its maximum value? (Hint: R has a maximum when $\log R$ has a maximum.)
Problem 18. The result of one set of infrared measurements in human calf (leg) muscle gave a total scattering coefficient $\mu_s = 8.3$ cm^{-1} and an absorption coefficient $\mu_a = 0.176$ cm^{-1}.

(a) What fraction of the photons have not scattered in passing through a layer that is 8 μm thick? (This corresponds roughly to the size of a cell.)
(b) On average, how many scattering events take place for each absorption event?
(c) What is the cross section for scattering per molecule? For this estimate, assume the muscle consists entirely of water, with molecular weight 18 and density 10^3 kg m^{-3}.

Problem 19. Consider light with fluence rate φ_0 continuously and uniformly irradiating a half-infinite slab of tissue having an absorption coefficient μ_a and a reduced scattering coefficient μ_s'. Divide the photons into two types: the incident ballistic photons that have not yet interacted with the tissue, and the diffuse photons undergoing multiple scattering. The diffuse photon fluence rate, φ, is governed by the steady state limit of the photon diffusion equation (Eq. 14.27). The source of diffuse photons is the scattering of ballistic photons, so the source term in Eq. 14.27 is $s = \mu_s' \exp(-z/\lambda_{\text{unatten}})$, where z is the depth below the tissue surface. At the surface ($z = 0$), the diffuse photons obey the boundary condition $\varphi = 2Dd\varphi/dz$.[17]

(a) Derive an analytical expression for the diffuse photon fluence rate in the tissue, $\varphi(z)$.
(b) Plot $\varphi(z)$ versus z for $\mu_a = 0.08$ mm^{-1} and $\mu_s' = 4$ mm^{-1}.
(c) Evaluate λ_{unatten} and λ_{diffuse} for these parameters.

Section 14.7

Problem 20. Carry out the averages leading to Eq. 14.31.
Problem 21. If yellow light from a source has a coherence time of 10^{-8} s, how many cycles are there in the wave?
Problem 22. What coherence time is needed for a spatial resolution of 1 μm?
Problem 23. An infrared transition involves an energy of 0.1 eV. What are the corresponding frequency and wavelength? If the Raman effect is observed with light at 550 nm, what will be the frequencies and wavelengths of each Raman line?
Problem 24. A Raman spectrum has a line at 500 nm with subsidiary lines at 400 and 667 nm. What is the wavelength of the corresponding infrared line?

Section 14.8

Problem 25. Sodium is introduced into a flame at 2500 K. What fraction of the atoms are in their first excited state? In their ground state? (Remember that the characteristic sodium line is yellow.) If the flame temperature changes by 10 K, what is the fractional change in the population of each state? Which method of measuring sodium concentration is more stable to changes in flame temperature: measuring the intensity of an emitted line or measuring the amount of absorption?
Problem 26. (a) Show that the maximum of the thermal radiation function $W_\lambda(\lambda, T)$ occurs at a wavelength such that $e^x(5 - x) = 5$, where $x = hc/(\lambda_{\max} k_B T)$. Verify that $x = 4.9651$ is a solution of this transcendental equation, so that

$$T\lambda_{\max} = \frac{hc}{4.9651 k_B}.$$

(b) Similarly, show that

$$\frac{\nu_{\max}}{T} = \frac{2.82144 k_B}{h}$$

and that $\lambda_{\max}\nu_{\max} = 0.57c$.
Problem 27. Let $W_\nu(\nu) = A\nu(\nu_0 - \nu)$ for $\nu < \nu_0$, and $W_\nu(\nu) = 0$ otherwise.

(a) Plot $W_\nu(\nu)$ versus ν.
(b) Calculate the frequency corresponding to the maximum of $W_\nu(\nu)$, called $\nu_{\max}$.

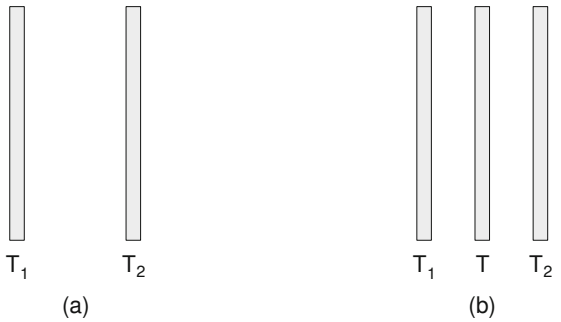
[17] The derivation of this boundary condition is found in Haskell et al. (1994). See also Roth (2008).

(c) Let $\lambda_0 = c/\nu_0$ and $\lambda_{max} = c/\nu_{max}$. Write λ_{max} in terms of λ_0.
(d) Integrate $W_\nu(\nu)$ over all ν to find W_{tot}.
(e) Use Eqs. 14.36 and 14.37 to calculate $W_\lambda(\lambda)$.
(f) Plot $W_\lambda(\lambda)$ versus λ.
(g) Calculate the wavelength corresponding to the maximum of $W_\lambda(\lambda)$, called λ^*_{max}, in terms of λ_0.
(h) Compare λ_{max} and λ^*_{max}. Are they the same or different? If λ_0 is 400 nm, calculate λ_{max} and λ^*_{max}. What part of the electromagnetic spectrum is each of these in?
(i) Integrate $W_\lambda(\lambda)$ over all λ to find W^*_{tot}. Compare W_{tot} and W^*_{tot}. Are they the same or different?

Problem 28. Integrate Eq. 14.33 over all wavelengths to obtain the Stefan–Boltzmann law, Eq. 14.34. You will need the integral

$$\int_0^\infty \frac{x^3 dx}{e^x - 1} = \frac{\pi^4}{15}.$$

Problem 29. Two parallel surfaces of area S have unit emissivity and are at temperatures T_1 and T_2 [$T_1 > T_2$, panel (a)]. They are large compared to their separation, so that all radiation emitted by one surface strikes the other. Assume that radiation is emitted and absorbed only by the two surfaces that face each other. Let P_0 be the energy lost per unit time by body 1. A new sheet of perfectly absorbing material is introduced between bodies 1 and 2, as shown in panel (b). It comes to equilibrium temperature T. Let P be the net energy lost by surface 1 in this case. Find P/P_0 in terms of T_1 and T_2.



Problem 30. The sun has a radius of 6.9×10^8 m. The earth is 149.5×10^9 m from the sun. Treat the sun as a thermal radiator at 5800 K and calculate the energy from the sun per unit area per unit time striking the upper atmosphere of the earth (the solar constant). State the result in W m^{-2} and cal cm^{-2} min^{-1}.

Problem 31. If all the energy received by the earth from the sun is lost as thermal radiation (a poor assumption because a significant amount is reflected from cloud cover), what is the equilibrium temperature of the earth?

Section 14.9

Problem 32. Show that an approximation to Eq. 14.41 for small temperature differences is $w_{tot} = SK_{rad}(T - T_s)$. Deduce the value of K_{rad} at body temperature. Hint: Factor $T^4 - T_s^4 = (T - T_s)(\cdots)$. You should get $K_{rad} = 6.76$ W m^{-2} K^{-1}.

Problem 33. What fractional change in $W_\lambda(\lambda, T)$ for thermal radiation from the human body results when there is a temperature change of 1 K at 5 μm? 9 μm? 15 μm?

Section 14.10

Problem 34. (a) Suppose that the threshold for erythema caused by sunlight with $\lambda = 300$ nm is 30 J m^{-2}. Does this suggest that the result is thermal (an excessive temperature increase) or something else, like the photoelectric effect or photodissociation? Make some reasonable assumptions to estimate the temperature rise.

(b) The energy in sunlight at all wavelengths reaching the earth is 2 cal cm^{-2} min^{-1}. Suppose that the total body area exposed is 0.6 m^2. What would be the temperature rise per minute for a 60 kg person if there were no heat-loss mechanisms? Compare the rate of energy absorption to the basal metabolic rate, about 100 W.

Problem 35. Suppose that the energy fluence rate of a parallel beam of ultraviolet light that has passed through thickness x of solution is given by $\psi = \psi_0 e^{-\mu_a x}$. (Scattering is ignored.) The absorption coefficient μ_a is related to the concentration C (molecules cm^{-3}) of the absorbing molecules in the solution by $\mu_a = eC$. Biophysicists working with ultraviolet light define the dose rate to be the power absorbed per molecule of absorber. (This is a different definition of dose than is used in Chap. 15.) Calculate the dose rate for a thin layer ($\mu_a x \ll 1$).

Problem 36. A beam of photons passes through a monatomic gas of molecular weight A and absorption cross section σ. Ignore scattering. The gas obeys the ideal gas law, $pV = Nk_BT$.
(a) Find the attenuation coefficient in terms of σ, p, and any other necessary variables.
(b) Generalize the result to a mixture of several gases, each with cross section σ_i, partial pressure p_i, and N_i molecules.

Problem 37. The attenuation of a beam of photons in a gas of pressure p is given by $d\Phi = -\Phi(\sigma p/k_BT)\,dx$, where σ is the cross section, k_B the Boltzmann constant, T the absolute temperature, and x the path length. Suppose that the pressure is given as a function of altitude y by $p = p_0 e^{-mgy/k_BT}$. What is the total attenuation by the entire atmosphere?

Problem 38. Consider a beam of photons incident on the atmosphere from directly overhead. The atmosphere contains several species of molecules, each with partial pressure p_i. The absorption coefficient is $\mu_a = (1/k_B T) \sum_i \sigma_i p_i$. If each constituent of the atmosphere varies with height y as $p_i(y) = p_{0i} \exp(-m_i g y / k_B T)$, show that the fluence rate striking the earth is given by an expression of the form $e^{-\alpha}$ and find α.

Section 14.11

Problem 39. Consider a tissue with a specific heat of 3.6 J kg^{-1} K^{-1}, a density of 1000 kg m^{-3}, and a thermal conductivity of 0.5 W m^{-1} K^{-1}. Assume the specific heat of blood is the same, and that the tissue perfusion is 4.17×10^{-6} m^3 kg^{-1} s^{-1}. Find the thermal diffusivity, the time for the heat to flow 1 cm, and the thermal penetration depth.

Section 14.12

Problem 40. Suppose that a sphere radiates uniformly from its surface according to Lambert's cosine law: $L = L_0$. By considering area $dS = 2\pi r^2 \sin\theta \, d\theta$ on the surface of a sphere, find the power radiated per steradian in the direction of the z axis and the total power radiated.

Problem 41. Show that the exitance, total power per unit area radiated from a surface obeying Lambert's cosine law, is $W_r = \pi L_0$.

Problem 42. How many photons per second correspond to 1 lm at 555 nm for photopic vision? At 510 nm for scotopic vision?

Section 14.13

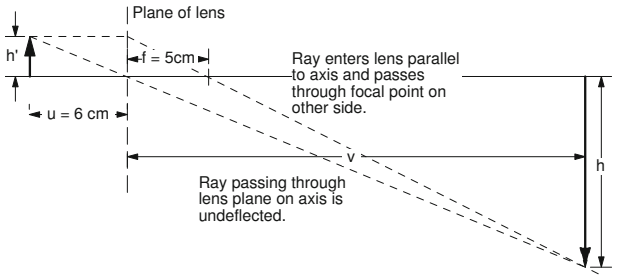
Problem 43. A person is nearsighted, and the relaxed eye focuses at a distance of 50 cm. What is the strength of the desired corrective lens in diopters?

Problem 44. What is the distance of closest vision for an average person with normal vision at age 20? Age 40? Age 60?

Problem 45. A person of age 40 is fitted with bifocals with a strength of +1 diopter more than the correction for distance vision. What are the closest and farthest distances of focus without the bifocal lens and with it? By the time the person is age 50, what are they with and without the same lens?

Problem 46. You can make a rough measurement of your own eye's properties. Tape a piece of paper with some pattern on it on the wall. Cover one eye. Move away from the wall until the pattern starts to blur. Measure the distance to the wall in meters. Calculate the vergence of the object, U. Assume that the F of your relaxed eye is 59 diopters. Calculate V for your eye. Now find the closest distance at which you can see the paper. Calculate the accommodation of your eye.

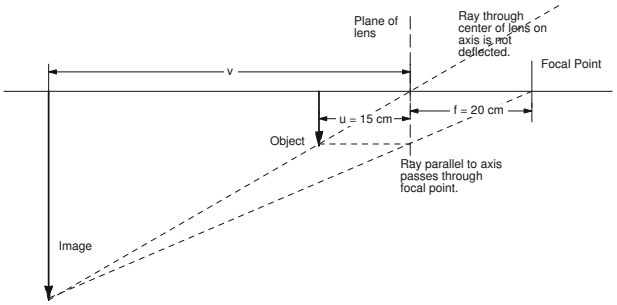
Problem 47. An object is placed 6 cm from a converging lens with a 5-cm focal length.



(a) Use the thin-lens equation (Eq. 14.64) to calculate the image distance.

(b) The magnification of the image is given by $m = -v/u$. (A negative magnification implies an inverted image.) What is the magnification for the image in part (a)? A value $|m| > 1$ implies a"magnified" image. This is how a slide projector works.

Problem 48. An object is placed 15 cm from a converging lens with a focal length of 20 cm.



(a) Use the thin-lens equation (Eq. 14.64) to calculate the image distance. Your value should be negative, corresponding to a "virtual image".

(b) The magnification of the image is again given by $m = -v/u$. What is the magnification for the image in part (a)? This is how a magnifying glass works.

Problem 49. Combine the results of Problems 47 and 48. Consider two lenses, the first with focal length 5 cm and the second with focal length 20 cm, separated by 45 cm. The object is 6 cm in front of the first lens. The image from the first lens is the object for the second.

(a) Calculate the image distance and magnification of the image created by the first lens (called the objective).

(b) Use the first image as the object for the second lens (called the eyepiece), and calculate the image distance and magnification of the second image.

(c) The total magnification is the product of the magnifications of the objective and eyepiece. What is the total magnification? This is how the compound microscope works. The objective lens acts like a slide projector, and the eyepiece acts like a magnifying glass. Very large total magnifications can be obtained when the object is just to the left of the focal point of the objective, and the first image is just to the right of the focal point of the eyepiece.

Problem 50. Snell's law, $n_1 \sin\theta_1 = n_2 \sin\theta_2$, gives an interesting result if light passes from a medium with a higher index of refraction to one with a lower index of refraction, $n_1 > n_2$. Assume light passes from glass ($n_1 = 1.5$) to air ($n_2 = 1.0$).

(a) If θ_1 is 30°, what is θ_2?
(b) If θ_1 is 40°, what is θ_2?
(c) If θ_1 is 50°, what is θ_2?
This is really a tricky question, because for θ_1 greater than some *critical angle*, θ_c, θ_2 exceeds 90°, and light cannot pass into the second medium. Instead all the light is reflected and remains within the first medium.
(d) Calculate the critical angle for *total internal reflection* from glass to air.

Total internal reflection allows thin glass fibers to act as fiberoptic "light pipes," which can be used to transmit signals. Bundles of such optical fibers are used in endoscopes to see inside the body.

Problem 51. Table 14.7 shows that most of the converging power of the eye occurs at the air-cornea interface. When a person is under water, this must be supplied by the water-cornea surface. The index of refraction of the cornea is closer to that of water than to that of air. What are the implications for seeing under water? What are the implications for the vision of aquatic animals? (For more information on the difference between the eyes of aquatic and terrestrial animals, see Denny 1993.)

Section 14.14

Problem 52. How many photons per 0.1 s enter the eye from a 100 W light bulb 1000 ft away? Assume the pupil is 6 mm in diameter. How far away can a 100 W bulb be seen if there is no absorption in the atmosphere? Use a luminous efficiency of 17 lm W^{-1} and then assume an equivalent light source at 555 nm.

Problem 53. The table below shows the radiance of some extended sources. Without worrying about obliquity factors (assume that all the light is at normal incidence), calculate the number of photons entering a receptive field of 0.17° diameter when the pupil diameter is 6 mm and the integration time is 0.1 s. Assume a conversion efficiency of 100 lm W^{-1} and then assume that all the photons are at 555 nm.

Source	Radiance (lm m^{-2} sr^{-1})
White paper in sunlight	25000
Clear sky	3200
Surface of the moon	2900
White paper in moonlight	0.03

Problem 54. A piece of paper is illuminated by dim light so that its radiance is 0.01 lm m^{-2} sr^{-1} in the direction of a camera. A camera lens 1 cm in diameter is 0.6 m from the paper. The sheet of paper is 10 × 10 cm. The shutter of the camera is open for 1 ms. Assume all the light is at 555 nm. How many photons from the paper enter the lens of the camera while the shutter is open?

Problem 55. If three or more photons must be absorbed by a visual receptor field for the observer to see a flash, what fraction of the flashes are seen if the average number of photons absorbed in a receptor field per flash is four?

Problem 56. Assume that an average of d photons are detected and that the photons are Poisson distributed. What must d be to detect a signal that is a 1% change in d, if the signal-to-noise ratio must be at least 5?

Problem 57. Suppose that the average number of photons striking a target during an exposure is m. The probability that x photons strike during a similar exposure is given by the Poisson distribution. What is the probability that an organism responds to an exposure of radiation in each of the following cases?

(a) The response of the organism requires that a single target within the organism be hit by two or more photons.
(b) The response of the organism requires that two targets within the organism each be struck by one or more photons during the exposure.

References

AAPM 57 (1996) Recommended nomenclature for physical quantities in medical applications of light. American Association of Physicists in Medicine (AAPM) Report No. 57. College Park, MD

Armstrong CM (2012) The truth about terahertz radiation. IEEE Spectr 49:36–41

Armstrong BK, Kricker A (2001) The epidemiology of UV induced skin cancer. J Photochem Photobiol B Biol 63:8–18

Barlow HB (1956) Retinal noise and absolute threshold. J Opt Soc Am 46:634–639

Barlow HB (1957) Incremental thresholds at low intensities considered as signal/noise discriminations. J Physiol 136:469–488

Barysch MJ, Hofbauer GF, Dummer R (2010) Vitamin D, ultraviolet exposure, and skin cancer in the elderly. Gerontology 56:410–413. doi:10.1159/000315119

Bergmanson JPG, Söderberg PG (1995) The significance of ultraviolet radiation for eye diseases. A review with comments on the efficacy of UV-blocking contact lenses. Ophthalmic Physiol Opt 15(2):83–91

Berne BJ, Pecora R (1976) Dynamic light scattering: with applications to chemistry, biology, and physics. Wiley, New York
Blume S (1993) Social process and the assessment of a new imaging technique. Intl J Technol Assess Health Care 9(3):335–345
Bramson MA (1968) Infrared radiation. Plenum, New York
Brezinski ME (2006) Optical coherence tomography. Elsevier, Amsterdam
Buka RL (2004) Sunscreens and insect repellents. Curr Opin Pediatr 16:378–384
Charman WN (2009) Optics of the eye. In: Bass M (ed) Handbook of optics, 3rd ed, Vol. III. McGraw-Hill, New York. (Sponsored by the Optical Society of America.)
Cornsweet TN (1970) Visual perception. Academic, New York
de Boer JF Srinivas SM Nelson JS Milner TE Ducros MG (2002) Polarization-sensitive optical coherence tomography. In: Bouma BE, Tearney GJ (eds) Handbook of optical coherence tomography. Marcel Dekker, New York, pp 274–298
Denny MW (1993) Air and water: the biology and physics of life's media. Princeton University Press, Princeton
Diem M (1993) Introduction to modern vibrational spectroscopy. Wiley, New York
Diffey BL (1991) Solar ultraviolet radiation effects on biological systems. Phys Med Biol 36(3):299–328
Diffey BL, Cheeseman J (1992) Sun protection with hats. Brit J Dermatol 127(1):10–12
Duderstadt JJ, Hamilton LJ (1976) Nuclear reactor analysis. Wiley, New York
Eisberg R, Resnick R (1985) Quantum physics of atoms, molecules, solids, nuclei and particles, 2nd ed. Wiley, New York
Esenaliev RO, Larin KV, Larina IV, Motamedi M (2001) Noninvasive monitoring of glucose concentration with optical coherence tomography. Opt Lett 26(13):992–994
Farmer J (1997) Blood oxygen measurement. In: Webster JG (ed) Design of pulse oximeters. Institute of Physics, Bristol
Fercher AF, Drexler W, Hitzenberger CK, Lasser T (2003) Optical coherence tomography: principles and applications. Rep Prog Phys 66:239–303
Fitzgerald AJ, Berry E, Zinovev NN, Walker GC, Smith MA, Chamberlain JM (2002) An introduction to medical imaging with coherent terahertz frequency radiation. Phys Med Biol 47:R67–R84
Fraden J (1991) Noncontact temperature measurements in medicine. In: Wise DL (ed) Bioinstrumentation and biosensors. Dekker, New York, pp 511–549
Gasiorowicz S (2003) Quantum physics, 3rd ed. Wiley, New York
Giasson CJ, Quesnel N-M, Boisjoly H (2005) The ABCs of ultraviolet-blocking contact lenses: an ocular panacea for ozone loss? Int Ophthalmol Clin 45(1):117–139
Greivenkamp JE, Schwiegerling J, Miller JM, Mellinger MD (1995) Visual acuity modeling using optical raytracing of schematic eyes. Am J Ophthalmol 120(2):227–240
Grossweiner L (1994) The science of phototherapy. CRC, Boca Raton
Hanlon EB, Manoharan R, Koo T-W, Shafer KE, Motz JT, Fitzmaurice M, Kramer JR, Itzkan I, Dasari RR, Feld MS (2000) Prospects for *in vivo* Raman spectroscopy. Phys Med Biol 45:R1–R59
Hao O-Y, Stanfield J, Cole C, Appa Y, Rigel D (2012) High-SPF sunscreens (SPF $\geq$ 70) may provide ultraviolet protection above minimal recommended levels by adequately compensating for lower sunscreen user application amounts Capsule Summary. J Am Acad Dermatol 67(6):1220–1227. doi:10.1016/j.jaad.2013.04.002
Halliday D, Resnick R, Krane KS (1992) Fundamentals of physics, 4th ed., extended edition, Vol. 2 . Wiley, p 1022
Haskell RC, Svassand LO, Tsay T-T, Feng T-C, McAdams MS, Tromberg BJ (1994) Boundary conditions for the diffusion equation in radiative transfer. J Opt Soc Am A 11(10):2727–2741
Heald MA (2003) Where is the "Wien peak"? Am J Phys 71(12):1322–1323
Hecht S, Shlaer S, Pirenne MH (1942) Energy, quanta, and vision. J Gen Physiol 25:819–840
Hielscher A, Jacques SL, Wang L, Tittel FK (1995) The influence of boundary conditions on the accuracy of diffusion theory in time-resolved reflectance spectroscopy of biological tissues. Phys Med Biol 40:1957–1975
Huang D, Swanson EA, Lin CP , Schuman JS, Stinson WG, Chang W, Hee MR, Flotte T, Gregory K, Puliafito CA, Fujimoto JG (1991) Optical coherence tomography. Science 254:1178–1181
Johnsen S (2011) The optics of life: a biologist's guide to light in nature. Princeton University Press, Princeton
Kevan PG, Chittka L, Dyer AG (2001) Limits to the salience of ultraviolet: lessons from colour vision in bees and birds. J Exp Biol 204:2571–2580
Kimlin MG, Parisi AV (1999) Ultraviolet radiation penetrating vehicle glass: a field based comparative study. Phys Med Biol 44:917–926
Kligman LH (1989) Photoaging: manifestations, prevention, and treatment. Clin Geriatr Med 5:235–251
Knobler R, Barr ML, Couriel DR, Ferrara JLM, French LE, Jaksch P, Reinisch W, Rook AH, Schwarz T, Greinix H (2009) Extracorporeal photopheresis: past, present, and future. J Am Acad Dermatol 61:652–665
Kripke ML (2003) The ABCs of sunscreen protection factors. J Invest Dermatol 121(1):vii–viii
Lahiri BB, Bagavathiappan S, Jayakumar T, Philip J (2012) Medical applications of infrared thermography: a review. Infrared Phys Technol 55:221–235
Lazovich D, Vogel RI, Berwick M, Weinstock MA, Anderson KE, Warshaw EM (2010) Indoor tanning and risk of melanoma: a case-control study in a highly exposed population. Cancer Epidemiol Biomarkers Prev 19(6):1557–1568
Liu H, Boas DA, Zhang Y, Yodh AG, Chance B (1995) Determination of optical properties and blood oxygenation in tissue using continuous NIR light. Phys Med Biol 40:1983–1993
MacNeill BD, Lowe HC, Takano M, Fuster V, Jang I-K (2003) Intravascular modalities for detection of vulnerable plaque. Arterioscler Thromb Vasc Biol 23:1333–1342
Madronich S (1993) The atmosphere and UV-B radiation at ground level. In: Young AR et al. (eds) Environmental UV photobiology. Plenum, New York, pp 1–39
McDonagh AF (1985) Light effects on transport and excretion of bilirubin in newborns. Ann NY Acad Sci 453:65–72
Moyal DD, Fourtanier AM (2002) Effects of UVA radiation on an established immune response in humans and sunscreen efficacy. Exp Dermatol 11(Suppl. 1):28–32
Murphy MR, Oellrich RG (1990) A new method of phototherapy: nursing perspectives. J Perinatol 10(3):249–251
Nickell S, Hermann M, Essenpreis M, Farrell TJ, Krämer U, Patterson MS (2000) Anisotropy of light propagation in human skin. Phys Med Biol 45:2873–2886
O'Sullivan N, Tait C (2014) Tanning bed and nail lamp use and the risk of cutaneous malignancy: a review of the literature. Australas J Dermatol 55:99–106. doi:10.1111/ajd.12145
Patterson MS, Chance B, Wilson BC (1989) Time resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties. Appl Opt 28(12):2331–2336
Peng Q, Juzeniene A, Chen J, Svaassand LO, Warloe T, Giercksky K-E, Moan J (2008) Lasers in medicine. Rep Prog Phys 71:056701. doi:10.1088/0034-4885/71/5/056701
Pillsbury DM, Heaton CL (1980) A manual of dermatology, 2nd ed. Saunders, Philadelphia, p 5
Pogue BW, Patterson MS (1994) Frequency-domain optical absorption spectroscopy of finite tissue volumes using diffusion theory. Phys Med Biol 39:1157–1180
Rieke F, Baylor DA (1998) Single-photon detection by rod cells of the retina. Rev Mod Phys 70(3):1027–1036

Rodieck RW (1998) The first steps in seeing. Sunderland, Sinauer
Rose A (1973) Vision: human and electronic. New York, Plenum
Roth BJ (2008) Photon density measured over a cut surface: implications for optical mapping of the heart. IEEE Trans Biomed Eng 55(8):2102–2104
Saint-Jalmes H, Lebec M, Beaurepaire E, Dubois A, Boccara AC (2002) Full-field optical coherence microscopy. In: Bouma BE, Tearney GJ (eds) Handbook of optical coherence tomography. Marcel Dekker, New York , pp 299–333
Sakitt B (1972) Counting every quantum. J Physiol (London) 223:131–150
Schmidt CW (2012) UV radiation and skin cancer: the science behind age restrictions for tanning beds. Environ Health Perspect 120(8):A308–A313
Schmitt JM (1999) Optical coherence tomography (OCT): a review. IEEE J Sel Top Quantum Electron 5:1205–1215
Schroeder DV (2000) An introduction to thermal physics. Addison Wesley Longman, San Francisco
Serway RA, Jewett JW (2013) Principles of physics, 5th ed. Brooks/Cole, Boston
Sevick EM, Chance B, Leigh J, Nioka S, Maris M (1991) Quantitation of time- and frequency-resolved optical spectra for the determination of tissue oxygenation. Analyt Biochem 195:330–351
Smye SW, Chamberlain JM, Fitzgerald AJ, Berry E (2001) The interaction between terahertz radiation and biological tissue. Phys Med Biol 46:R101–R112
Soffer BH, Lynch DK (1999) Some paradoxes, errors, and resolutions concerning spectral optimization of human vision. Am J Phys 67(11):946–953
Stark L, Theodoridis GC (1973) Information theory in physiology. In: Brown JHU, Gans DS (eds) Engineering principles in physiology, Vol. 1. Academic, New York , pp 13–31
Steketee J (1973) Spectral emissivity of skin and pericardium. Phys Med Biol 18:686–694
Stern RS (2007) Psoralen and ultraviolet—A therapy for psoriasis. N Engl J Med 357:682–690
Teramura T, Mizuno M, Asano H, Naito N, Arakane K, Miyachi Y (2012) Relationship between sun-protection factor and application thickness in high-performance sunscreen: double application of sunscreen is recommended. Clin Exp Dermatol 37:904–908. doi:10.1111/j.1365-2230.2012.04388.x
Verheye S, Diamantopoulos L, Serruys PW, van Langenhove G (2002) Intravascular imaging of the vulnerable atherosclerotic plaque: spotlight on temperature measurement. J Cardiovasc Risk 9:247–254
Vreugdenburg TD Willis CD Mundy L Hiller JE (2013). A systematic review of elastography, electrical impedance scanning, and digital infrared thermography for breast cancer screening and diagnosis. Breast Cancer Res Treat 137:665–676. doi:10.1007/s10549-012-2393-x
Webster JG (1997) Design of pulse oximeters. Institute of Physics Publishers, Bristol
Wieben O (1997) Light absorbance in pulse oximetry. In: Webster JG (ed) Design of pulse oximeters. Institute of Physics Publishers, Bristol
Wilson BC, Patterson MS (2008) The physics, biophysics and technology of photodynamic therapy. Phys Med Biol 53:R61–R109
Yamashita Y, Maki A, Koizumi H (2001) Wavelength dependence of the precision of noninvasive optical measurement of oxy-, deoxy-, and total-hemoglobin concentration. Med Phys 28(6):1108–1114
Yasuno Y, Makita S, Sutoh Y, Itoh M, Yatagai T (2002) Birefringence imaging of human skin by polarization-sensitive spectral interferometric optical coherence tomography. Opt Lett 27(20):1803–1805
Yazdanfar S, Rollins AM, Izatt JA. Imaging and velocimetry of the human retinal circulation with color Doppler optical coherence tomography. Opt Lett 25(19):1448–1450
Yodh A, Chance B (1995) Spectroscopy and imaging with diffusing light. Phys Today 48(3):34–41
Zalewski EF (2009) Radiometry and photometry In: Bass M (ed) Handbook of Optics, 3rd ed. McGraw-Hill, New York. (Sponsored by the Optical Society of America)
Zhang X-C (2002) Terahertz wave imaging: horizons and hurdles. Phys Med Biol 47:3667–3677
Zhu TC, Finlay JC (2008) The role of photodynamic therapy (PDT) physics. Med Phys 35:3127–3126

Interaction of Photons and Charged Particles with Matter

15

An x-ray image records variations in the passage of x rays through the body because of scattering and absorption. A side effect of making the image is the absorption of some x-ray or charged particle energy by the body. Radiation therapy depends on the absorption of large amounts of x-ray energy by a tumor. Diagnostic procedures in nuclear medicine (Chap. 17) introduce a small amount of radioactive substance in the body. Radiation from the radioactive nuclei is then detected. Some of the energy from the photons or charged particles emitted by the radioactive nucleus is absorbed in the body. To describe all of these effects requires that we understand the interaction of photons and charged particles with matter.

In Chap. 14 we discussed the transport of photons of ultraviolet and lower energy—a few electron volts or less. Now we will discuss the transport of photons of much higher energy—10 keV and above. We will also discuss the movement through matter of charged particles such as electrons, protons, and heavier ions. These high energy photons and charged particles are called *ionizing radiation*, because they produce ionization in the material through which they pass. The distinction is blurred, since ultraviolet light can also ionize.

A charged particle moving through matter loses energy by local ionization, disruption of chemical bonds, and increasing the energy of atoms it passes near. It is said to be *directly ionizing*. Photons passing through matter transfer energy to charged particles, which in turn affects the material. These photons are *indirectly ionizing*.

Photons and charged particles interact primarily with the electrons in atoms. Section 15.1 describes the energy levels of atomic electrons. Section 15.2 describes the various processes by which photons interact; these are elaborated in the next four sections, leading in Sect. 15.7 to the concept of a photon attenuation coefficient. Attenuation is extended to compounds and mixtures in Sect. 15.8.

An atom is often left in an excited state by a photon interaction. The mechanisms by which it loses energy are covered in Sect. 15.9. The energy that is transferred to electrons can cause radiation damage. The transfer process is described in Sects. 15.10 and 15.15–15.17.

Section 15.11 introduces the charged-particle stopping power, which is the rate of energy loss by a charged particle as it passes through a material. Extensions of this concept, which are important in radiation damage, are the linear energy transfer and the restricted collision stopping power, introduced in Sect. 15.12. A charged particle travels a certain distance through the material as it loses its kinetic energy. This leads in Sect. 15.13 to the concept of range. Charged particles also lose energy by emitting photons. The radiation yield is also discussed in Sect. 15.13. Insight into the process of radiation damage is gained by examining track structure in Sect. 15.14.

The last three sections return to the movement of energy from a photon beam to matter. The discussion requires an understanding of both photon interactions and charged-particle stopping power and range.

15.1 Atomic Energy Levels and X-ray Absorption

A neutral atom has a nuclear charge $+Ze$ surrounded by a cloud of Z electrons. As was described in Chap. 14, each electron has a definite energy, characterized by a set of five quantum numbers: n, l, s (which is always $\frac{1}{2}$), j, and m_j. (Instead of j and m_j, the numbers m_l and m_s are sometimes used.) There are restrictions on the values of the numbers:

$n = 1, 2, 3, \ldots$	the principal quantum number
$l = 0, 1, 2, \ldots, n-1$	the orbital angular momentum quantum number
$s = \frac{1}{2}$	the spin quantum number

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_15, © Springer International Publishing Switzerland 2015

$j = l - \frac{1}{2}$ or $l + \frac{1}{2}$, except that	the total angular
$j = \frac{1}{2}$ when $l = 0$	momentum quantum number
$m_j = -j, -(j-1), \ldots,$	"z component" of the total
$(j-1), j$	angular momentum

The dependence of the electron energy on m_j is very slight, unless the atom is in a magnetic field.

In each atom, only one electron can have a particular set of values of the quantum numbers. Since the atoms we are considering are not in a magnetic field, electrons with different values of m_j but the same values for n, l, and j will all be assumed to have the same energy. Electrons with different values of n are said to be in different *shells*. The shell for $n = 1$ is called the K shell; those for $n = 2, 3, 4, \ldots$ are labeled $L, M, N, \ldots$. Different values of l and j for a fixed value of n are called *subshells*, denoted by roman numeral subscripts on the shell labels. The maximum number of electrons that can be placed in a subshell is $2(2l + 1)$. Each electron bound to the atom has a certain negative energy, with zero energy defined when the electron is just unbound, that is, at rest infinitely far away from the atom. Table 15.1 lists the energy levels of electrons in tungsten. Some of the levels in Table 15.1 are shown in Fig. 15.1. The scale is logarithmic. Since the energies are negative, the magnitude increases in the downward direction. Tables of atomic energy levels can be found many places, including www.csrri.iit.edu/periodic-table.html .

The *ionization energy* is the energy required to remove the least-tightly-bound electron from the atom. For tungsten, it is about 6 eV. If one plots the ionization energy or the chemical valence of atoms as a function of Z, one finds abrupt changes when the last electron's value of n or l changes.

In contrast to this behavior of the outer electrons, the energy of an inner electron with fixed values of n and l varies smoothly with Z. To a first approximation, the two innermost K electrons are attracted to the nuclear charge Ze. The energy of the level can be estimated using Eq. 14.9 for hydrogen, with the nuclear charge e replaced by Ze:

$$E_K = -\frac{13.6Z^2}{1^2}. \qquad (15.1)$$

The two electrons also repel each other and experience some repulsion by electrons in other shells. This effect is called *charge screening*. Experiment (measuring values of E_K) shows that the effective charge seen by a K electron is approximately $Z_{\text{eff}} \approx Z - 2$ for heavy elements, so that for K electrons ($n = 1$),

$$E_K \approx -13.6(Z-2)^2 \qquad \text{(in eV)}. \qquad (15.2)$$

Table 15.1 Energy levels for electrons in a tungsten atom ($Z = 74$)

n	l	j	Number of electrons	X-ray label	Energy (eV)
1	0	$\frac{1}{2}$	2	K	−69 525
2	0	$\frac{1}{2}$	2	L_{I}	−12 100
	1	$\frac{1}{2}$	2	L_{II}	−11 544
	1	$\frac{3}{2}$	4	L_{III}	−10 207
3	0	$\frac{1}{2}$	2	M_{I}	−2 820
	1	$\frac{1}{2}$	2	M_{II}	−2 575
	1	$\frac{3}{2}$	4	M_{III}	−2 281
	2	$\frac{3}{2}$	4	M_{IV}	−1 872
	2	$\frac{5}{2}$	6	M_{V}	−1 809
4	0	$\frac{1}{2}$	2	N_{I}	−594
	1	$\frac{1}{2}$	2	N_{II}	−490
	1	$\frac{3}{2}$	4	N_{III}	−424
	2	$\frac{3}{2}$	4	N_{IV}	−256
	2	$\frac{5}{2}$	6	N_{V}	−244
	3	$\frac{5}{2}, \frac{7}{2}$	14	$N_{\text{VI,VII}}$	−35
5	0	$\frac{1}{2}$	2	O_{I}	−77
	1	$\frac{1}{2}$	2	O_{II}	−47
	1	$\frac{3}{2}$	4	O_{III}	−36
	2	$\frac{3}{2}, \frac{5}{2}$	4	$O_{\text{IV,V}}$	−6
6	0	$\frac{1}{2}$	2	P_{I}	

The screening is greater for electrons with larger values of n, which may be thought of as being in "orbits" of larger radius.

15.2 Photon Interactions

There are a number of different ways in which a photon can interact with an atom. The more important ones will be considered here. It is convenient to adopt a notation (γ, bc) where γ represents the incident photon and b and c are the results of the interaction. For example (γ, γ) represents initial and final photons having the same energy; in a (γ, e) interaction the photon is absorbed and only an electron emerges. This section describes the common interactions and the energy balance for each case.

15.2.1 Photoelectric Effect

In the *photoelectric effect,* (γ, e), the photon is absorbed by the atom and a single electron, called a *photoelectron*, is

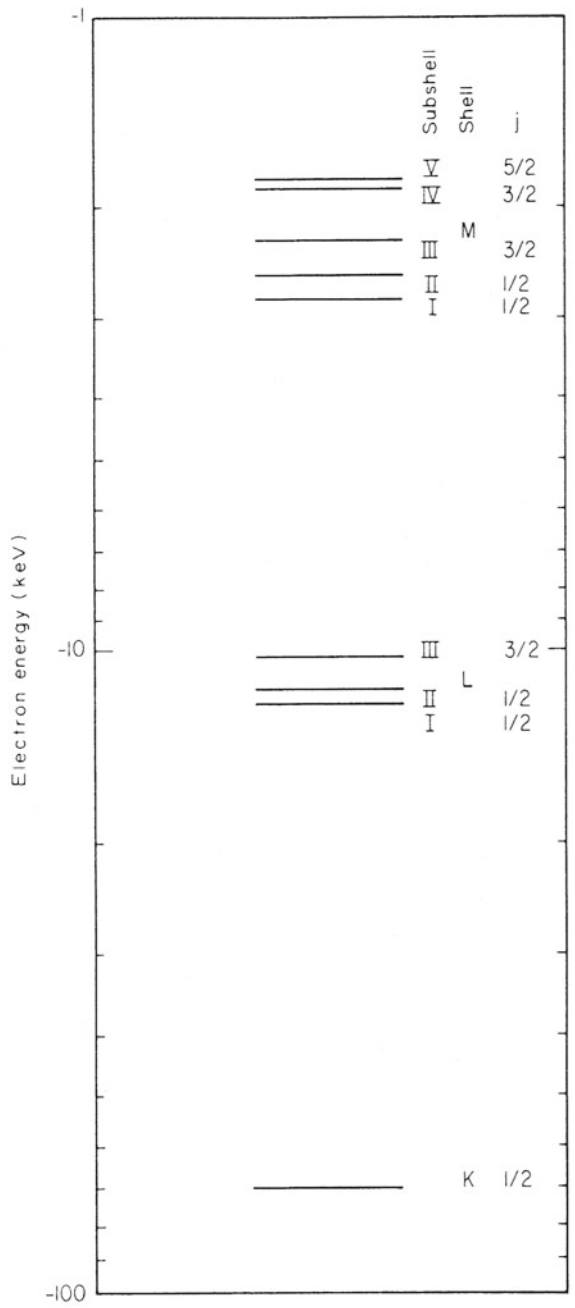


Fig. 15.1 Energy levels for electrons in tungsten

ejected. The initial photon energy $h\nu_0$ is equal to the final energy. The recoil kinetic energy of the atom is very small because its mass is large, so the final energy is the kinetic energy of the electron, T_{el}, plus the excitation energy of the atom. The excitation energy is equal to the binding energy of the ejected electron, B. The energy balance is therefore

$$h\nu_0 = T_{\text{el}} + B. \tag{15.3}$$

The atom subsequently loses its excitation energy. The deexcitation process described in Sect. 15.9 involves the emission of additional photons or electrons. The photoelectric cross section is τ.

15.2.2 Compton and Incoherent Scattering

In *Compton scattering*, $(\gamma, \gamma' e)$, the original photon disappears and a photon of lower energy and an electron emerge. The statement of energy conservation is

$$h\nu_0 = h\nu + T_{\text{el}} + B.$$

Usually the photon energy is high enough so that B can be neglected, and this is written as

$$h\nu_0 = h\nu + T_{\text{el}}. \tag{15.4}$$

The Compton cross section for scattering from a single electron is σ_C. *Incoherent scattering* is Compton scattering from all the electrons in the atom, with cross section σ_{incoh}.

15.2.3 Coherent Scattering

Coherent scattering is a (γ, γ) process in which the photon is elastically scattered from the entire atom. That is, the internal energy of the atom does not change. The recoil kinetic energy of the atom is very small (see Problem 8), and it is a good approximation to say that the energy of the incident photon equals the energy of the scattered photon:

$$h\nu_0 = h\nu. \tag{15.5}$$

The cross section for coherent scattering is σ_{coh}.

15.2.4 Inelastic Scattering

It is also possible for the final photon to have a different energy from the initial photon (γ, γ') without the emission of an electron. The internal energy of the target atom or molecule increases or decreases by a corresponding amount. Again, the recoil kinetic energy of the atom is negligible. Examples are fluorescence and Raman scattering. In fluorescence, if one waits long enough, additional photons are emitted, in which case the reaction could be denoted as $(\gamma, \gamma'\gamma'')$, or $(\gamma, 2\gamma)$, or even $(\gamma, 3\gamma)$.

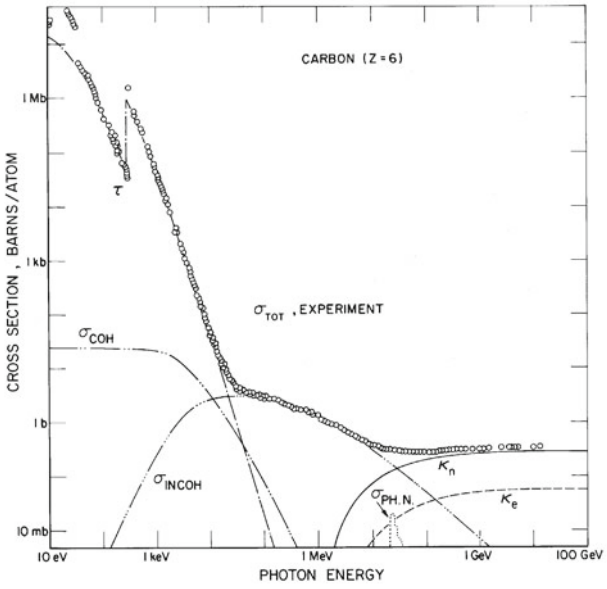


Fig. 15.2 Total cross section for the interactions of photons with carbon vs photon energy. The photoelectric cross section is τ, the coherent scattering cross section σ_{coh}, the total Compton cross section σ_{incoh}, and the nuclear and electronic (triplet) pair production are κ_n and κ_e. The photonuclear scattering cross section PHN is also shown. The cross section is given in barns: $1\,\text{b} = 10^{-28}\,\text{m}^2$. Reprinted with permission from Hubbell et al. (1980). Copyright 1980, American Institute of Physics. Figure courtesy of J. H. Hubbell

15.2.5 Pair Production

Pair production takes place at high energies. This is a (γ, e^+e^-) reaction. Since it takes energy to create the (negative) electron and the positive electron or *positron*, their rest energies must be included in the energy balance equation:

$$h\nu_0 = T_+ + m_e c^2 + T_- + m_e c^2 = T_+ + T_- + 2m_e c^2. \quad (15.6)$$

The cross section for pair production is κ.

15.2.6 Energy Dependence

Figure 15.2 shows the cross section for interactions of photons with carbon for photon energies from 10 to 10^{11} eV. At the lowest energies the photoelectric effect dominates. Between 10 keV and 10 MeV Compton scattering is most important. Above 10 MeV pair production takes over. There is a small bump at about 20 MeV due to nuclear effects, but its contribution to the cross section is only a few percent of that due to pair production. The four important effects are discussed in the next four sections.

15.3 The Photoelectric Effect

In the photoelectric effect a photon of energy $h\nu_0$ is absorbed by an atom, and an electron of kinetic energy $T_{\text{el}} = h\nu_0 - B$ is ejected. B is the magnitude of the binding energy of the electron and depends on which shell the electron was in. Therefore it is labeled B_K, B_L, and so forth. The cross section for the photoelectric effect, τ, is a sum of terms for each shell:

$$\tau = \tau_K + \tau_L + \tau_M + \cdots . \quad (15.7)$$

As the energy of a photon beam is decreased, the photoelectric cross section increases rapidly. For photon energies too small to remove an electron from the K shell, the cross section for the K-shell photoelectric effect is zero. Even though photons do not have enough energy to remove an electron from the K shell, they may have enough energy to remove L-shell electrons. The cross section for L electron photoelectric effect is much smaller than that for K electrons, but it increases with decreasing energy until its threshold energy is reached. This energy dependence is shown for lead in Fig. 15.3, which plots the cross section for the photoelectric effect, incoherent Compton scattering, and coherent scattering. The K absorption edge for the photoelectric effect is seen. The photoelectric effect below the K absorption edge is due to $L, M, \ldots$ electrons; above this energy the K electrons also participate. Above 0.8 MeV in lead Compton scattering becomes more important than the photoelectric effect.

The energy dependence of the photoelectric effect cross section is between E^{-2} and E^{-3}. An approximation to the Z and E dependence of the photoelectric cross section near 100 keV is (Attix 1986, p. 140)

$$\tau \propto Z^4 E^{-3}. \quad (15.8)$$

Once an atom has absorbed a photon and ejected a photoelectron, it is in an excited state. The atom will eventually lose this excitation energy by capturing an electron and returning to its ground state. The deexcitation processes are described in Sect. 15.9.

15.4 Compton Scattering

15.4.1 Kinematics

Compton scattering is a $(\gamma, \gamma' e)$ process. The equations that are used to relate the energy and angle of the emerging photon and electron, as well as the equations that give the cross section for the scattering, are usually derived assuming that the electron is free and at rest. We turn first to the kinematics.

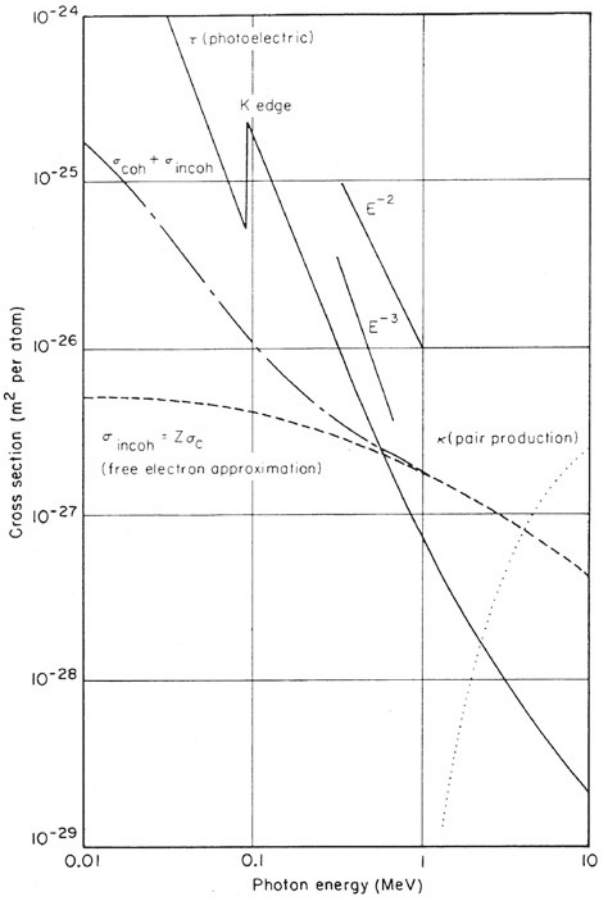


Fig. 15.3 Cross sections for the photoelectric effect and incoherent and coherent scattering from lead. The binding energies of the K and L shells are 0.088 and 0.0152 MeV. Plotted from Table 3.22 of Hubbell (1969)

A photon has energy E and momentum p, related by

$$E = h\nu = pc. \tag{15.9}$$

This is a special case of a more general relationship from special relativity:

$$E^2 = (pc)^2 + (m_0c^2)^2. \tag{15.10}$$

In these equations E is the total energy of the particle, p its momentum, m_0 the *rest mass* of the particle (measured when it is not moving), and m_0c^2 is the *rest energy*.[1] For a photon, which can never be at rest, $m_0 = 0$. Equation 15.9 can also be derived from the classical electromagnetic theory of light.

The conservation of energy and momentum can be used to derive the relationship between the angle at which the scattered photon emerges and its energy. A detailed knowledge of the forces involved is necessary to calculate the relative number of photons scattered at different angles; in fact, this calculation must be done using quantum mechanics. Figure 15.4 shows the geometry of the scattering. The electron emerges with momentum p, kinetic energy T, and total energy $E = T + m_ec^2$. It emerges at an angle ϕ with the direction of the incident photon. The scattered photon emerges at angle θ with a reduced energy and a corresponding frequency ν' which is lower than ν_0, the frequency of the incident photon. Conservation of momentum in the direction of the incident photon gives



Fig. 15.4 Momentum relationships in Compton scattering. **a** Before. **b** After. The photon emerges at angle θ, the electron at angle ϕ

$$\frac{h\nu_0}{c} = \frac{h\nu'}{c}\cos\theta + p\cos\phi,$$

while conservation of momentum at right angles to that direction gives

$$\frac{h\nu'}{c}\sin\theta = p\sin\phi.$$

Conservation of energy gives

$$h\nu_0 = h\nu' + T.$$

The equation $E = T + m_ec^2$ can be combined with Eq. 15.10 to give

$$(pc)^2 = T^2 + 2m_ec^2T.$$

The last four equations can then be combined and solved for various unknowns.

The wavelength of the scattered photon is

$$\lambda' - \lambda_0 = \frac{c}{\nu'} - \frac{c}{\nu_0} = \frac{h}{m_ec}(1 - \cos\theta). \tag{15.11}$$

The wavelength shift (but not the frequency or energy shift) is independent of the incident wavelength. The quantity h/m_ec has the dimensions of length and is called the *Compton wavelength* of the electron. Its numerical value is

$$\lambda_C = \frac{h}{m_ec} = 2.427 \times 10^{-12}\ \text{m} = 2.427\ \text{pm}. \tag{15.12}$$

[1] Since this is one of the few relativistic results we will need, it is not developed here. A discussion can be found in any book on special relativity.

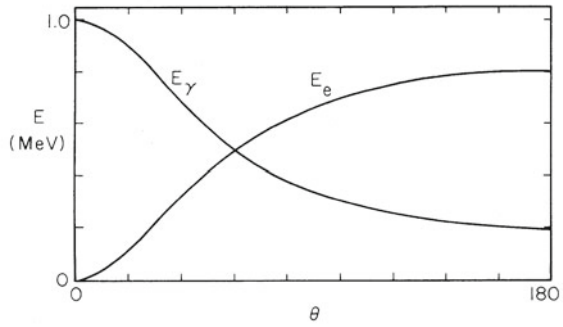


Fig. 15.5 The energy of the emerging photon and recoil electron as a function of θ, the angle of the emerging photon, for a 1-MeV incident photon

If Eq. 15.11 is solved for the energy of the scattered photon, the result is

$$h\nu' = \frac{h\nu_0}{1 + x(1 - \cos\theta)}, \tag{15.13}$$

where x is the energy of the incident photon in units of $m_e c^2 = 511$ keV:

$$x = \frac{h\nu_0}{m_e c^2}. \tag{15.14}$$

The energy of the recoil electron is $T = h\nu_0 - h\nu'$:

$$T = \frac{h\nu_0 x(1 - \cos\theta)}{1 + x(1 - \cos\theta)}. \tag{15.15}$$

Figure 15.5 shows the energy of the scattered photon and the recoil electron as a function of θ, the angle of emergence of the photon. The sum of the two energies is 1 MeV, the energy of the incident photon.

15.4.2 Cross Section: Klein–Nishina Formula

The inclusion of dynamics, which allows us to determine the relative number of photons scattered at each angle, is fairly complicated. The quantum-mechanical result is known as the *Klein–Nishina* formula (Attix 1986). The result depends on the polarization of the photons. For unpolarized photons, the cross section per unit solid angle for a photon to be scattered at angle θ is

$$\frac{d\sigma_C}{d\Omega} = \frac{r_e^2}{2}\left[\frac{1 + \cos^2\theta + \dfrac{x^2(1 - \cos\theta)^2}{1 + x(1 - \cos\theta)}}{[1 + x(1 - \cos\theta)]^2}\right], \tag{15.16}$$

where

$$r_e = \frac{e^2}{4\pi\epsilon_0 m_e c^2} = 2.818 \times 10^{-15}\,\text{m},$$

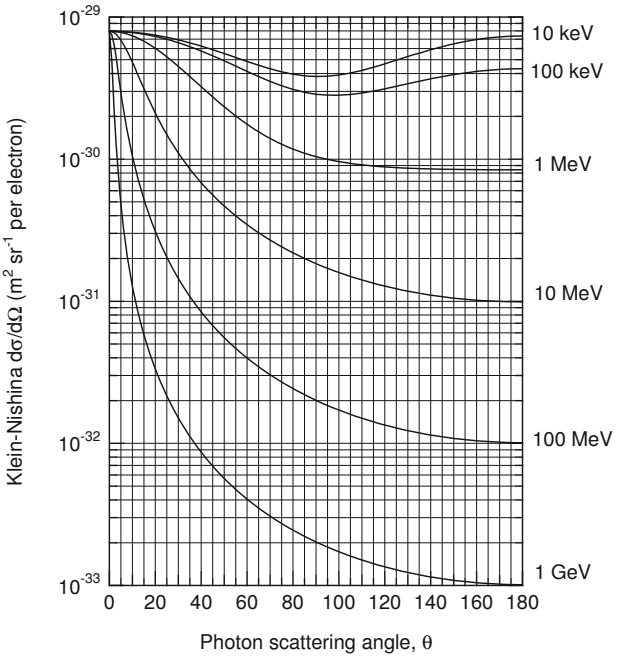


Fig. 15.6 Differential cross section for Compton scattering of unpolarized photons from a free electron, calculated from Eq. 15.16. The incident photon energy for each curve is shown on the right

is the *classical radius* of the electron. The cross section is plotted in Fig. 15.6. It is peaked in the forward direction at high energies. As $x \to 0$ (long wavelengths or low energy) it approaches

$$\frac{d\sigma_C}{d\Omega} = \frac{r_e^2(1 + \cos^2\theta)}{2}, \tag{15.17}$$

which is symmetric about 90°.

Equation 15.16 can be integrated over all angles to obtain the total Compton cross section for a single electron:

$$\sigma_C = 2\pi r_e^2\left[\frac{1 + x}{x^2}\left(\frac{2(1 + x)}{1 + 2x} - \frac{\ln(1 + 2x)}{x}\right) + \frac{\ln(1 + 2x)}{2x} - \frac{1 + 3x}{(1 + 2x)^2}\right]. \tag{15.18}$$

As $x \to 0$, this approaches

$$\sigma_C \to \frac{8\pi r_e^2}{3} = 6.652 \times 10^{-29}\,\text{m}^2. \tag{15.19}$$

Figure 15.7 shows σ_C as a function of energy.

The classical analog of Compton scattering is *Thomson scattering* of an electromagnetic wave by a free electron. The electron experiences the electric field $\mathbf{E}$ of an incident plane electromagnetic wave and therefore has an acceleration $-e\mathbf{E}/m$. Accelerated charges radiate electromagnetic waves, and the energy radiated in different directions can be

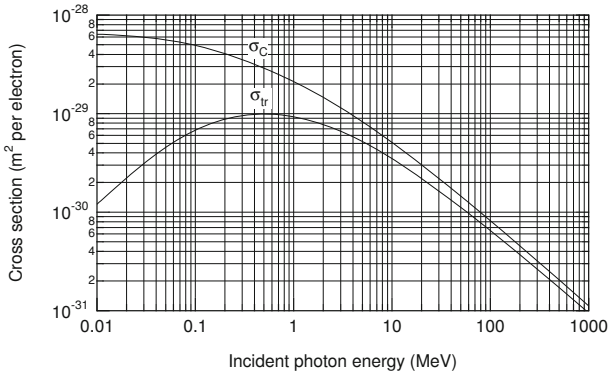


Fig. 15.7 The total cross section σ_C for Compton scattering by a single electron and the cross section for energy transfer $\sigma_{\rm tr} = f_C\sigma_C$

calculated, giving Eqs. 15.17 and 15.19. (See, for example, Jackson 1999, Chap. 14.) In the classical limit of low photon energies and momenta, the energy of the recoil electron is negligible.

15.4.3 Incoherent Scattering

The Compton cross section is for a single electron. For an atom containing Z electrons, the maximum value of the incoherent cross section occurs if all Z electrons take part in the Compton scattering:

$$\sigma_{\rm incoh} \le Z\sigma_C.$$

For carbon $Z\sigma_C = 4.0\times10^{-28}$ m^2. This value is approached by $\sigma_{\rm incoh}$ near 10 keV. At low energies $\sigma_{\rm incoh}$ falls below this maximum value because the electrons are bound and not at rest. This falloff can be seen in Fig. 15.2. It is appreciable for energies as high as 7–8 keV, even though the K-shell binding energy in carbon is only 283 eV. The electron motion and binding in the target atom also cause a small spread in the energy of the scattered photons (Carlsson et al. 1982).

Departures of the angular distribution and incoherent cross section from Z times the Klein–Nishina formula are discussed by Hubbell et al. (1975) and by Jackson and Hawkes (1981).

15.4.4 Energy Transferred to the Electron

We will need to know the average energy transferred to an electron in a Compton scattering. Equation 15.15 gives the electron kinetic energy as a function of photon scattering angle. The transfer cross section is defined to be

$$\sigma_{\rm tr} = \int_0^\pi \frac{d\sigma_C}{d\Omega}\frac{T(\theta)}{h\nu_0}2\pi\sin\theta\,d\theta = f_C\sigma_C. \qquad (15.20)$$

This can be integrated. The result is (see Attix 1986, p. 134)

$$\sigma_{\rm tr} = 2\pi r_e^2\left[\frac{2(1+x)^2}{x^2(1+2x)} - \frac{1+3x}{(1+2x)^2} - \frac{(1+x)(2x^2-2x-1)}{x^2(1+2x)^2} - \frac{4x^2}{3(1+2x)^3} - \left(\frac{1+x}{x^3} - \frac{1}{2x} + \frac{1}{2x^3}\right)\ln(1+2x)\right]. \qquad (15.21)$$

This quantity is also plotted in Fig. 15.7. Equation 15.21 is a rather nasty equation to evaluate, particularly at low energies, because many of the terms nearly cancel.

15.5 Coherent Scattering

A photon can also scatter elastically from an atom, with none of the electrons leaving their energy levels. This (γ,γ) process is called *coherent scattering* (sometimes called *Rayleigh scattering*), and its cross section is $\sigma_{\rm coh}$. The entire atom recoils; if one substitutes the atomic mass in Eqs. 15.14 and 15.15, one finds that the atomic recoil kinetic energy is negligible.

The primary mechanism for coherent scattering is the oscillation of the electron cloud in the atom in response to the electric field of the incident photons. There are small contributions to the scattering from nuclear processes. The cross section can be calculated classically as an extension of Thomson scattering, or it can be done using various degrees of quantum-mechanical sophistication (Kissel et al. 1980).

The coherent cross section is peaked in the forward direction because of interference effects between electromagnetic waves scattered by various parts of the electron cloud. The peak is narrower for elements of lower atomic number and for higher energies. Coherent and incoherent scattering cross sections are shown in Fig. 15.8 for 100-keV photons scattering from carbon, calcium and lead. Also shown for comparison is $Z(d\sigma/d\Omega)_{KN}$.

If the wavelength of the incident photon is large compared to the size of the atom, then all Z electrons behave like a single particle with charge $-Ze$ and mass Zm_e. The classical radius is replaced by $Z^2e^2/4\pi\epsilon_0 Zm_ec^2$. From Eqs. 15.17 and 15.19, one can see that the cross section in this limit is Z^2 times the single-electron value: $Z^2\sigma_C$. The limiting value for carbon is 2.39×10^{-27} m^2, which can be compared to the low energy limit for $\sigma_{\rm coh}$ in Fig. 15.2.

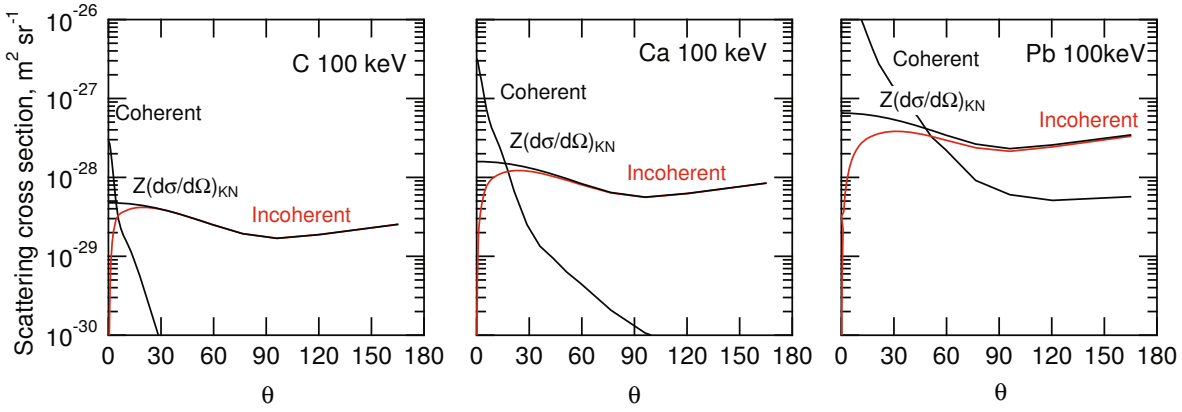


Fig. 15.8 The coherent and incoherent differential cross sections as a function of angle for 100-keV photons scattering from carbon, calcium, and lead. Calculated from Hubbell et al. (1975)

15.6 Pair Production

A photon with energy above 1.02 MeV can produce a particle–antiparticle pair: a negative electron and a positron. Conservation of energy requires that

$$h\nu_0 = \underbrace{T_- + m_e c^2}_{\text{electron}} + \underbrace{T_+ + m_e c^2}_{\text{positron}} = T_+ + T_- + 2m_e c^2. \tag{15.22}$$

Since the rest energy ($m_e c^2$) of an electron or positron is 0.51 MeV, pair production is energetically impossible for photons below $2m_e c^2 = 1.02$ MeV.

One can show, using $h\nu_0 = pc$ for the photon, that momentum is not conserved by the positron and electron if Eq. 15.22 is satisfied. Pair production always takes place in the Coulomb field of another particle (usually a nucleus) that recoils to conserve momentum. The nucleus has a large mass, so its kinetic energy $p^2/2m$ is small compared to the terms in Eq. 15.22. The cross section for this (γ, e^+e^-) reaction involving the nucleus is κ_n.

An additional contrbution to the cross section, κ_e, arises when the incident photon energy exceeds $4m_e c^2 = 2.04$ MeV, the threshold for pair production in which a free electron (rather than a nucleus) recoils to conserve momentum. Because ionization and free-electron pair production are $(\gamma, e^-e^-e^+)$ processes, this is usually called *triplet production*. Extensive data are given in Hubbell et al. (1980).

The cross section for both processes is $\kappa = \kappa_n + \kappa_e$. The energy dependence of κ can be seen in Figs. 15.1 and 15.2.

15.7 The Photon Attenuation Coefficient

Consider the arrangement shown in Fig. 15.9a, in which a beam of photons is collimated so that a narrow beam strikes a detector. A scattering material is then introduced in the beam.

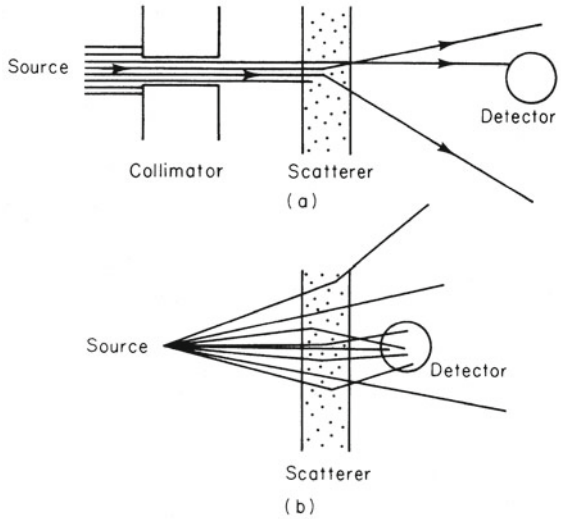


Fig. 15.9 Measurements with narrow-beam geometry (**a**) and broad-beam geometry (**b**)

Some of the photons pass through the material without interaction. Others are scattered. Still others disappear because of photoelectric effect or pair-production interactions. If we measure only photons that remain in the unscattered beam, the loss of photons is called *attenuation* of the beam. Attenuation includes both scattering and absorption. We record as still belonging to the beam only photons that did not interact; they still travel in the forward direction with the original energy. This is called a *narrow-beam geometry* measurement. It is an idealization, because photons that undergo Compton or coherent scattering through a small angle can still strike the detector. Figure 15.9b shows a source, scatterer, and detector geometry that is much more difficult to interpret. In this case photons that are initially traveling in a different direction are scattered into the detector. These are called *broad-beam geometry* experiments.

In narrow-beam geometry, the total cross section is related to the total number of particles that have interacted in the scatterer. Let N be the number of particles that have not undergone any interaction in passing through scattering material of thickness z. We saw in Sect. 14.5 that the number of particles that have not interacted decreases in thickness dz by

$$dN = -\frac{\sigma_{\text{tot}} N_A \rho}{A} N\, dz,$$

so that

$$\frac{dN}{dz} = -\mu_{\text{atten}} N,$$

where

$$\mu_{\text{atten}} = \frac{N_A \rho \sigma_{\text{tot}}}{A}. \tag{15.23}$$

In these equations ρ is the mass density of the target material and A is its atomic weight.[2] The number of particles that have undergone no interaction decays exponentially with distance:

$$N(z) = N_0 e^{-\mu_{\text{atten}} z}. \tag{15.24}$$

The quantity μ_{atten} is called the *total linear attenuation coefficient*.

In a broad-beam geometry configuration the total number of photons reaching the detector includes secondary photons and is larger than the value given by Eq. 15.24.

The units in Eqs. 15.23 and 15.24 are worth discussing. Avogadro's number is 6.022×10^{23} entities mol^{-1}. If the density ρ is in kg m^{-3} and σ_{tot} is in m^2, then A must be expressed in kg mol^{-1} and μ_{atten} is in m^{-1}. On the other hand, it is possible to express ρ in g cm^{-3}, σ_{tot} in cm^2, and A in g mol^{-1}, so that μ_{atten} is in cm^{-1}. As an example, consider carbon, for which $A = 12.011 \times 10^{-3}$ kg mol^{-1} $= 12.011$ g mol^{-1}. If $\sigma_{\text{tot}} = 1.269 \times 10^{-28}$ m^2 atom^{-1} $= 1.269 \times 10^{-24}$ cm^2 atom^{-1}, then either

$$\begin{aligned}\mu_{\text{atten}} &= \frac{(6.022 \times 10^{23}\,\text{atom mol}^{-1})(2.000 \times 10^{3}\,\text{kg m}^{-3})}{12.011 \times 10^{-3}\,\text{kg mol}^{-1}} \\ &\quad \times (1.269 \times 10^{-28}\,\text{m}^2\,\text{atom}^{-1}) \\ &= 12.7\,\text{m}^{-1}\end{aligned}$$

or

$$\begin{aligned}\mu_{\text{atten}} &= \left(\frac{(6.022 \times 10^{23}\,\text{atom mol}^{-1})(2.000\,\text{g cm}^{-3})}{12.011\,\text{g mol}^{-1}}\right) \\ &\quad \times (1.269 \times 10^{-24}\,\text{cm}^2\,\text{atom}^{-1}) \\ &= 0.127\,\text{cm}^{-1}.\end{aligned}$$

[2] The atomic weight is potentially confusing. Sometimes A has no units (as in labeling an nuclear isotope), sometimes it is in grams per mole, and sometimes it is in kilograms per mole).

The total cross section for photon interactions is

$$\sigma_{\text{tot}} = \sigma_{\text{coh}} + \sigma_{\text{incoh}} + \tau + \kappa. \tag{15.25a}$$

In many situations the coherently scattered photons cannot be distinguished from those unscattered, and σ_{coh} should not be included:

$$\sigma_{\text{tot}} = \sigma_{\text{incoh}} + \tau + \kappa. \tag{15.25b}$$

Tables usually include total cross sections and attenuation coefficients both with and without coherent scattering.

It is possible to regroup the terms in Eqs. 15.23 and 15.24 in a slightly different way:

$$dN = -N \frac{N_A \sigma_{\text{tot}}}{A} \rho dz.$$

The quantity $N_A \sigma_{\text{tot}}/A$ is the *mass attenuation coefficient*, μ_{atten}/ρ (m^2 kg^{-1}):

$$\frac{\mu_{\text{atten}}}{\rho} = \frac{N_A \sigma_{\text{tot}}}{A}. \tag{15.26}$$

The exponential attenuation is then

$$N(\rho z) = N_0 e^{-(\mu_{\text{atten}}/\rho)(\rho z)}. \tag{15.27}$$

The mass attenuation coefficient has the advantage of being independent of the density of the target material, which is particularly useful if the target is a gas. It has an additional advantage if Compton scattering is the dominant interaction. If $\sigma_{\text{tot}} = Z\sigma_C$, then

$$\frac{\mu_{\text{atten}}}{\rho} = \frac{Z\sigma_C N_A}{A}.$$

Since Z/A is nearly $1/2$ for all elements except hydrogen, this quantity changes very little throughout the periodic table. This constancy is not true for the photoelectric effect or pair production. Figure 15.10 plots the mass attenuation coefficient vs energy for three substances spanning the periodic table. It is nearly independent of Z around 1 MeV where Compton scattering is dominant. The K and L absorption edges can be seen for lead; for the lighter elements they are below 10 keV. Figure 15.11 shows the contributions to μ_{atten}/ρ for air from the photoelectric effect, incoherent scattering, and pair production. Tables of mass attenuation coefficients are provided by the National Institute of Standards and Technology (NIST) at http://www.nist.gov/pml/data/xcom/index.cfm.

15.8 Compounds and Mixtures

The usual procedure for dealing with mixtures and compounds is to assume that each atom scatters independently.

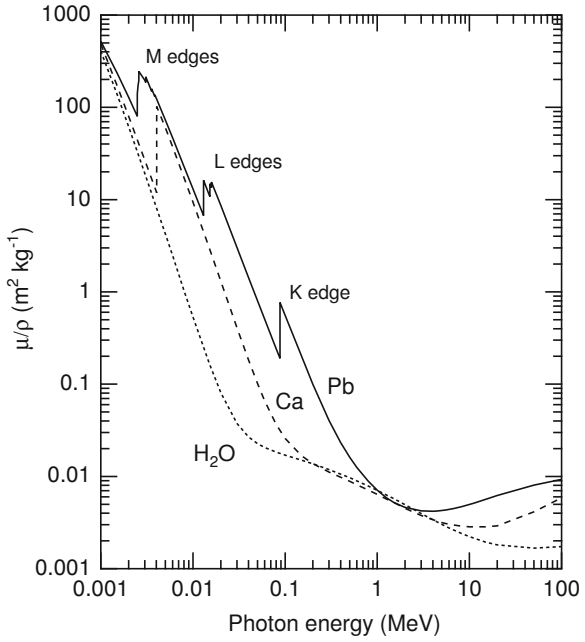


Fig. 15.10 Mass attenuation coefficient vs. energy for lead, calcium, and water. Near 1 MeV the mass attenuation coefficient is nearly independent of Z. (Plotted from data provided by NIST: http://www.nist.gov/pml/data/xcom/index.cfm)

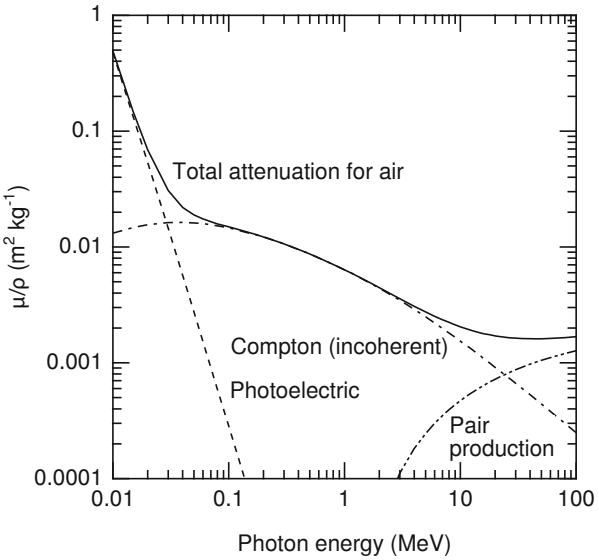


Fig. 15.11 Mass attenuation coefficient vs energy for air. (Plotted from data provided by NIST: http://www.nist.gov/pml/data/xcom/index.cfm)

If the cross section for element i summed over all the interaction processes of interest is denoted by σ_i, then Eq. 14.19 is replaced by

$$\frac{\overline{n}}{N} = \sum_i (N_T)_i \sigma_i = \left(\sum_i (N_{TV})_i \sigma_i \right) dz, \qquad (15.28)$$

where $(N_T)_i$ is the number of target atoms of species i *per unit projected area* of the target and $(N_{TV})_i$ is the number of target atoms *per unit volume*. The sum is taken over all elements in the compound or mixture.

It is possible to replace the sum by the product of the cross section per molecule multiplied by the number of molecules per unit volume. The cross section per molecule is the sum of the cross sections for all the atoms in the molecule. To see that this is so, note that a volume of scatterer V contains a total mass $M = \rho V$. The mass of each element is M_i and the mass fraction is $w_i = M_i/M$. The total number of atoms of species i in volume V is the number of moles times Avogadro's number:

$$(N_{TV})_i = \frac{M_i N_A}{A_i V} = \frac{w_i}{A_i} \rho N_A. \qquad (15.29)$$

The mass fraction of element i in a compound containing a_i atoms per molecule with atomic mass A_i is

$$w_i = \frac{a_i A_i}{A_{\text{mol}}}, \qquad (15.30)$$

where A_{mol} is the molecular weight. Therefore

$$\sum_i (N_{TV})\sigma_i = \left(\sum_i \frac{a_i \sigma_i}{A_{\text{mol}}} \right) \rho N_A$$
$$= \left(\sum_i a_i \sigma_i \right) \frac{\rho N_A}{A_{\text{mol}}} = \sigma_{\text{mol}} (N_{TV})_{\text{mol}}. \qquad (15.31)$$

The factor $(N_{TV})_{\text{mol}} = \rho N_A / A_{\text{mol}}$ is the number of molecules per unit volume. When a target entity (molecule) consists of a collection of subentities (atoms), we can say that in this approximation (all subentities interacting independently), the cross section per entity is the sum of the cross sections for each subentity. For example, for the molecule CH_4, the total molecular cross section is $\sigma_{\text{carbon}} + 4\sigma_{\text{hydrogen}}$ and the molecular weight is $[(4 \times 1) + 12 = 16] \times 10^{-3}$ kg mol^{-1}.

15.9 Deexcitation of Atoms

After the photoelectric effect, Compton scattering, or triplet production, an atom is left with a hole in some electron shell. An atom can be left in a similar state when an electron is knocked out by a passing charged particle or by certain transformations in the atomic nucleus that are discussed in Chap. 17.

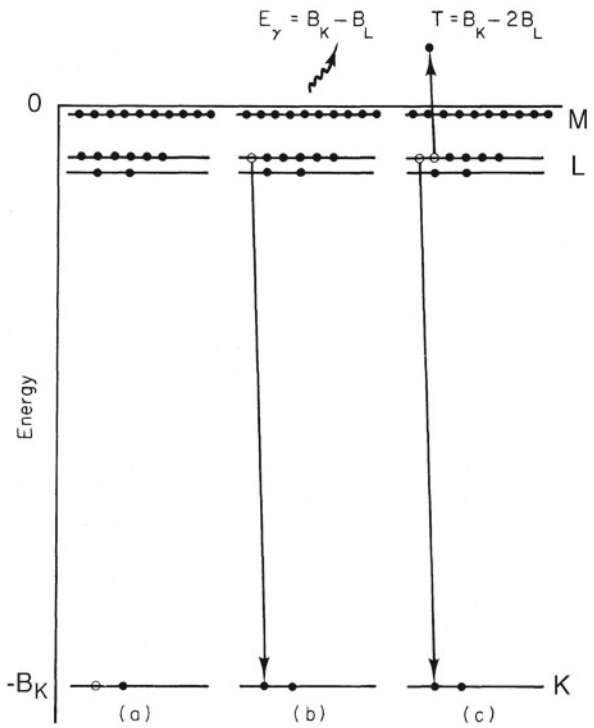


Fig. 15.12 Two possible mechanisms for the deexcitation of an atom with a hole in the K shell. **a** The atom with the hole in the K shell. **b** An electron has moved from the L shell to the K shell with emission of a photon of energy $B_K - B_L$. **c** An electron has moved from the L shell to the K shell. The energy liberated is taken by another electron from the L shell, which emerges with energy $B_K - 2B_L$. This electron is called an Auger electron

The hole in the shell can be filled by two competing processes: a *radiative transition*, in which a photon is emitted as an electron falls into the hole from a higher level, or a *nonradiative* or *radiationless transition*, such as the emission of an *Auger electron* from a higher level as a second electron falls from a higher level to fill the hole. Both processes are shown in Fig. 15.12. In the radiative transition, the energy of the photon is equal to the difference in binding energies of the two levels. For the example of Fig. 15.12b, the photon energy is $B_K - B_L$. The emission of an L-shell Auger electron is shown in Fig. 15.12c: its energy is $T = (B_K - B_L) - B_L = B_K - 2B_L$. Table 15.2 shows the energy changes that occur after a hole is created in an atom by photoelectric excitation. It is worth understanding each table entry in detail. Two different paths for deexcitation are shown: one for photon emission and one for ejection of an Auger electron. The sum of the photon, electron, and atomic excitation energies does not change.

The photon that is emitted is called a *characteristic photon* or a *fluorescence photon*. Its energy is given by the difference of two electron energy levels in the atom. There is an historical nomenclature for these photons. Because a hole moving to larger values of n corresponds to a decrease in the total energy of an atom, it is customary to draw the energy levels for holes instead of electrons, as in Fig. 15.13. Transitions in which the hole is initially in the $n = 1$ state give rise to the K series of x rays, those in which the initial hole is in the $n = 2$ state give rise to the L series, and so on. Greek letters (and their subscripts) are used to denote the shell (and subshell) of the final hole. The transitions shown in Fig. 15.13 are consistent with certain selection rules which can be derived using quantum theory:

$$\Delta l = \pm 1, \qquad \Delta j = 0, \pm 1. \tag{15.32}$$

We saw in Eqs. 15.1 and 15.2 that the position of a level could be estimated by the Bohr formula corrected for screening. The energy of the K_α line—which depends on screening for both the initial ($n = 2$) and final ($n = 1$) values of $n$can be fitted empirically by

$$E_{K_\alpha} = \left(\tfrac{3}{4}\right)(13.6)(Z-1)^2. \tag{15.33}$$

After creation of a hole in the K shell, it is random whether the atom deexcites by emitting a photon or an Auger electron. The probability of photon emission is called the *fluorescence yield*, W_K. The Auger yield is $A_K = 1 - W_K$. For a vacancy in the L or higher shells, one must consider the fluorescence yield for each subshell, defined as the number of photons emitted with an initial state corresponding to a hole in a subshell, divided by the number of holes in that subshell. The situation is further complicated by the fact that radiationless transitions can take place within the subshell, thereby altering the number of vacancies in each subshell. These are called *Coster–Kronig transitions,* and they are also accompanied by the emission of an electron. For example, a hole in the L_I shell can be filled by an electron from the L_III shell with the ejection of an M-shell electron. A *super–Coster–Kronig transition* involves electrons all within the same shell, for example, a hole in the M_I shell filled by an electron from the M_II shell with the ejection of an electron from the M_IV shell.

One can define an average fluorescence yield $\overline{W}_L$, $\overline{W}_M$, etc. for each shell, but it is not a fundamental property of the atom, since it depends on the vacancy distribution in the subshells. Bambynek et al. (1972) review the physics of atomic deexcitations and present theoretical and experimental data for the fundamental parameters. They show that $\overline{W}_L$ is less sensitive to the initial vacancy distribution than one might expect, because of the rapid changes in hole distribution caused by the Coster–Kronig transitions. Hubbell et al. (1994) provide a more recent review. Figure 15.14 shows values for W_K, $\overline{W}_L$, and $\overline{W}_M$ as a function of Z. One can see from this figure that radiationless transitions are much more important

Table 15.2 Energy changes in the photoelectric effect and in subsequent deexcitation

Process	Total photon energy	Total electron energy	Atom excitation energy	Sum
Before photon strikes atom	$h\nu$	0	0	$h\nu$
After photoelectron is ejected (Fig. 15.12)	0	$h\nu - B_K$	B_K	$h\nu$
Case 1: Deexcitation by the emission of a K and an L photon				
Emission of K fluorescence photon (Fig. 15.12b)	$B_K - B_L$	$h\nu - B_K$	B_L	$h\nu$
Emission of L fluorescence photon	$B_K - B_L, B_L$	$h\nu - B_K$	0	$h\nu$
Case 2: Deexcitation by emission of an Auger electron from the L shell				
Emission of Auger electron (Fig. 15.12c)	0	$h\nu - B_K, B_K - 2B_L$	$2B_L$	$h\nu$
First L-shell hole filled by fluorescence	B_L	$h\nu - B_K, B_K - 2B_L$	B_L	$h\nu$
Second L-shell hole filled by fluorescence	B_L, B_L	$h\nu - B_K, B_K - 2B_L$	0	$h\nu$

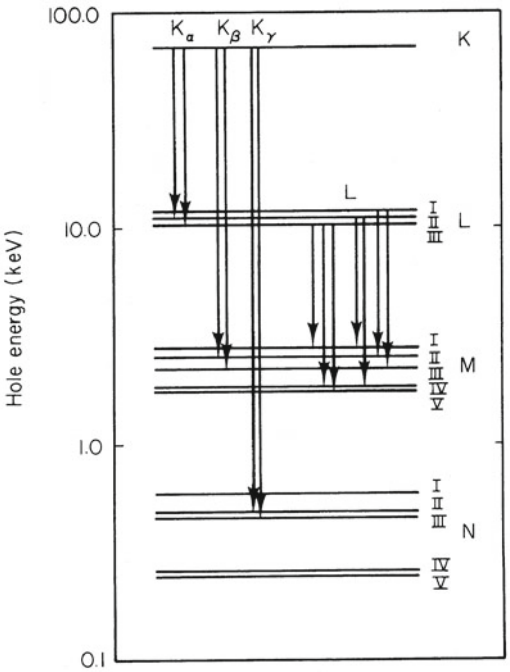


Fig. 15.13 Energy-level diagram for holes in tungsten, and some of the x-ray transitions

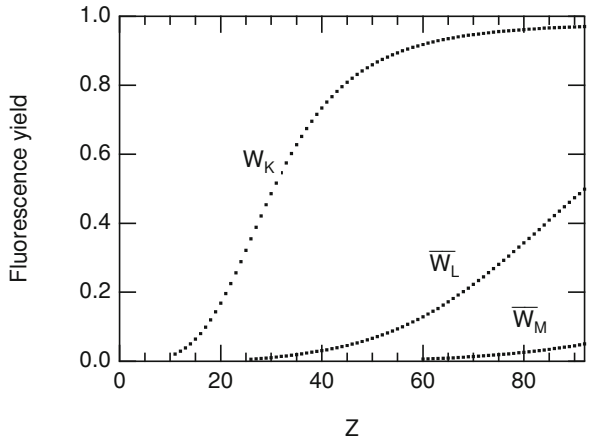


Fig. 15.14 Fluorescence yields for K-, L-, and M-shell vacancies as a function of atomic number Z. Points are from Table 8 of Hubbell et al. (1994)

(the fluorescent yield is much smaller) for the L shell than for the K shell. They are nearly the sole process for higher shells. The deexcitation is often called the *Auger cascade*.

The Auger cascade produces many vacancies in the outer shells of the atom, and some of these may be filled by electrons from other atoms in the same molecule. This process can break molecular bonds. Moreover, the Auger and Coster–Kronig electrons from the higher shells can be quite numerous. They are of such low energy that they travel only a fraction of a cell diameter. This must be taken into consideration when estimating cell damage from radiation. The effect of radiationless transitions is quite important for certain radioactive isotopes that are administered to a patient, particularly when they are bound to the cellular DNA. We will discuss them further in Chap. 17.

15.10 Energy Transfer from Photons to Electrons

The attenuation coefficient gives the rate at which photons interact and leave the primary beam as they pass through the material. If a beam of monoenergetic photons of energy $E = h\nu$ and particle fluence Φ passes through a thin layer dx of material, the number of particles per unit area that interact in the layer, $-d\Phi$, is proportional to the fluence and the attenuation coefficient: $-d\Phi = \Phi\mu_{\text{atten}}\,dx$. The energy fluence is $\Psi = h\nu\Phi$. The reduction of energy fluence of unscattered photons is $-d\Psi = -h\nu\,d\Phi$. For a thick absorber we can say that the number of unscattered photons and the energy carried by unscattered photons decay as

$$\Phi_{\text{unscatt}} = \Phi_0 e^{-\mu_{\text{atten}}x}, \qquad \Psi_{\text{unscatt}} = \Psi_0 e^{-\mu_{\text{atten}}x}. \quad (15.34)$$

The total energy flow is much more complicated. Every photon that interacts contributes to a pool of secondary photons of lower energy and to a pool of electrons and positrons. Figure 15.15 shows the processes by which energy can move between the photon pool and the electron–positron pool.

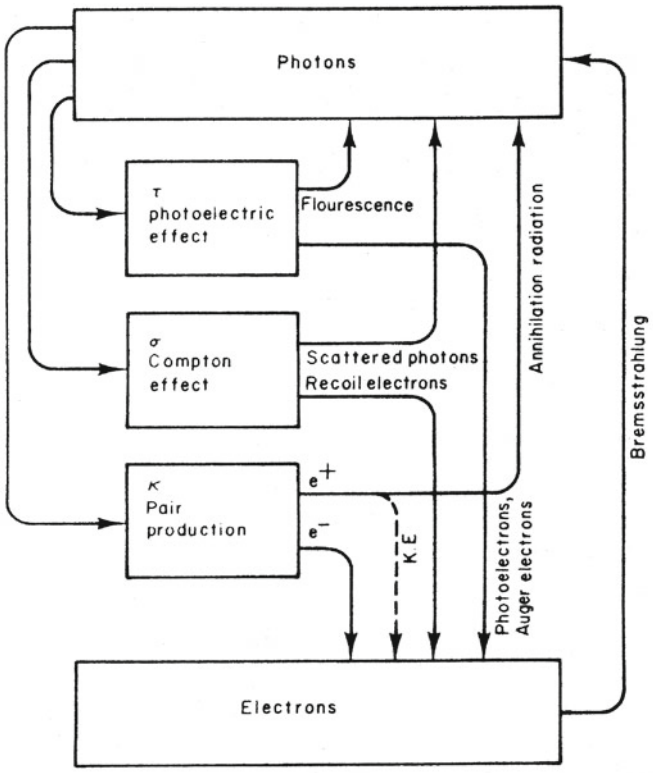


Fig. 15.15 Routes for the transfer of energy between photons and electrons. The four lines going to the lower box represent electrons; the others represent photons

Energy that remains as secondary photons, such as those resulting from fluorescence or Compton scattering, can travel long distances from the site of the initial interaction. Ionizing particles (photoelectrons, Auger electrons, Compton recoil electrons, and electron–positron pairs) usually lose their energy relatively close to where they were produced. We will see in Sect. 15.13 that for primary photons below 10 MeV, the mean free path of the secondary electrons is usually short compared to that of the photons. Damage to cells is caused by local ionization or excitation of atoms and molecules. This damage is done much more efficiently by the electrons than by the photons.

The *mass energy transfer coefficient* $\mu_{\rm tr}/\rho$ is a measure of the energy transferred from primary photons to charged particles in the interaction. If N monoenergetic photons of energy E strike a thin absorber of thickness dx, the amount of energy transferred to charged particles is defined to be

$$\overline{dE_{\rm tr}} = NE\,\mu_{\rm tr}\,dx,$$

so that

$$\frac{\mu_{\rm tr}}{\rho} = \frac{1}{\rho NE}\frac{\overline{dE_{\rm tr}}}{dx}. \tag{15.35}$$

We can relate $\mu_{\rm tr}$ to $\mu_{\rm atten}$. Suppose the material contains a single atomic species and that f_i is the average fraction of the photon energy that is transferred to charged particles in process i. (Different values of i denote the photoelectric effect, incoherent scattering, coherent scattering, and pair production.) Multiplying the number of photons that interact by their energy E and by f_i gives the energy transferred. Comparison with Eq. 15.23 shows that

$$\frac{\mu_{\rm tr}}{\rho} = \frac{N_A}{A}\sum_i f_i\sigma_i. \tag{15.36}$$

Coherent scattering produces no charged particles, so

$$\frac{\mu_{\rm tr}}{\rho} = \frac{N_A}{A}\left(\tau f_\tau + \sigma_{\rm incoh} f_C + \kappa f_\kappa\right). \tag{15.37}$$

Fraction f_τ for the photoelectric effect can be written in terms of δ, the average energy emitted as fluorescence radiation per photon absorbed. The quantity δ is calculated taking into account all atomic energy levels and the fluorescence yield for each shell. The average electron energy is $h\nu - \delta$, so

$$f_\tau = \frac{h\nu - \delta}{h\nu} = 1 - \frac{\delta}{h\nu}. \tag{15.38}$$

We can estimate δ by assuming that τ_K is the dominant term in the photoelectric cross section, Eq. 15.7. The probability that the hole in the K shell is filled by fluorescence is W_K. The energy of the photon is $B_K - B_L$ or $B_K - B_M$, and so on. A hole is left in a higher shell, which may decay by photon or Auger-electron emission. The latter is much more likely for the higher shells. Therefore, nearly all of the photons emitted have energy $B_K - B_L$, so we have the approximate relationship

$$\delta \approx W_K\,(B_K - B_L). \tag{15.39}$$

For Compton scattering, the fraction of the energy transferred to electrons is implicit in Eqs. 15.20 and 15.21. The transfer cross section $f_C\sigma_C$, is plotted in Fig. 15.7.

For pair production, energy in excess of $2m_ec^2$ becomes kinetic energy of the electron and positron. The fraction is

$$f_\kappa = 1 - \frac{2m_ec^2}{h\nu}. \tag{15.40}$$

All of these can be combined to estimate $\mu_{\rm tr}$.

We will see in Sect. 15.11 that charged particles traveling through material can radiate photons through a process known as *bremsstrahlung*. The *mass energy-absorption coefficient* $\mu_{\rm en}$ takes this additional effect into account. It is defined as

$$\frac{\mu_{\rm en}}{\rho} = \frac{\mu_{\rm tr}}{\rho}(1-g), \tag{15.41}$$

where g is the fraction of the energy of secondary electrons that is converted back into photons by bremsstrahlung in the

material. The fraction of the energy converted to photons depends on the energy of the electrons. Since the average electron energy is different in the three processes, we can write (again assuming noninteracting atoms in the target material)

$$\frac{\mu_{\text{en}}}{\rho} = \frac{N_A}{A} \sum_i f_i \sigma_i (1 - g_i). \tag{15.42}$$

In addition to bremsstrahlung, there is another process that converts charged-particle energy back into photon energy. Positrons usually come to rest and then combine with an electron to produce *annihilation radiation*. Occasionally, a positron annihilates while it is still in flight, thereby reducing the amount of positron kinetic energy that is available to excite atoms. While not mentioned in the International Commission on Radiation Units and Measurements (ICRU) Report 33 (1980) definition, this effect has been included in the tabulations of μ_{en}/ρ by Hubbell (1982). Seltzer (1993) reviews the calculation of μ_{tr}/ρ and μ_{en}/ρ.

The energy-transfer and energy-absorption coefficients differ appreciably when the kinetic energies of the secondary charged particles are comparable to their rest energies, particularly in high-Z materials. The ratio $\mu_{\text{en}}/\mu_{\text{tr}}$ for carbon falls from 1.00 when $h\nu = 0.5$ MeV to 0.96 when $h\nu = 10$ MeV. For lead at the same energies it is 0.97 and 0.74. Tables are given by Attix (1986). The difference between the attenuation and the energy-absorption coefficients is greatest at energies where Compton scattering predominates, since the scattered photon carries away a great deal of energy. Figure 15.16 compares μ_{atten}/ρ and μ_{en}/ρ for water.

Attenuation and energy-transfer coefficients are found in Hubbell and Seltzer (1996). These tables are also available on the web at http://www.nist.gov/pml/data/xraycoef/index.cfm . Another data source is a computer program provided by Boone and Chavez (1996).

We will return to these concepts in Sect. 15.15 to discuss the dose, or energy per unit mass deposited in tissue or a detector. First, we must discuss energy loss by charged particles.

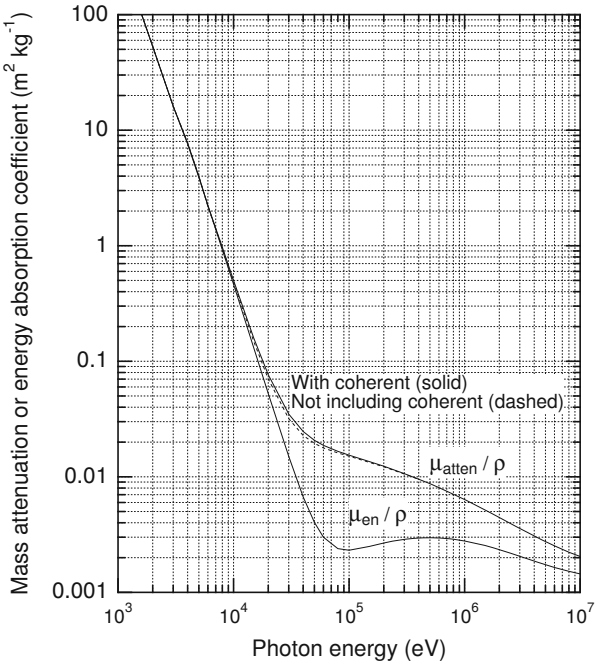


Fig. 15.16 Coherent and incoherent attenuation coefficients and the mass energy absorption coefficient for water. (Plotted from data in Hubbell 1982)

15.11 Charged-Particle Stopping Power

The behavior of a particle with charge ze and mass M_1 passing through material is very different from the behavior of a photon. When a photon interacts, it usually disappears: either being completely absorbed as in the photoelectric effect or pair production, or being replaced by a photon of different energy traveling in a different direction as in Compton scattering. The exception is coherent scattering, where a photon of the same energy travels in a different direction. A charged particle has a much larger interaction cross section than a photon—typically 10^4–10^5 times as large. Therefore the "unattenuated" charged-particle beam falls to zero almost immediately.

Each interaction usually causes only a slight decrease in the particle's energy, and it is convenient to follow the charged particle along its path. Figure 15.27 shows the tracks of some α particles (helium nuclei) in photographic emulsion. The spacing of the fiducial marks at the bottom is 10 μm. Each particle entered at the bottom of the figure and stopped near the top. Figures 15.28 and 15.29 show the tracks of electrons. Figure 15.28 is in photographic emulsion, while Fig. 15.29 is in water. We will be discussing these tracks in detail in Sect. 15.14. For now, we need only note that the α-particle tracks are fairly straight, with some deviation near the end of the track. The electrons, being lighter, show considerably more scattering.[3]

[3] This distinction between photons and charged particles represents two extremes on a continuum, and we must be careful not to adhere to the distinction too rigidly. A photon may be coherently scattered through a small angle with no loss of energy, while a charged particle may occasionally lose so much energy that it can no longer be followed.

It is convenient to speak of how much energy the charged particle loses per unit path length, the *stopping power*, and its *range*—roughly, the total distance it travels before losing all its energy. The stopping power is the expectation value of the amount of kinetic energy T lost by the projectile per unit path length. The term "power" is historical. The units of stopping power are J m^{-1} not J s^{-1}. The *mass stopping power* is the stopping power divided by the density of the stopping material and is analogous to the mass attenuation coefficient (often we will say stopping power when we actually mean mass stopping power):

$$S = -\frac{dT}{dx}, \qquad \frac{S}{\rho} = -\frac{1}{\rho}\frac{dT}{dx}. \tag{15.43}$$

In the energy-loss process, the projectile interacts with the target atom. The projectile loses energy W, which becomes kinetic energy or internal excitation energy of the atom. Internal excitation may include ionization of the atom. If the atoms in the material act independently, the cross section per atom for an interaction that results in an energy loss between W and $W + dW$ is $(d\sigma/dW)dW$. The results of Sect. 14.5 can be used to write the probability that a projectile loses an amount of energy between W and $W + dW$ while traversing a thickness dx of a substance of atomic mass number A and density ρ:

$$\text{(probability)} = \frac{\overline{n}}{N} = \frac{N_A\rho}{A}\,dx\,\frac{d\sigma}{dW}\,dW. \tag{15.44}$$

The average total energy loss is

$$dT = \frac{N_A\rho}{A}dx\int_0^{W_{\max}} W\,\frac{d\sigma}{dW}\,dW, \tag{15.45}$$

and the mass stopping power is

$$\frac{S}{\rho} = \frac{N_A}{A}\int_0^{W_{\max}} W\,\frac{d\sigma}{dW}\,dW. \tag{15.46}$$

The integral is sometimes called the *stopping cross section* ϵ. Its units are J m^2.

Figure 15.17 shows the mass stopping power for protons, α particles ($z = 2$, $M_\alpha = 4M_p$), and electrons and positrons ($z = \pm 1$) in carbon as a function of energy. We see a number of features of these curves:

1. All of the stopping power curves have roughly the same shape, rising with increasing energy, reaching a peak, and then falling. (The electron and positron curves peak at a lower energy than is shown in the figure.)
2. There is a region where the stopping power falls approximately as $1/T$.
3. At still higher energies the curves rise again. This can be seen for the electron and positron curves above 1 MeV. Similar increases occur in the proton and α-particle curves at higher energies than are plotted here.

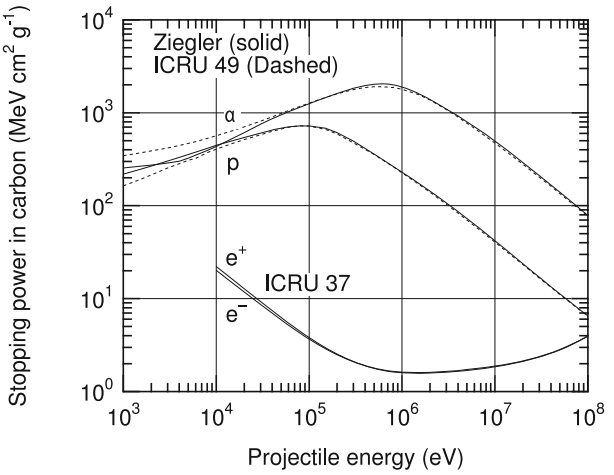


Fig. 15.17 The mass stopping power for electrons (e^-), positrons (e^+), protons (p), and α particles in carbon vs kinetic energy. Plotted from data in ICRU 37, ICRU 49, and the program SRIM (Stopping and Range of Ions in Matter), version 96.04 (see Ziegleret al. 1985)

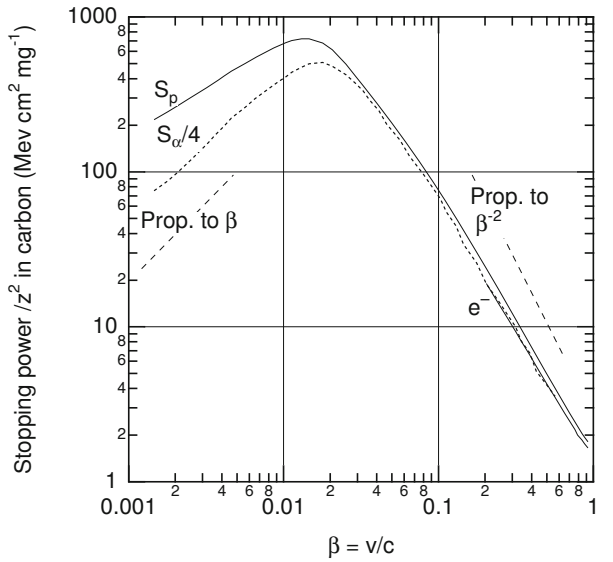


Fig. 15.18 The scaled stopping power. The stopping power in carbon is plotted vs the speed $\beta = v/c$ of the projectile for electrons, protons, and α particles. The α-particle stopping power has been divided by 4, the square of the particle charge z. Proton and α-particle stopping powers are from the program SRIM (see caption for Fig. 15.17). The electron stopping power is from ICRU Report 37 (1984)

The similarities suggest that the stopping power curves for different projectiles may be related. Figure 15.18 shows the similarities more clearly. The stopping powers are plotted vs particle speed in the form $\beta = v/c$. At low energies ($\beta \ll 1$) β is related to kinetic energy by

$$\beta = \left(\frac{2T}{Mc^2}\right)^{1/2}. \tag{15.47}$$

For larger values of β, the relativistically correct expression

$$\beta = \left[1 - \left(\frac{1}{T/Mc^2 + 1}\right)^2\right]^{1/2}, \tag{15.48}$$

was used to convert Fig. 15.17 to Fig. 15.18. The α-particle stopping power in Fig. 15.17 has been divided by the square of the α-particle charge number $z^2 = 4$. All three curves of $(1/z^2)S/\rho$ vs β are described by very similar functions for $\beta > 0.04$, though the electron and α-particle curves are about 10 % below the proton curve.[4] At low speeds the scaled α-particle curve falls significantly below the proton curve. The reason, the formation of an electron cloud on the α particle, is discussed below.

It is not difficult to understand the basic shape of the stopping power curve. Most of the energy loss is from the projectile to the electrons of the target atom. Since the electrons are bound to the target nucleus, the speed with which the projectile passes the target is important. Imagine pushing slowly on a swing with a force that gradually increases and then decreases. The net force on the swing is the vector sum of the external force exerted $\mathbf{F}_{\text{ext}}$, the vertical pull of gravity, and the tension in the ropes and equals the swing's mass times acceleration. For small horizontal displacements x from equilibrium, the vector sum of the weight and the tension in the string is horizontal and nearly proportional to x. It points toward the equilibrium position, and for small displacements is approximately a linear restoring force. If the proportionality constant is k, $ma = F_{\text{ext}} - kx$. This is the equation of motion for an undamped harmonic oscillator (Chap. 10 and Appendix F). If the force builds up slowly, there is a very small acceleration, and the swing angle changes so that $F_{\text{ext}} \approx kx$. As the force decreases the swing returns to its resting position. All of the work that was done to displace the swing is now returned as work by the swing on the source of the external force. No net energy has been imparted to the swing. This is called an *adiabatic* process or approximation, a slightly different use of the term than in Chap. 3.

At the other extreme, the force could be applied for a very short time, building up to a peak and falling quickly. In this case, the swing does not have time to move and $F_{\text{ext}} = ma$. This can be integrated to give

$$\int F_{\text{ext}}\, dt = m \int a\, dt = m(v_{\text{final}} - v_{\text{initial}}). \tag{15.49}$$

The swing acquires a velocity and hence some kinetic energy. The integral of force with respect to time is called the *impulse*, and this limit is the *impulse approximation*.

[4] A value $\beta = 0.04$ corresponds to a kinetic energy of 400 eV for electrons, 800 keV for protons, and 3.2 MeV for α particles.

The two limits depend on whether the duration of the force is long or short compared to the natural period of the swing. The atomic electrons are bound, and they have a natural period that is the circumference of their orbit divided by their speed v_{electron}. The length of time that a projectile exerts a force on the electrons is roughly the diameter of the atom divided by the projectile speed. Ignoring factors of 2π, we see that the passage of the projectile will be adiabatic if

$$\frac{d_{\text{atom}}}{v_{\text{projectile}}} \gg \frac{d_{\text{atom}}}{v_{\text{electron}}}$$

or $v_{\text{projectile}} \ll v_{\text{electron}}$. The impulse approximation will be valid if $v_{\text{projectile}} \gg v_{\text{electron}}$.

This is sufficient to explain the shape of the stopping-power curves in Fig. 15.18. When the projectile has very low energy it moves past the atom so slowly that the electrons have time to rearrange themselves[5] and then return to their original state as the projectile leaves, restoring to the projectile the energy that they received while rearranging. As the projectile speed increases, the process is no longer adiabatic, first for the more slowly moving outer electrons and then for more and more of the inner atomic electrons as the speed increases. At the other extreme, when the projectile speed becomes high enough, we can think of the process in terms of the impulse approximation. The faster the projectile moves by, the shorter the time the force is applied and the smaller the energy transfer. The energy transfer is most effective, and the peak of the stopping power occurs, when the speed of the projectile is about equal to the speed of the atomic electrons in the target.

The cross section $d\sigma/dW$ in Eqs. 15.44–15.46 is the sum of cross sections for three possible processes. We have already described the stopping power due to interactions of the projectile with the target electrons, S_e. There is another contribution to the stopping power from interactions of the projectile with the target nucleus, S_n. It is also possible for the energy loss to involve the radiation of a photon, so we also have radiative stopping power, S_r. Because these are independent processes, the total stopping power and the cross section are each the sum of three terms:

$$\begin{aligned} \frac{S}{\rho} &= \frac{S_e}{\rho} + \frac{S_n}{\rho} + \frac{S_r}{\rho}, \\ \frac{d\sigma}{dW} &= \left(\frac{d\sigma}{dW}\right)_e + \left(\frac{d\sigma}{dW}\right)_n + \left(\frac{d\sigma}{dW}\right)_r. \end{aligned} \tag{15.50}$$

To compare these processes, we need to consider the maximum energy that can be transferred, as well as the relative

[5] Classically, if the electrons go around the nucleus many times while the projectile moves by, the shape of their orbits can change in response to the projectile. Quantum mechanically, the shape of the wave function can change, but the quantum numbers do not change.

Table 15.3 Maximum energy transfer and relative importance of nuclear and radiative interactions for various projectiles and targets

Projectile	Target	Nuclear W_{max} (eV)	Electron W_{max} (eV)	S_n/S	S_r/S
Electron, 100 keV	Hydrogen	240	50,000		0.01 %
	Carbon	20	50,000		0.09 %
	Lead	1	50,000		2.2 %
Electron, 1 MeV	Hydrogen	4300	500,000		0.13%
	Carbon	360	500,000		0.65 %
	Lead	20	500,000		11.5 %
Proton, 10 keV	Hydrogen	5000	20	1.7 %	
	Carbon	2800	20	1.6 %	
	Lead	200	20	1.5 %	
Proton, 100 keV	Hydrogen	50,000	220	0.17 %	
	Carbon	28,400	220	0.15 %	
	Lead	1900	220	0.24 %	
Proton, 1 MeV	Hydrogen	500,000	2200	0.11 %	
	Carbon	280,000	2200	0.07 %	
	Lead	19,000	2200	0.09 %	
α particle, 10 keV	Hydrogen	6400	5	27 %	
	Carbon	7500	5	12 %	
	Lead	700	5	10 %	
α particle, 100 keV	Hydrogen	64,000	50	1.6 %	
	Carbon	75,000	50	1.1%	
	Lead	7400	50	1.8 %	
α particle, 1 MeV	Hydrogen	640,000	500	0.13 %	
	Carbon	750,000	500	0.12 %	
	Lead	74,000	500	0.20 %	

probability of each process. The maximum possible energy transfer W_{max} can be calculated using conservation of energy and momentum. For a collision of a projectile of mass M_1 and kinetic energy T with a target particle of mass M_2 which is initially at rest, a nonrelativistic calculation gives

$$W_{max} = \frac{4TM_1M_2}{(M_1 + M_2)^2}. \tag{15.51}$$

The analogous relativistic equation (needed, for example, when the projectile is an electron) is

$$W_{max} = \frac{2(2 + T/M_1c^2)TM_1M_2}{M_1^2 + 2(1 + T/M_1c^2)M_1M_2 + M_2^2}. \tag{15.52}$$

The values of W_{max} for representative projectiles and targets are shown in Table 15.3, along with the percentage of the stopping power due to nuclear collisions. For electrons, the table also shows the percentage of the stopping power due to radiative transitions. The percentages are calculated from ICRU Report 49 (1993). Electrons can scatter from nuclei, but the amount of recoil energy transferred to the nucleus is very small. Although electrons undergo a great deal of nuclear scattering, which results in a tortuous path through material, they lose very little energy in a nuclear scattering. The heavier projectiles can lose relatively more energy in each nuclear collision than in each electron collision. For a given kind of projectile, nuclear stopping is more important at lower energies, because less energy can be transferred to an electron. The heavier the projectile for a given energy, the more important the nuclear term becomes, for the same reason.

The collision of electrons with electrons is a special case. Equation 15.51 or 15.52 gives $W_{max} = T$. Consider the collision of two billiard balls of the same mass. If the projectile misses the target, it continues straight ahead with its original energy and $W = 0$. If it hits the target head on, it comes to rest and the target travels in the same direction with the same energy that the projectile had—a situation indistinguishable from the complete miss. It is customary (but arbitrary) in the case of identical particles to say that the particle with higher energy is the projectile, so $W_{max} = T/2$. This adjustment has been made in Table 15.3 for electrons on electrons and protons on protons.

Radiation is only important for electrons and occurs in a certain fraction of the elastic electron scatterings from the target nucleus. Nuclear scattering gives the electron a fairly large acceleration. Classically, an accelerated charged particle radiates electromagnetic waves. This process is called *bremsstrahlung*—braking or deceleration radiation. The energy radiated is proportional to the square of the acceleration, so bremsstrahlung is only important for light projectiles. There is also a contribution from electron–electron or positron–electron scattering. The electron–electron contribution vanishes at low energies, although the positron–electron

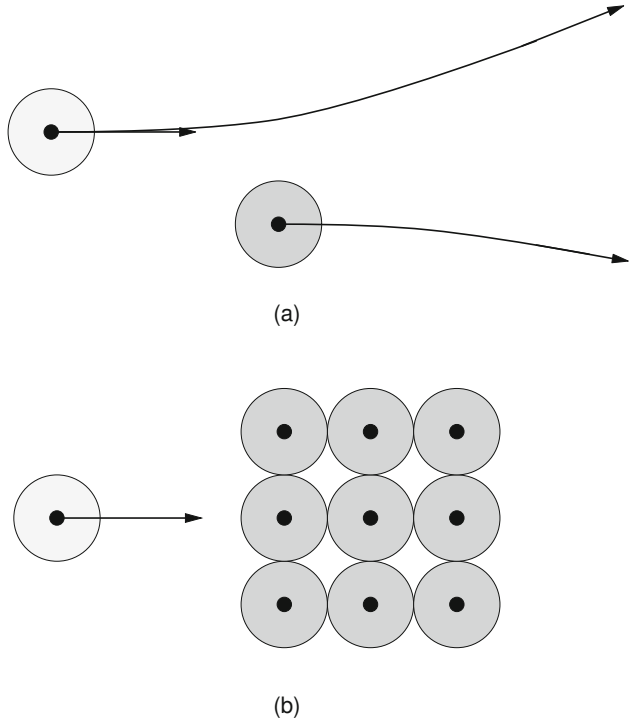


Fig. 15.19 A projectile, which may or may not carry an electron cloud, moves past a target atom. **a** In a gas the projectile interacts with one atom at a time. **b** In a liquid or a solid, neighboring atoms may influence the interaction

bremsstrahlung does not.[6] We will see in Chap. 16 that bremsstrahlung is an important component of the x-ray spectrum produced when a beam of electrons strikes a target. Even so, the fraction of the electron energy that is converted to radiation is small.

An atom has a radius of a few times 10^{-10} m. The nucleus of the atom is much smaller, about 10^{-15} m, and contains most of the atom's mass. The atom's size is determined by the electron cloud around the nucleus. Figure 15.19a shows a projectile entering at the left and traveling to the right through a gas. It interacts with one target atom at a time. The solid black dots represent the nuclei of the projectile and the target atom. The shaded circles represent the electron clouds. The projectile may or may not have an electron cloud, which is shown with lighter shading. Figure 15.19b shows the interaction with a solid or liquid in which the target atoms are tightly packed, and it may not be accurate to say that the projectile interacts with only one atom at a time.

[6] This difference can be understood classically. In the first approximation, the radiation by a charge is proportional to the product of the charge times its acceleration, qa. For two interacting electrons, $a_1 = -a_2$, $q_1 = q_2$, and the sum of these two terms vanishes. For an electron and a positron $a_1 = -a_2$, $q_1 = -q_2$, and the two terms add.

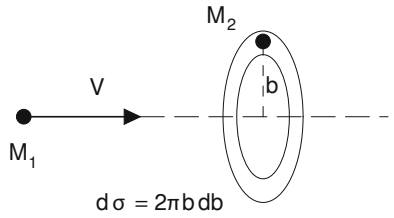


Fig. 15.20 The impact parameter is the perpendicular distance from the target particle to a line extended from the projectile in the direction of its velocity before the interaction

Classically, the motion of a charged projectile past a charged target depends on the charges and masses of the particles, the initial velocity or kinetic energy of the projectile, and the *impact parameter* b, which is the perpendicular distance from a line through the initial velocity of the projectile to the target, as shown in Fig. 15.20. The classical cross section for having an impact parameter between b and $b + db$ is the area of the ring, $2\pi b\, db$. If we could relate b to the energy loss W, we would have the cross section $d\sigma/dW$ of Eq. 15.46.

The energy-loss process is quite complicated, and the cross section cannot be calculated exactly. A great deal of experimental and theoretical work on stopping powers has been done, extending from 1899 to the present time. The history is nicely reviewed by Ziegler et al. (1985). Much of the recent work on stopping powers has been motivated by the use of ion implantation to make semiconductors, the analysis of materials using ion beams, and medical applications. Currently stopping powers of low-energy heavy ions can be calculated with an accuracy of better than 10 %. For high-speed light ions the accuracy is better than 2 %.

15.11.1 Interaction with Target Electrons

We first consider the interaction of the projectile with a target electron, which leads to the *electronic stopping power*, S_e. Many authors call it the *collision stopping power*, S_{col}. There can be interactions in which a single electron is ejected from a target atom or interactions with the electron cloud as a whole (a *plasmon* excitation). The stopping power at higher energies, where it is nearly proportional to β^{-2}, has been modeled by Bohr, by Bethe, and by Bloch (see the review by Ahlen 1980). The Bethe–Bloch model is also valid for relativistic energies. A nonrelativistic model for high energies was developed by J. Lindhard and his colleagues (see references in Ziegler et al. 1985). It allows more accurate calculations of which electrons in the target receive energy from the projectile.

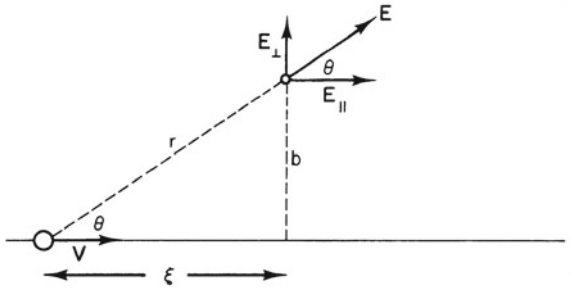


Fig. 15.21 A heavy particle of charge ze, mass M, and velocity $\mathbf{V}$ moves past a stationary electron

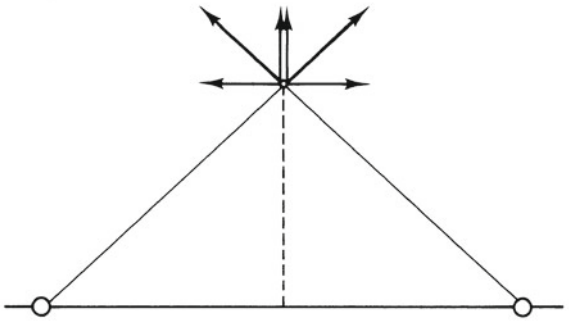


Fig. 15.22 Why the parallel component of $\mathbf{p}$ is zero. For every point where the projectile gives a particular $\mathbf{E}_\parallel$, there is a symmetric point where $\mathbf{E}_\parallel$ is equal but opposite. The components $\mathbf{E}_\perp$ are in the same direction in both places, so the perpendicular component of $\mathbf{p}$ does not vanish

We can gain considerable insight into the high-energy loss process by making a classical calculation of the cross section for transferring energy to an electron using the impulse approximation. This is a simplification of the Bethe–Bloch model. In our model, a heavy projectile passes by a free electron that is at rest. Momentum is transferred from the projectile to the electron. Because of its large mass, the projectile's velocity does not change appreciably, but the lighter electron acquires an appreciable velocity and kinetic energy. If the momentum transferred to the electron is $\mathbf{p}$, its kinetic energy is $p^2/2m_e$. That kinetic energy must have been lost by the projectile.

Figure 15.21 shows a particle of mass M, charge ze, and velocity $V = \beta c$ moving past a stationary electron. The impact parameter b is the perpendicular distance from the electron to the path of the projectile. The distance from the projectile to the electron is r, and the distance along the path to the point of closest approach is ξ. The momentum transferred to the electron is $\mathbf{p} = \int \mathbf{F}dt = -e\int \mathbf{E}dt$. By symmetry, there is no component of $\mathbf{p}$ parallel to the path of the projectile. The reason is shown in Fig. 15.22. For each location of the projectile that gives a parallel component of $\mathbf{F}$ in one direction, there is a position an equal distance on the other side of the point of closest approach that gives a component of $\mathbf{F}$ with the same magnitude but in the opposite direction. The perpendicular component of $\mathbf{F}$ is the same for both locations, so there is a net perpendicular component of momentum transfer. The magnitude of the perpendicular component of $\mathbf{E}$ is

$$E_\perp = E\sin\theta = \frac{ze\sin\theta}{4\pi\epsilon_0 r^2} = \frac{zeb}{4\pi\epsilon_0 r^3} = \frac{ze}{4\pi\epsilon_0}\frac{b}{\left(\xi^2+b^2\right)^{3/2}}.$$

The perpendicular impulse is

$$\int F_\perp dt = -e\int E_\perp (dt/d\xi)d\xi.$$

If the fraction of energy lost by the projectile is small, then $dt/d\xi = 1/\beta c$ does not change during the collision. The magnitude of the impulse is therefore

$$\begin{aligned} p &= -\frac{e}{V}\int E_\perp d\xi = -\frac{ze^2 b}{4\pi\epsilon_0\beta c}\int_{-\infty}^{\infty}\frac{d\xi}{(\xi^2+b^2)^{3/2}} \\ &= -\frac{ze^2 b}{4\pi\epsilon_0\beta c}\lim_{x\to\infty}\left[\frac{\xi}{b^2(\xi^2+b^2)^{1/2}}\right]_{-x}^{x} \\ &= -\frac{2ze^2}{4\pi\epsilon_0\beta c b}. \end{aligned}$$

The smaller the impact parameter, the greater the momentum transfer to the electron. The kinetic energy acquired by the electron is

$$W = \frac{p^2}{2m_e} = \frac{2z^2e^4}{(4\pi\epsilon_0)^2 m_e c^2\beta^2 b^2}.$$

The factor $e^4/(4\pi\epsilon_0)^2 m_e c^2$ depends only on the charge and mass of the electron. It can be written as $r_e^2 m_e c^2$, where r_e is the classical radius of the electron (Eq. 15.17). The factor has the numerical value

$$r_e^2 m_e c^2 = 6.50\times10^{-43}\,\mathrm{J\,m^2} = 4.06\times10^{24}\,\mathrm{eV\,m^2}.$$

Using this notation the energy transfer per target electron is

$$W = \frac{2z^2 r_e^2 m_e c^2}{\beta^2 b^2}. \tag{15.53}$$

Note that W does not depend on the mass of the heavy projectile, but only on its speed. As the speed becomes less, the energy transfer becomes greater, because the projectile takes longer to move past the electron and the force is exerted for a longer time (as long as the time is still short enough so that the impulse approximation remains valid).

If the electrons are uniformly distributed, the cross section for each electron is $d\sigma = (d\sigma/dW)dW = 2\pi b\,db$. This can be written, with the help of Eq. 15.53, in terms of W:

$$\frac{d\sigma}{dW}dW = \frac{4\pi z^2 r_e^2 m_e c^2}{2\beta^2}\frac{dW}{W^2}. \tag{15.54}$$

This expression diverges as W approaches zero, corresponding to very large impact parameters. However, the assumption that the target electrons are free fails in this limit, so that there is some effective lower limit $W_{\min}$. Also, the greater the impact parameter, the longer the electron will experience the force exerted by the projectile (though it will be weaker). If the time is too long, the electron can move in response to the force and not absorb as much energy; the impulse approximation is no longer valid. We have already seen that there is a maximum energy transfer $W_{\max}$. Multiplying the cross section by W, integrating from $W_{\min}$ to $W_{\max}$, and noting that there are Z electrons per target atom, we obtain

$$\frac{S_e}{\rho} = \frac{4\pi N_A r_e^2 m_e c^2}{\beta^2}\frac{Z}{A}z^2 \ln\left(\frac{W_{\max}}{W_{\min}}\right). \tag{15.55}$$

The factor $4\pi N_A r_e^2 m_e c^2$ has the value 30.707 eV m^2 mol^{-1} = 0.307 07 MeV cm^2 mol^{-1}.

A quantum-mechanical calculation gives a result of essentially the same form as Eq. 15.55. The logarithmic term includes both ionization and plasmon excitation[7] and is called the *stopping number per atomic electron* $L(\beta, z, Z)$:

$$\frac{S_e}{\rho} = \frac{4\pi r_e^2 m_e c^2}{\beta^2} N_A \frac{Z}{A} z^2 L(\beta, z, Z). \tag{15.56}$$

For heavy charged particles L has the form

$$L(\beta, z, Z) = L_0 + zL_1 + z^2 L_2,$$
$$L_0 = \ln\left(\frac{\beta^2}{1-\beta^2}\right) + \ln\left(\frac{2m_e c^2}{I(Z)}\right) - \beta^2 - \frac{C}{Z} - \frac{\delta}{2}. \tag{15.57}$$

Equation 15.56 with $L = L_0$ is often called the *Bethe–Bloch formula*. The second term in L_0 depends on $I(Z)$, the ionization potential of the atoms in the absorber, averaged over all the electrons in the atom. Values of $I(Z)$ have been calculated theoretically and also derived from measurements of the stopping power. They range from 14.8 eV for hydrogen to 884 eV for uranium. The value 14.8 eV is greater than the ground-state energy of hydrogen, 13.6 eV, because the ejected electron has some average kinetic energy. Published values of I can vary considerably, depending on whether the other correction terms are present. For example, values of I in the literature for hydrogen range from 11 to 20 eV. Discussions of values for I and the various terms in L can be found in ICRU Report 49 (1993), in Ahlen (1980), and in Attix (1986). The term $\delta/2$ corrects for the *density effect*. The calculation above assumed that the electron experienced the full electric field of the projectile. However, other electrons in the absorber move slightly, polarizing the absorber and reducing the field. This effect becomes important at high energies as the electric field is distorted by relativistic effects. It also depends on the density of the absorber. A small density effect persists in conductors even at low energies; however, it is usually incorporated into the value of $I(Z)$. For the projectile energies we are considering, the density effect is most important for electrons.

An alternative nonrelativistic treatment by Lindhard and colleagues allows the use of accurate atomic electron density distributions and also considers the effect of electrons in neighboring atoms.[8] In the Lindhard model the stopping power is

$$\frac{S_e}{\rho} = \frac{N_A}{A}\int z^2 I(V, \rho_e)\rho_e 4\pi r^2 dr, \tag{15.58}$$

where z is the projectile charge, I is the *stopping interaction strength* in J m^2 (more often in eV pm^2),[9] ρ_e is the electron density in the atom (in units of the electron charge), and $4\pi r^2 dr$ is the volume element. Integration of ρ_e over all volume gives Z, the atomic number of the target. Comparison of Eqs. 15.58 and 15.46 shows that the integral in Eq. 15.58 is the stopping cross section per target atom.

Figure 15.23 shows how the Lindhard model explains why the stopping power falls below the $1/\beta^2$ curve at lower projectile velocities. Each panel shows the electron density in copper, $4\pi r^2\rho_e$, and the interaction strength I. Their product, the solid line, is the integrand in Eq. 15.58. The integral is taken from 0 to 0.14 nm [1.4 Å (angstrom)]. The K, L, and M shells of copper can be seen in the electron density curve. Figure 15.23a is for a 10-MeV proton or some other heavy ion with the same speed. The projectile is moving fast enough so that all electrons except those in the K shell interact with it. Contrast this with Fig. 15.23b, which is for a 100-keV proton. The projectile speed is much less, and the interaction is almost exclusively with the outer electrons.[10]

[7] A plasmon excitation is due to the interaction of the projectile with the entire electron cloud of the atom.

[8] The electron density functions are calculated using quantum mechanics. The problem is to find the electron distribution by solving Schrödinger's equation with the potential distribution due to the nucleus and the potential due to the electron charge distribution for which one is solving. This self-consistent computation is called the *Hartree–Fock approximation*.

[9] I is not the same as the average ionization energy of Eq. 15.57.

[10] The solid line representing the integrand does not fall to zero at 0.12 nm = 1.2 Å because of the effect of electrons from neighboring atoms. In a solid there are no regions where the electron density is zero.

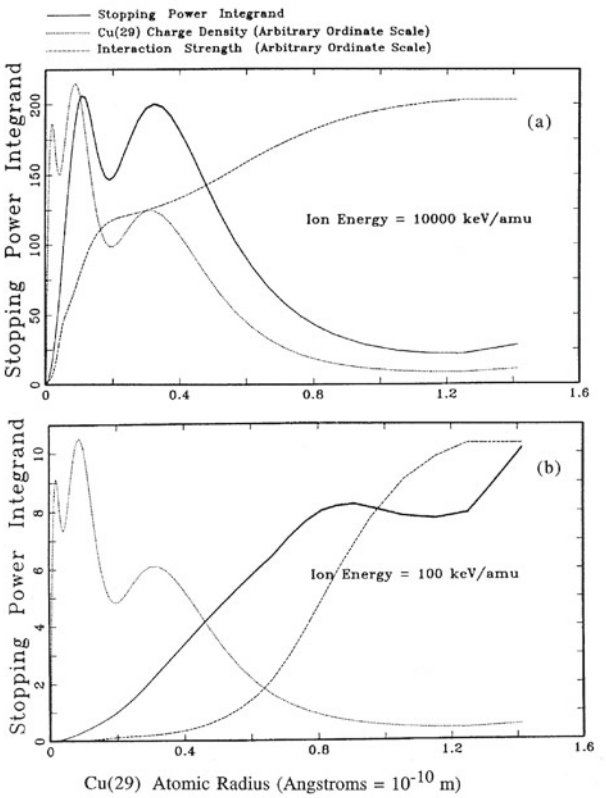


Fig. 15.23 Calculation of the stopping power at low energies involves integrating the product of the electron charge distribution in the target atom and the interaction strength function, which depends on the projectile speed. The *dotted line* shows the electron charge density for copper. The *solid line* shows the integrand. **a** For 10-MeV protons, all electrons but those in the K shell contribute. **b** For 100-keV protons the interaction function has changed, and only the outermost electrons contribute. Note the much different ordinate scales in **a** and **b**. (Provided by J. F. Ziegler)

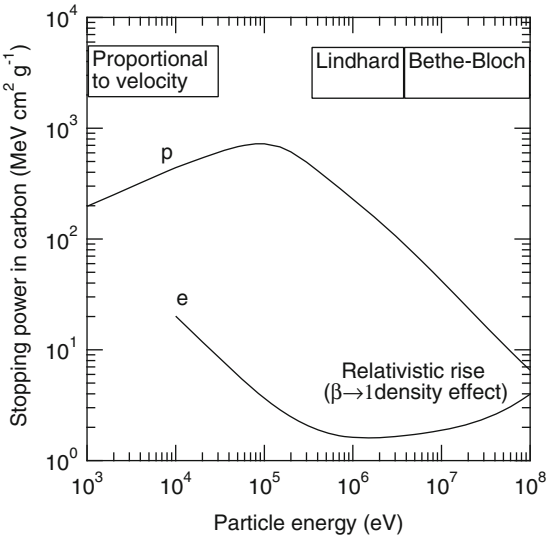


Fig. 15.24 Proton and electron stopping power vs energy in carbon, showing the regions in which various models are valid

Both the Bethe–Bloch and Lindhard models fail at low energy, because the electrons are not free and many of the interactions are adiabatic. Some models reviewed by Ziegler et al. (1985) predict a stopping power proportional to projectile velocity. This has been found to be true in general, though not for all elements. The experiments are quite difficult because of the thinness of the targets, contamination, etc. Figure 15.24 shows the regions where the various models apply for protons. For electrons, relativistic effects are important above about 500 keV. The rise in stopping power at high energies is due to the density effect (polarization of the electrons).

Another important effect at low energies is that the slowly moving ion can capture electrons, decreasing the value of z^2. Ziegler et al. (1985) discuss the scaling of data for different projectiles and the appropriate effective charge values. The average projectile charge follows a universal curve when plotted as a function of the appropriate combination of the speeds of the projectile and target electrons. They, and the ICRU Report 49 (1993), assume that for protons the effective charge is always unity. The theoretical justification is that the radius of the electron orbit in hydrogen is larger than the interatomic separation in solids.

15.11.2 Scattering from the Nucleus

The projectile can also scatter from the target atom as a whole. The recoil kinetic energy of the atom is lost by the projectile. Since the nucleus contains most of the mass, the kinematics are those of the bare projectile and the target nucleus, and this process is called *nuclear scattering*, with stopping power S_n. (Sometimes it is called *elastic scattering*, with a subscript that can cause it to be confused with electron interactions.)

Just as with Compton scattering, knowing the angle through which the projectile is scattered defines the amount of energy transferred to the target. The angle depends on the impact parameter. The problem can be solved for a given impact parameter if the force between the projectile and target is a function only of their separation and one knows the potential energy of their separation. The details are found in Ziegler et al. (1985). We will simply comment on the contributions to the potential energy. They are

1. The Coulomb force between the projectile and the target nucleus.
2. The Coulomb force between the projectile and the electron cloud of the target atom.
3. The Coulomb attraction between the target nucleus and any electrons surrounding the projectile.

4. The Coulomb repulsion between the electron clouds of the target and the projectile.
5. A term due to the Pauli exclusion principle if the projectile is an ion with an electron cloud. To see how it arises, suppose that both the projectile and target have both of their possible K-shell electrons. If the nuclei get close enough, they effectively form a single nucleus that cannot have four K-shell electrons. Therefore two of the electrons have to move to unfilled shells. This requires energy that comes from the kinetic energy of the projectile. This is called *Pauli promotion*. Even though the electrons have time to return to their original orbits for a slow projectile, the effect changes the overall potential and hence the projectile orbit and the probability of a particular energy transfer.
6. An *exchange term* that also arises from the Pauli principle, related to whether the spins of the projectile and target electrons are parallel or antiparallel.

Because nuclear scattering is relatively unimportant for the charged particles we are considering and because it does not lead to ionization, we will not describe any details of the calculations.

15.11.3 Stopping of Electrons

Equations similar to Eq. 15.56 are obtained for electrons and positrons. Recall that energy loss in nuclear scattering is negligible for positrons and electrons because they are so light, and that bremsstrahlung transfers some of the electron kinetic energy to radiation. Electrons and positrons are assumed to collect no screening charge. Even at low energies, the electron velocities are high enough so that the Bethe–Bloch model is used. The collision stopping power for electrons is[11]

$$\frac{S_e}{\rho} = 4\pi N_A r_e^2 m_e c^2 \frac{1}{\beta^2}\frac{Z}{A} L_{\pm}. \tag{15.59}$$

The subscript $\pm$ indicates that stopping number per electron is slightly different for electrons and for positrons. The exact forms can be found in Attix (1986) or in ICRU Report 37 (1984). In both cases L depends on $I(Z)$ and the density effect. An accurate calculation of the shell correction for electrons has not been made; therefore ICRU Report 37 omits the shell correction from the tables for electrons and positrons. This omission makes the use of Eq. 15.58 less accurate for electrons below 10 keV. The best values of S_e/ρ for electrons and positrons are obtained from theoretical calculations using Eq. 15.59 and values of $I(Z)$ determined from proton data.

[11] The literature often replaces the 4π by 2π for electrons and makes L twice as large.

15.11.4 Compounds

In dealing with compounds, it is frequently assumed that each atom in the target interacts independently with the projectile, as we assumed for photons. The stopping power per molecule is then equal to the sum of the stopping powers for each atom in the molecule. This leads to a formula analogous to Eq. 15.31, known as the *Bragg rule*:

$$\frac{S}{\rho} = \sum_i w_i \left(\frac{S}{\rho}\right)_i. \tag{15.60}$$

This equation applies to the collision, radiative, nuclear, and total stopping powers. The approximation is quite inaccurate near the peak of the stopping power curve, where the errors can be greater than a factor of 2. This is not surprising, given the behavior of the scattering function I in Fig. 15.23b. Most of the energy loss is to outer electrons—the conduction electrons if the substance is a metal.

In a semiconductor there are gaps in the energy levels, and this precludes the low-energy transfers. As a result, the stopping power is lower in semiconductors. In crystals, channeling can occur: the stopping power depends on the orientation of the trajectory with the crystal symmetry axis.

Carbon poses a particular problem. It is an important element in the body, and it has chemical bonds that range from metallic to insulating in nature. Various investigators have shown variations in stopping power of 30 % for ions in pure carbon, depending on how it was fabricated. Graphite can be made with different electrical conductivities, and there are associated differences in stopping power. Ziegler and Manoyan (1988) have applied charge-scaling techniques to several organic carbon compounds by considering separately the stopping due to closed atomic shells (cores) and the remaining bonds between different pairs of atoms.

ICRU Reports 37 (1984) and 49 (1993) handle departures from the Bragg rule in the first approximation by using different values of I for electrons in compounds. The density effect is important for electrons and also does not follow the Bragg rule.

Stopping-power values are found in ICRU Report 37 for positrons and electrons. ICRU Report 49 has stopping powers for protons and α particles. These data are also found on the web: http://www.nist.gov/pml/data/star/index.cfm. A computer program for protons and ions, SRIM (Stopping and Range of Ions in Matter) is described by Ziegler et al. (1985) and is available at www.srim.org. It is updated every few years.

15.12 Linear Energy Transfer and Restricted Collision Stopping Power

In modeling the effect of ionizing radiation on targets, whether they be radiation detectors, photographic emulsions, cells, or parts of cells, one often wants to know how much of a charged particle's energy is absorbed "locally," that is, within some region around a particle's trajectory. An accurate calculation is difficult, since some of the electrons produced may leave the region of interest. Also, the energy absorbed in some region of interest around a particle track comes both from energy lost by the particle while traversing that track segment and also from photons and charged particles produced elsewhere by the projectile. (This is discussed in detail in ICRU Report 16 1970.)

An approximation to the desired quantity is the *linear energy transfer* (LET) or the *restricted linear collision stopping power* L_Δ. It is defined as the ratio dT/dx, where dx is the distance traveled by the particle and dT is the mean energy loss to electrons that results in energy transfers less than some specified Δ. This use of the symbol L should not be confused with the stopping number of Eqs. 15.56–15.59. The quantity L_Δ can be calculated by replacing $W_{\max}$ by Δ in the expression for the stopping power. The value of Δ is usually specified in electron volts.

The electron stopping power S_e is numerically equal to L_∞. However, S_e is defined in terms of the energy *lost by the particle*, while L_∞ is defined in terms of energy *imparted to the medium*.

Note that although the quantity actually of interest may be the energy imparted within some region around the trajectory, this definition is based on energy transfers less than Δ. A quantity based on the region of interest would be easier to measure; L_Δ is easier to calculate.

ICRU Report 37 calculates L_Δ for positrons and electrons for values of Δ down to 1 keV. The report points out that such calculations are inaccurate for smaller values of Δ, even in light elements. ICRU Report 16 provides values of L_Δ for protons and heavy ions.

15.13 Range, Straggling, and Radiation Yield

We can see in Fig. 15.27 that the α particles, entering from the bottom with the same energy, all travel about the same distance before coming to rest. This distance is called the *range* of the α particles. It will be defined more precisely below.

We can estimate the range in the following way. The stopping power represents an average energy loss per unit path length. The actual energy loss fluctuates about the mean values given by the stopping power. If these fluctuations are neglected and the projectiles are assumed to lose energy continuously along their tracks at a rate equal to the stopping power, then one is making the *continuous-slowing-down approximation* (CSDA). In this approximation one can calculate the range, the distance a particle with initial energy T_0 travels before coming to rest or reaching some final kinetic energy T_f. A factor ρ is introduced to express the range in mass per unit area:

$$R_{\mathrm{CSDA}}(T_0, T_f) = \rho \int dx = \rho \int_{T_f}^{T_0} \frac{dT}{S_e + S_n + S_r}. \tag{15.61}$$

ICRU Report 37 (1984) discusses the problem of carrying the integration to $T_f = 0$.

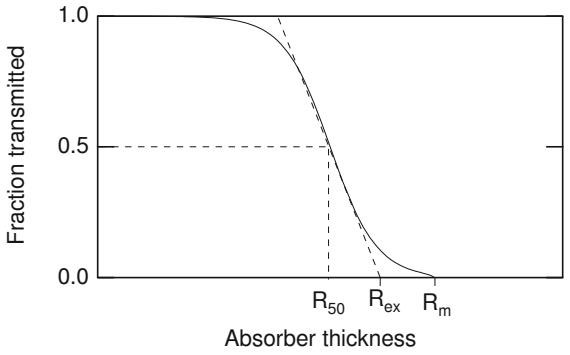


Fig. 15.25 Plot of the number of particles passing through an absorber vs its thickness to show the definition of various ranges. R_{50} is the median range, R_{ex} is the extrapolated range, and R_m is the maximum range

The CSDA range is not directly measurable. Measurements of the fraction $F(R)$ of monoenergetic particles in a beam that passes through an absorber of thickness R gives a curve like that of Fig. 15.25. Various ranges can be defined using this curve. The most easily measured is the median range R_{50}, corresponding to an absorber thickness that transmits 50 % of the incident particles. The extrapolated range R_{ex} is obtained by extrapolating the linear portion of the curve to the abscissa. The maximum range R_m is the thickness that just stops all of the particles; it is, of course, very difficult to measure. If $F(R)$ is known accurately one can define a *mean range* $\overline{R} = \int R(-dF/dR)dR / \int (-dF/dR)dR$. If the shape of the transmission curve is perfectly symmetrical about the mean, then R_{50} is equal to $\overline{R}$, even though they are conceptually quite different. For heavy projectiles $\overline{R}$ (usually approximated by R_{50}) provides the best estimate of R_{CSDA}.

The fluctuations in the range are called *straggling*. The straggling distribution has also been calculated. The track of a heavy projectile such as an α particle is fairly straight, because the various scattering interactions result only in small

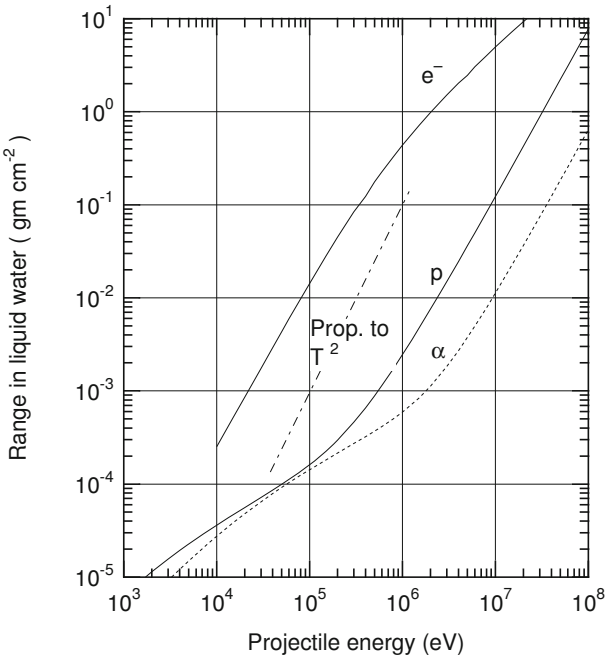


Fig. 15.26 Range of electrons, protons, and α particles in liquid water. Data are from ICRU Reports 37 (1984) and 49 (1993). Note that for water the range in gm cm^{-2} is the same as the range in cm



Fig. 15.27 Tracks of 22-MeV α particles in photographic emulsion. The α particles enter at the *bottom of the page* and come to rest near the top. The *small square fiducial marks* at the bottom are 10 μm apart. The features of the tracks are discussed in the text. (From Powell et al. 1959). Reproduced by permission of Prof. D. H. Perkins

angular deviations. The straggling results primarily from the fact that Sdx represents only an average energy loss in path length dx. The fluctuations can be integrated to give the spread in range; see Ahlen (1980) or ICRU Report 37 (1984) or ICRU Report 49 (1993) or the computer program SRIM (Ziegler et al. 1985).

Electrons and positrons are so light that they undergo large-angle scattering (occasionally from an electron, more often from an atomic nucleus). The resulting electron trajectories are quite tortuous, as can be seen in Figs. 15.28 and 15.29. The median or mean range for an electron is considerably less than R_{CSDA}. For electrons and positrons the extrapolated range R_{ex} corresponds most closely to R_{CSDA}, at least in materials with atomic number up to silver (Tung et al. 1979). Figure 15.26 shows ranges in water. At medium energies the dependence on energy is approximately T^2.

Tables of ranges are found in the references cited above or at the NIST web site http://www.nist.gov/pml/data/star/index.cfm.

The *radiation yield*, Y, is the fraction of the initial particle (usually electron) kinetic energy T_0 that is converted to bremsstrahlung photons as the electron comes to rest in the medium in question. The yield is calculated using the continuous-slowing-down approximation as (neglecting S_n)

$$Y(T_0) = \frac{1}{T_0}\int_0^{T_0} \frac{S_r(T)dT}{S_e(T) + S_r(T)}. \qquad (15.62)$$

15.14 Track Structure

We can gain insight into the interaction processes by examining tracks in photographic emulsions or in cloud chambers. Figures 15.27 and 15.28 are taken from a classic atlas of tracks in nuclear emulsions (Powell et al. 1959). They show the difference between the interaction of heavy and light particles in matter. Figure 15.27 shows the tracks of four cosmic-ray α particles, each of which entered the bottom of the figure and stopped near the top. The fiduciary marks along the bottom are 10 μm apart. Each track is about 195 μm long, corresponding to an initial α-particle energy of about 22 MeV. The emulsion has a density of 3.6×10^3 kg m^{-3}. Each black dot is a sensitive silver halide grain about 0.6 μm in diameter. At the beginning of the track, S is about 70 keV μm^{-1} or 42 keV per grain; 10 μm from the end of the track it is 200 keV μm^{-1} or 120 keV per grain. The energy that must be deposited in a grain to render it developable is about 2.8 keV. The amount of energy

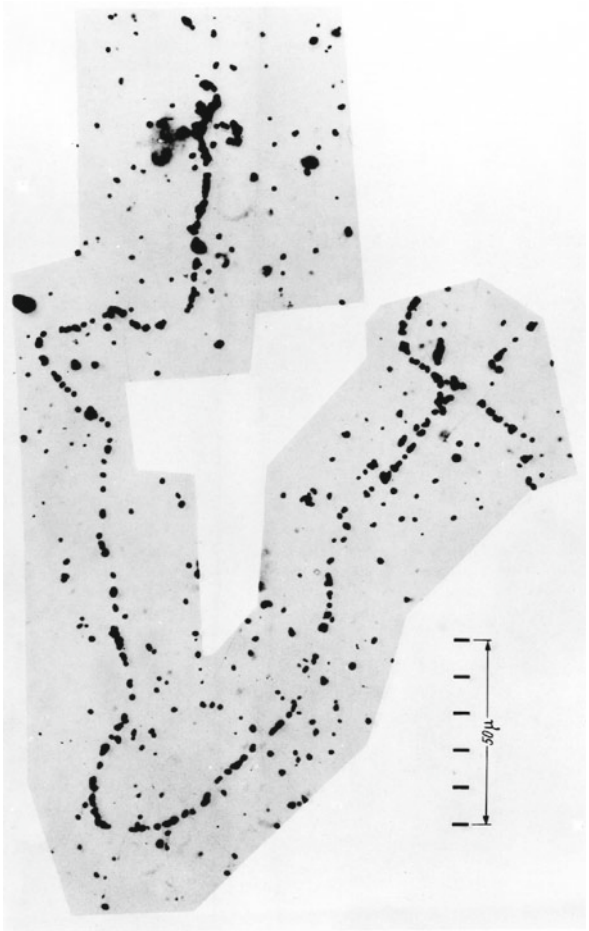


Fig. 15.28 Tracks of electrons in emulsion. An electron–positron pair was produced in the lower left corner. Each particle has an energy of about 250 keV. The details are discussed in the text. (From Powell et al. 1959. Reproduced by permission of Professor D. H. Perkins)

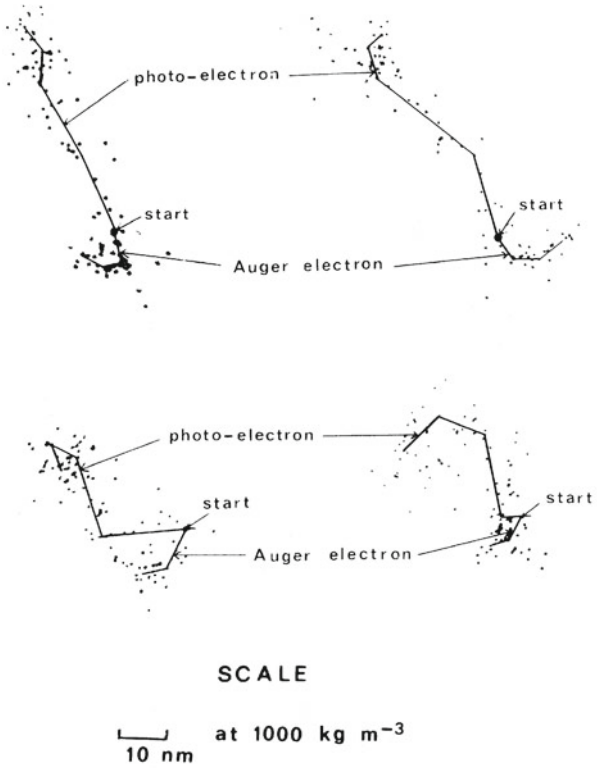


Fig. 15.29 Tracks of ≈1 keV electrons in a cloud chamber. An equivalent scale in water or tissue has been added. Photoelectrons and Auger electrons can be seen. The lines were drawn to guide the eye. (From Budd and Marshall 1983, pp. 19–32. Reproduced by permission of the Radiation Research Society)

deposited in each grain is so much larger than this that the track density is uniform. Small bumps of 1–4 grains can be seen occasionally along each track. Some of these are due to δ *rays*: electrons that have received enough energy to travel a few micrometers in the emulsion. Others are artifacts due to the general background fog. Multiple small-angle scattering causes small deviations in each track, which become greater as the α particle slows down.

In Figure 15.28 an electron–positron pair has been produced in the lower left corner of the emulsion by a 1.5-MeV photon. Each particle has a kinetic energy of about 250 keV. One immediately notices the tortuous path of both particles due to large-angle scattering. The stopping power near the beginning of the track is about 0.8 keV μm^{-1}, so that about 0.5 keV is deposited in each grain. About 30 μm from the end, the stopping power and the average amount of energy deposited in each grain are about 3 times larger. The upper track is considerably more dense near the end of its path. The failure of the other track to show this density increase could be due to annihilation of the positron in flight or to a large-angle scattering out of the emulsion.

Figure 15.29 shows the ionization produced by an electron at a much different scale. It was produced from a cloud chamber photograph of electron tracks in a low-density gas (Budd and Marshall 1983). The scale shows distances in liquid water or tissue that correspond to the same value of ρx, corrected for phase effects. Note that the scale shows 10 *nanometers*. An atomic diameter is 0.2–0.6 nm. In each case a photoelectron of energy between 950 and 1480 eV has been ejected from a gas atom in the cloud chamber. Auger electrons are also seen.

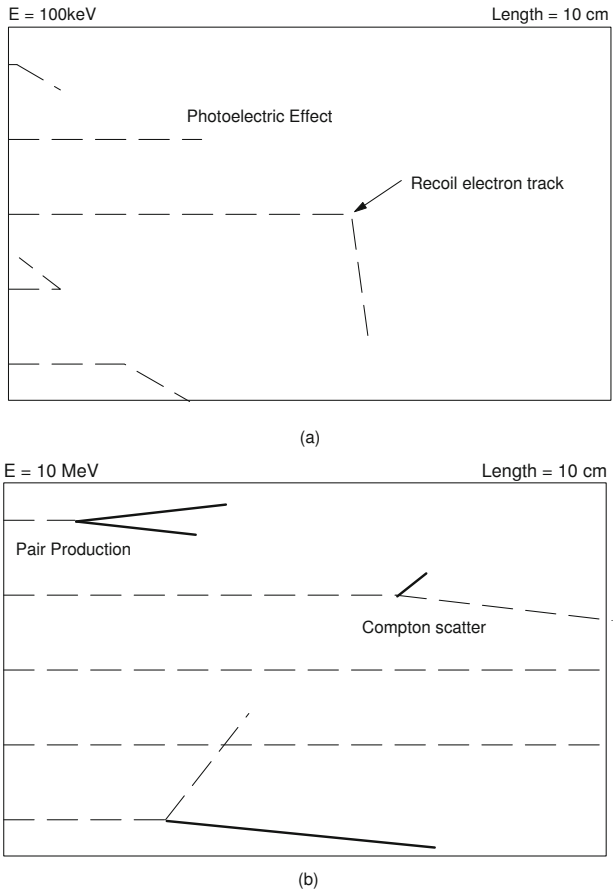


Fig. 15.30 A simulation of photons passing through a layer of water 10 cm thick. **a** The photon energy is 100 keV. One photon has a photoelectric interaction. The other four are Compton scattered. **b** The photon energy is 10 MeV. Two photons do not interact, one produces an electron–positron pair, and two Compton scatter

15.15 Energy Transferred and Energy Imparted; Kerma and Absorbed Dose

The response of a substance to radiation, whether it is the darkening of a photographic film, an electrical pulse in an ionization chamber, or the response of a tumor to radiation therapy, is due, directly or indirectly, to the ionization produced by charged particles that lose their kinetic energy in the substance through the stopping mechanisms we have just discussed. We now define some quantities that are used to describe the transfer of energy from photons to charged particles and the energy lost by charged particles due to ionization.

15.15.1 An Example

Before discussing the formal definition of these quantities, let us consider some examples of energy transfer by photons. Figure 15.30 shows some schematic interactions of photons in a sample of water 10 cm thick. They are drawn to scale.[12] In Fig. 15.30a five photons of energy 100 keV enter from the left. Photon tracks are dashed. One photon is absorbed by the photoelectric effect, and four are Compton scattered. The energy of the photoelectron and the Compton-scattered electrons is so low that the ranges are insignificant on this scale. In Fig. 15.30b the incident photons have 10 MeV energy. One has undergone pair production, two have Compton scattered, and two have passed through without interacting. The electron tracks are shown as thick solid lines. Their lengths are equal to the CSDA range of electrons or positrons of that energy. They are drawn as straight lines, even though the real tracks are tortuous.

One of the quantities of interest is the *energy transferred* to kinetic energy of charged particles in some mass of material. (We saw this briefly in the discussion surrounding Eq. 15.35.) Another is the *energy imparted* in some mass of material, which is the kinetic energy lost by charged particles as they come to rest. Figure 15.31 shows the distinction between the two quantities. It shows two photons from Fig. 15.30b: one that underwent pair production, and one that was Compton scattered. The water has been divided into ten slices, each 1 cm thick. No energy is transferred in the first slice. Energy is transferred by pair production in the second slice and by Compton scattering in the third slice. In each case the electron (or positron) produced loses kinetic energy in that slice and also in other slices. There is energy imparted in slices 2–8, even though the energy is transferred only in slices 2 and 3.

Consider now the actual numbers. In keeping with the literature,[13] we will call the energy transferred E_{tr}, even though we have been using T for kinetic energy. For pair production the energy transferred is

$$E_{\mathrm{tr}} = T_+ + T_- = h\nu_0 - 2m_e c^2$$
$$= 10 - 2 \times 0.511 = 8.978 \approx 9.0\,\mathrm{MeV}. \qquad (15.63)$$

The partition of energy between the electron and positron is stochastic. We assume for this example that about 60 % (5.4 MeV) goes to one member of the positron–electron pair and 40 % (3.6 MeV) to the other. These numbers are shown at the vertex of Fig. 15.31. From these energies the ranges can be determined. Measuring the distance from the end of the track to the boundary between each slice allows us to

[12] These examples were constructed with a pedagogical simulation program called MacDose (Hobbie 1992). The program is available at the book web site: www.oakland.edu/~roth/hobbie.htm. It runs on a Macintosh using OS-9 or earlier. There is also a 26-min video using MacDose that shows the concepts here in more detail (Hobbie 2009). It is free and available through iTunes. A more realistic but easily understood Monte Carlo simulation is described by Arqueros and Montesinos (2003).

[13] See ICRU Report 33 (1980) or Attix (1986).

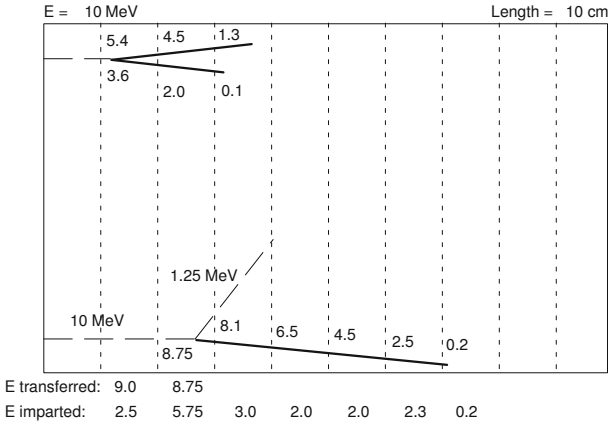


Fig. 15.31 The difference between energy transferred and energy imparted. Two of the photons from Fig. 15.30b are shown. The water has been divided into ten 1-cm slices. The numbers on the drawing show the charged-particle energy at the entrance to each slice. The energy transferred and the energy imparted in each slice are shown at the bottom

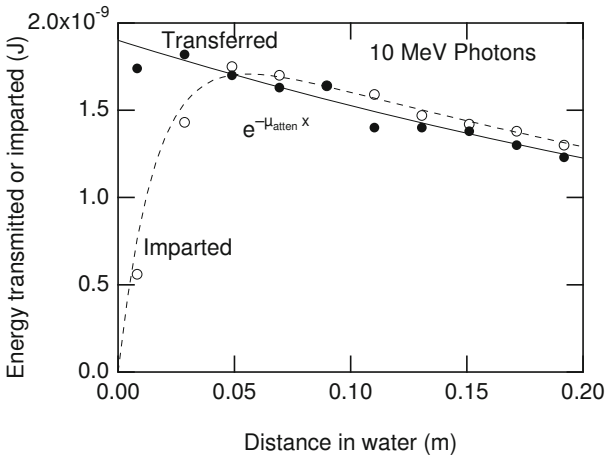


Fig. 15.32 Plot of energy transferred and energy imparted for a simulation using 40,000 photons of energy 10 MeV. The filled circles are the energy transferred in each slice, and the open circles are the energy imparted in each slice

determine the energy of each charged particle as it enters the slice. For the Compton scattering, 8.75 MeV is transferred to the recoil electron and the scattered photon has 1.25 MeV. The energy imparted by the 5.4-MeV particle is $5.4 - 4.5 = 0.9$ MeV in slice 2, $4.5 - 1.3 = 3.2$ MeV in slice 3, and 1.3 MeV in slice 4. Similar calculations can be done for the other charged particles. The energy transferred and the energy imparted in each slice are shown at the bottom of Fig. 15.31. This ignores any interaction of the 1.25-MeV Compton-scattered photon and assumes it leaves the volume of interest. Because for the 100-keV photons the range of the charged particles is small compared to 1 cm, the energy transferred and the energy imparted in each slice are the same in Fig. 15.30a.

Figure 15.32 shows a plot of the transferred and imparted energy for a uniform beam of 10-MeV photons all traveling to the right and striking a slab of water 20 cm thick. Both the energy transferred and the energy imparted are stochastic quantities. This simulation was done for 40,000 photons, and you can see the scatter in the points. The energy transferred falls exponentially as $\exp(-\mu_{\text{atten}} x)$.

15.15.2 Energy Transferred and Kerma

We found the energy transferred by calculating the energy of each electron or positron produced. The standard definition uses slightly different bookkeeping. It subtracts the energy of the photons leaving the volume of interest from those entering, and adds a term Q for the energy going into the volume due to changes in rest mass. For example, this is the $2m_e c^2$ of Eq. 15.63. The standard definition is

$$E_{\text{tr}} = (R_{\text{in}})_u - (R_{\text{out}})_u^{\text{nonr}} + Q. \tag{15.64}$$

The quantity R is radiant energy: the energy of particles (including photons) but not including rest energy. The subscript u means that it is the radiant energy of uncharged particles. The uncharged particles can be photons or neutrons.[14] Later we will use subscript c to denote the radiant energy of charged particles. The superscript "nonr" means that the quantity does not include radiant energy arising from bremsstrahlung or positron annihilation in flight from charged particles within the volume. The Q term is positive if mass is converted to energy (as in annihilation radiation) and negative if energy is converted to mass (as in pair production).

Using this method of calculating for Fig. 15.31 gives

$$E_{\text{tr}} = (R_{\text{in}})_u - (R_{\text{out}})_u^{\text{nonr}} + \sum Q = 10 - 0 - 2 \times 0.511 = 9.0\,\text{MeV}$$

for slice 2. For the third slice the equation gives

$$E_{\text{tr}} = (R_{\text{in}})_u - (R_{\text{out}})_u^{\text{nonr}} + \sum Q = 10 - 1.25 + 0 = 8.75\,\text{MeV}.$$

For the fourth slice, the uncharged radiant energy in is equal to the uncharged radiant energy out. In the fifth slice, if

[14] Neutrinos, which we will discuss in Chap. 17, travel such long distances without interacting that they are not considered in the calculations. Energy carried by neutrinos, which come from nuclear β decay, is assumed to have left the body.

the 1.25-MeV photon actually interacts as it appears to, we would have to include its energy transfer. In all the other slices the energy transferred is zero.

The energy transferred is a stochastic quantity, and so is the energy transferred per unit mass, dE_{tr}/dm. Its expectation value is the *kerma* (*k*inetic *e*nergy *r*eleased per unit *ma*ss):

$$K = \frac{d\overline{E_{\mathrm{tr}}}}{dm}. \tag{15.65}$$

If we consider monoenergetic photons of energy $h\nu$ and consider only the interaction of the primary photon beam (not any secondary photons, such as Compton-scattered photons or annihilation radiation), then the kerma is

$$K = \frac{\mu_{\mathrm{tr}}}{\rho}\Psi, \tag{15.66}$$

where Ψ is the energy fluence. To see why this is true, note that if the N photons are spread over area S, then $NE = \Psi S$ and $dm = \rho S dx$. The kerma is

$$K = \frac{d\overline{E_{\mathrm{tr}}}}{dm} = \frac{\Psi S \mu_{\mathrm{tr}} dx}{\rho S dx} = \frac{\mu_{\mathrm{tr}}}{\rho}\Psi.$$

15.15.3 Energy Imparted and Absorbed Dose

The *energy imparted,* E, is the net energy into the volume from all sources: uncharged particles, charged particles, and changes of rest mass:

$$E = (R_{\mathrm{in}})_u - (R_{\mathrm{out}})_u + (R_{\mathrm{in}})_c - (R_{\mathrm{out}})_c + \sum Q. \tag{15.67}$$

The *absorbed dose* is the expectation value of the energy imparted per unit mass:

$$D = \frac{d\overline{E}}{dm}. \tag{15.68}$$

It is measured in joules per kilogram or *gray* (Gy).

15.15.4 Net Energy Transferred, Collision Kerma and Radiative Kerma

Another quantity used in the literature is the *net energy transferred*. It subtracts from the energy transferred the energy that is reradiated (bremsstrahlung and radiation from positron annihilation in flight), even if the reradiation takes place outside the volume of interest. It is

$$E_{\mathrm{tr}}^{\mathrm{net}} = (R_{\mathrm{in}})_u - (R_{\mathrm{out}})_u^{\mathrm{nonr}} - R_u^r + \sum Q. \tag{15.69}$$

The *collision kerma* and *radiative kerma* are defined as expectation values per unit mass:

$$\begin{aligned} K_C &= \frac{d\overline{E_{\mathrm{tr}}^{\mathrm{net}}}}{dm} = K - K_r, \\ K_r &= \frac{d\overline{R_u^r}}{dm}. \end{aligned} \tag{15.70}$$

Considering only a primary beam of monoenergetic photons,

$$K_C = \frac{\mu_{\mathrm{en}}}{\rho}\Psi. \tag{15.71}$$

15.16 Charged-Particle Equilibrium

There are three equilibrium conditions that sometimes exist or are assumed to exist, that make it possible to calculate the relationship between energy transferred and energy imparted.

15.16.1 Radiation Equilibrium

The first and most restrictive condition is *radiation equilibrium*. It is a useful model when considering an extended radioactive source that is distributed uniformly throughout some volume V (such as the body or a particular organ). The source is assumed to emit its radiation isotropically. The energy released to neutrinos is ignored. A point of interest within the large volume is surrounded by a smaller volume v. The volume v must be far enough from the edge of V so that any radiation emitted from v is absorbed before reaching the surface of V. The entire volume V is assumed to be of the same atomic composition and density. Because everything is isotropic, on average for every photon or neutron or charged particle entering volume v, another identical one leaves. This means that

$$\left(\overline{R}_{\mathrm{in}}\right)_c = \left(\overline{R}_{\mathrm{out}}\right)_c \tag{15.72a}$$

and

$$\left(\overline{R}_{\mathrm{in}}\right)_u = \left(\overline{R}_{\mathrm{out}}\right)_u. \tag{15.72b}$$

The average energy imparted is

$$\overline{E} = \sum \overline{Q}. \tag{15.73}$$

When the conditions for radiation equilibrium are satisfied, the absorbed dose is the expectation value of the energy released by the radioactive material per unit mass. If there is no radioactive source, there is no energy imparted in radiation equilibrium.

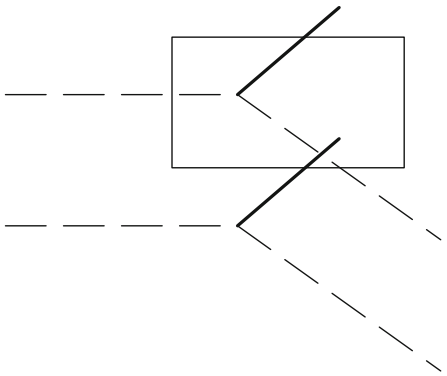


Fig. 15.33 One of the conditions for charged-particle equilibrium is that on average, for every charged particle of a certain energy leaving volume v traveling in a certain direction, a corresponding particle enters the volume

15.16.2 Charged-Particle Equilibrium

A less restrictive assumption is *charged-particle equilibrium*, in which only Eq. 15.72a is satisfied: the average amount of charged-particle radiant energy entering the region is the same as the average amount leaving. The assumption of charged particle equilibrium is a useful model in several cases, but we will consider only the case of an external beam of photons striking volume V. Again we consider what happens in a smaller volume v, separated from the boundary of V by a distance larger than the maximum range of any secondary charged particles produced by the external radiation. We also assume that the medium is homogeneous and that only a small fraction of the primary radiation interacts within the volume so attenuation can be neglected. Then the average number of charged particles produced per unit volume and per unit solid angle in any given direction is the same everywhere in the volume. Though the charged particles need not be produced isotropically, on average for every particle that leaves volume v, a corresponding one will enter it, as shown in Fig. 15.33. For charged-particle equilibrium, the average energy imparted is

$$\overline{E} = \left(\overline{R}_{\text{in}}\right)_u - \left(\overline{R}_{\text{out}}\right)_u + \sum \overline{Q}.$$

Comparing this with the average of Eq. 15.69 shows that the average net energy transferred is

$$\overline{E_{\text{tr}}^{\text{net}}} = \overline{E} + \left(\overline{R}_{\text{out}}\right)_u - \left(\overline{R}_{\text{out}}\right)_u^{\text{nonr}} - \overline{R_u^r}.$$

Now recall that $\left(\overline{R}_{\text{out}}\right)_u$ is the average value of all the uncharged radiation leaving volume v, $\left(\overline{R}_{\text{out}}\right)_u^{\text{nonr}}$ is the average value of all uncharged radiation leaving excluding bremsstrahlung and photons from annihilation in flight that occur within the volume, and $\overline{R_u^r}$ is the bremsstrahlung and annihilation-in-flight radiation from charged particles originating in v regardless of where it occurs. If there is charged-particle equilibrium, any radiative interaction by a charged particle after it leaves the volume will on average be replaced by an identical interaction inside v. If the volume is small enough so that all radiative loss photons escape from the volume before undergoing any subsequent interactions, then

$$\left(\overline{R}_{\text{out}}\right)_u = \left(\overline{R}_{\text{out}}\right)_u^{\text{nonr}} + \overline{R_u^r}.$$

Therefore, *for charged-particle equilibrium*, $\overline{E} = \overline{E_{\text{tr}}^{\text{net}}}$, and the dose is equal to the collision kerma:

$$D = K_C. \tag{15.74}$$

One situation where charged-particle equilibrium applies is for the thin slices in Fig. 15.30a. The electron ranges are so short (10 μm for a 25-keV electron) that a slice can be thin compared to $1/\mu$ and yet all the electrons produced stay within the volume.

The conditions for charged-particle equilibrium fail if the source of photons is too close (Ψ is not uniform because of $1/r^2$), close to a boundary (as between air and tissue or muscle and bone), for high-energy radiation (as in Fig. 15.32), or if there is an applied electric or magnetic field that alters the paths of the charged particles (as in some radiation detectors).

In Fig. 15.32, if we look at the situation far enough to the right, the energy imparted is proportional to the energy transferred. This situation is called *transient charged-particle equilibrium.*

The dose for a monoenergetic parallel beam of charged particles with particle fluence Φ passing through a thin layer can be calculated by making three assumptions:

1. The volume of interest is thin enough so that S_e remains constant.
2. Scattering can be neglected, so every particle passes straight through the layer.
3. The net kinetic energy carried out of the layer by δ rays is negligible, either because the layer is thick compared to the range of the δ rays or because the layer is immersed in a material of the same atomic number so that charged-particle equilibrium exists.

Then the energy lost in collisions in a layer of thickness dz is $E = \Phi(\text{area})(S_e/\rho)\rho dz$ and the mass is $\rho(\text{area})dz$, so the dose is

$$D = \frac{S_e}{\rho}\Phi. \tag{15.75}$$

Attix (1986, pp. 188–195) discusses corrections for situations where these assumptions are not valid.

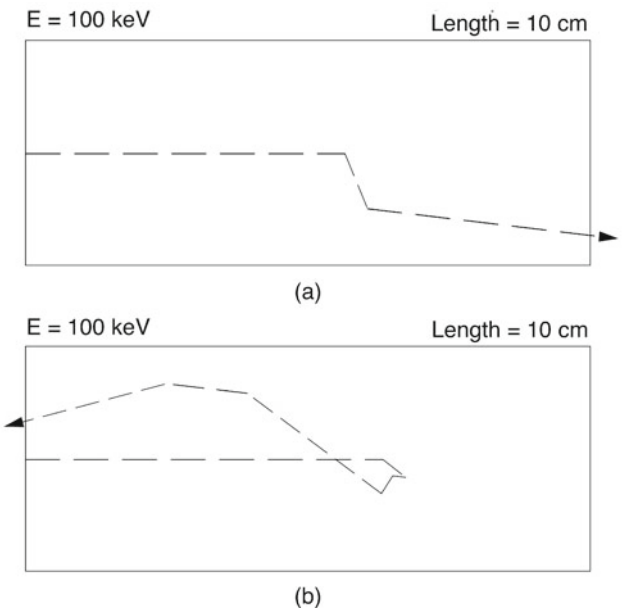


Fig. 15.34 Secondary photons also interact in this simulation. One 100-keV photon enters from the left in each panel. **a** The primary photon undergoes a Compton scattering. The Compton-scattered photon also undergoes a Compton scattering. The third photon escapes from the water. **b** The primary photon is Compton scattered. Each Compton-scattered photon undergoes another Compton scattering, until the sixth scattered photon leaves through the upstream surface of the water, traveling nearly in the direction from which the incident photon came

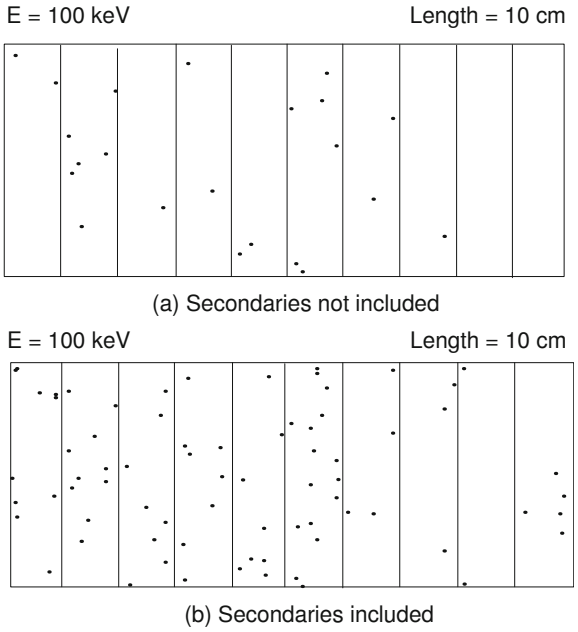


Fig. 15.35 Twenty-five 100-keV photons entered the water from the left. The dots represent recoil electrons from Compton scattering or photoelectrons. **a** Only the first interaction of the primary photon is considered. **b** Subsequent interactions are also considered

15.17 Buildup

We have been ignoring the interactions of secondary photons, primarily Compton-scattered photons and annihilation radiation. They can be quite significant. In fact, there can be a cascade of several generations of photons, though we will call them all *secondary photons*. Figure 15.34 compares two simulations in which the secondary photons are allowed to interact. In Fig. 15.34a there is one secondary interaction before the scattered photon escapes from the water. In Fig. 15.34b there are a total of six Compton scatterings before the secondary photon escapes.

All of these secondary photons produce electrons that contribute to the energy transferred and energy imparted. Figure 15.35 compares two cases where 25 photons of energy 100 keV enter the water from the left. The primary interactions are the same in both cases. In Fig. 15.35a the small dots represent the electrons produced by the interaction of the primary photons. In Fig. 15.35b the electrons produced by secondary and subsequent interactions are also shown. The energy transferred and energy imparted are much greater.

The *buildup factor* for any quantity is defined as the ratio of the quantity including secondary and scattered radiation

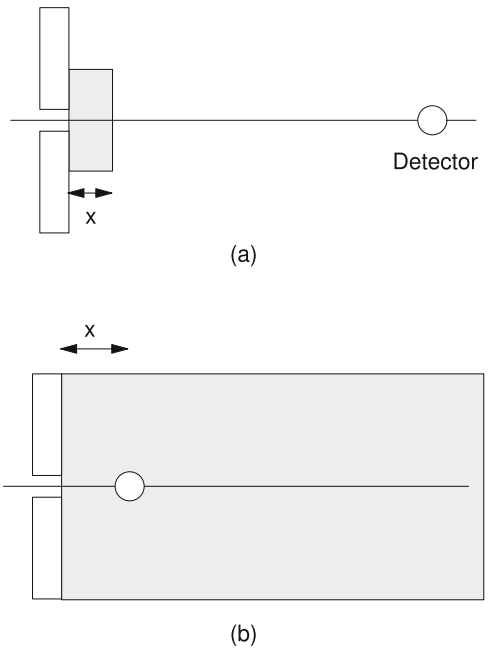


Fig. 15.36 Two different detector geometries. **a** The detector is at a fixed location and the absorber thickness is increased. **b** The detector is at a varying distance from the source in a water bath

to the quantity for primary radiation only. For example, if the primary beam has an energy fluence Ψ_0 at the surface, the energy fluence at depth x in the medium is

$$\Psi(x) = B(x)\Psi_0 e^{-\mu x}. \tag{15.76}$$

The buildup factor is quite sensitive to the geometry. Compare the two situations in Fig. 15.36. In Fig. 15.36a the detector is at a fixed location and the thickness of the absorber is increased. As the absorber thickness x approaches zero, the buildup factor approaches unity. In Fig. 15.36b the detector is at depth x in a water bath. Because of the backscattered radiation seen in Fig. 15.34b, $B(x) > 1$ as $x \to 0$. In this case, it is sometimes called the *backscatter factor*. For further discussion, see Attix (1986).

Symbols Used in Chap. 15

Symbol	Use	Units	First used page
a	Acceleration	m s^{-2}	440
a_i	Number of atoms of constituent i		434
b	Impact parameter	m	442
c	Velocity of light	m s^{-1}	428
d	Diameter	m	440
e	Charge on electron	C	425
f, f_C, f_i, f_κ, f_τ	Fraction of photon energy transferred to charged particles		431
g	Fraction of photon energy of secondary electrons converted back into photons by bremsstrahlung		437
h	Planck's constant	J s	427
j	Total angular momentum quantum number		425
k	Spring constant	N m^{-1}	440
l	Orbital angular momentum quantum number		425
m	Mass	kg	430
m_e	Electron rest mass	kg	428
m_j	Quantum number for the component of the total angular momentum along the z axis		425
m_0	Rest mass	kg	429
n	Principal quantum number		425
n	Number		434
p	Momentum	kg m s^{-1}	429
q	Charge	C	442
r	Position	m	444
r_e	"Classical" electron radius	m	430
s	Spin quantum number		425
t	Time	s	440
v	Velocity	m s^{-1}	439
w_i	Mass fraction of constituent i		434
x, y, z	Coordinate axes	m	433
x	Dimensionless energy ratio		430
z	Charge of projectile in multiples of e		439
A	Atomic mass number		433
A_i, A_{mol}	Atomic mass number of constituent i or molecule	(g mol)$^{-1}$ or (kg mol)$^{-1}$	434
A_K	Auger yield		435
B, B_K, etc.	Binding energy	eV or J	427
B	Buildup factor or backscatter factor		455
C	Shell correction coefficient		444
D	Absorbed dose	J kg^{-1} or Gy (gray)	452
E	Energy	J	426
E	Electric field	V m^{-1}	430
F	Force	N	440
F	Fraction of charged particles passing through an absorber		447
I	Average ionization energy	eV or J	444
I	Stopping interaction strength	J m^2	444
K, K_C	Kerma, collision kerma	J kg^{-1} or Gy (gray)	452
L	Stopping number per atomic electron		444
L_Δ	Restricted linear stopping power	J m^{-1}	447
M	Mass	kg	434
N	Number of particles		433
N_A	Avogadro's number	mol^{-1}	433
N_T	Number of target atoms per unit projected area	m^{-2}	434
N_{TV}	Number of target atoms per unit volume	m^{-3}	434
Q	Energy released from rest mass	J	451
R	Range	m	447
R_u, R_c	Radiant energy in the form of uncharged or charged particles	J	451
S	Area	m^2	452
S	Stopping power	J m^{-1}	439
S_e	Electron (collision) stopping power	J m^{-1}	440
S_n	Nuclear stopping power	J m^{-1}	440
S_r	Radiative stopping power	J m^{-1}	440
T	Kinetic energy	J	427
V	Volume	m^3	434
V	Velocity	m s^{-1}	443
$W_{K,L,M}$	Probability that a hole in the K, L, or M shell is filled by fluorescence		435
W	Energy lost in a single interaction	J	439
Y	Radiation yield		448
Z	Atomic number of target atom		425
β	v/c		439
δ	Average energy emitted as fluorescence radiation per photon absorbed	J	437
δ	Density-effect correction		444
ϵ	Stopping cross section	J m^2	439
ϵ_0	Electrical permittivity of free space	N^{-1} C^2 m^{-2}	430
θ, ϕ	Angles		429

κ	Pair production cross section	m^2	428
λ	Wavelength	m	429
μ, μ_{atten}	Attenuation coefficient	m^{-1}	433
μ_{en}	Energy absorption coefficient	m^{-1}	437
μ_{tr}	Energy transfer coefficient	m^{-1}	437
ν	Frequency	Hz	427
ξ	Position	m	443
ρ	Density	kg m^{-3}	433
σ_C	Total Compton cross section for one electron	m^2	427
σ_{coh}	Coherent Compton cross section for one atom	m^2	427
σ_{incoh}	Incoherent Compton cross section for one atom	m^2	427
σ_{tr}	Transfer cross section	m^2	431
σ_{tot}	Total cross section	m^2	433
τ	Photoelectric cross section	m^2	428
Δ	Energy transfer	J	447
Φ	Particle fluence	m^{-2}	436
Ψ	Energy fluence	J m^{-2}	436
Ω	Solid angle	sr	430

Problems

Section 15.1

Problem 1. The quantum numbers $m_s = \pm\frac{1}{2}$ and $m_l = l, l-1, l-2 \ldots, -l$ are sometimes used instead of j and m_j to describe an electron energy level. Show that the total number of states for given values of n and l is the same when either set is used.

Problem 2. Use Eq. 15.2 to estimate the K-shell energies for the following elements and compare them to the measured values of E_K.

Z	Element	Measured E_K (keV)
6	Carbon	0.284
20	Calcium	4.04
53	Iodine	33.2
82	Lead	88.0

Section 15.2

Problem 3. In the photoelectric effect, assume that the ejected electron has mass m and speed v and that the recoiling atom has mass M and speed V. Show that if the two particles have the same momentum, the kinetic energy of the atom is smaller than the kinetic energy of the electron by a factor (m/M). Calculate this factor for carbon and verify the claim in the text that the kinetic energy of the recoiling atom is small. Use non-relativistic expressions for the momentum and kinetic energy.

Section 15.3

Problem 4. The K-shell photoelectric cross section for 100-keV photons on lead ($Z = 82$) is $\tau = 1.76 \times 10^{-25}$ m^2 $atom^{-1}$. Estimate the photoelectric cross section for 60-keV photons on calcium ($Z = 20$).

Problem 5. Write Eq. 15.8 as $\tau = CZ^4E^{-3}$, where C is a proportionality constant. Estimate C from Fig. 15.3 (use $E = 100$ keV). Be sure to determine both the value of C and its units.

Problem 6. Describe how you could use different materials to determine the energy of monoenergetic x rays of energy about 50 keV by using changes in the attenuation coefficient. What materials would you use?

Section 15.4

Problem 7. Derive Eq. 15.11 from the preceding four equations.

Problem 8. Derive an equation for the direction of the recoil electron, ϕ, in terms of θ and λ_0.

Problem 9. A 1-MeV photon undergoes Compton scattering from a carbon target. The scattered photon emerges at an angle of 30°.

(a) What is the energy of the scattered photon? What is the energy of the recoil electron?

(b) What is the differential scattering cross section for scattering at an angle of 30° from one electron? From the entire carbon atom ($Z = 6$, $A = 12$)?

Problem 10. When $h\nu 0 >> mec^2$, what is the energy of a Compton-scattered photon at 180°? at 90°?

Problem 11. Integrate Eq. 15.16 over all possible scattering angles to obtain Eq. 15.18. Use the solid angle in spherical coordinates (Appendix L) and the substitution $u = 1 + x(1 - \cos\theta)$.

Problem 12. Integrate Eq. 15.17 over all possible scattering angles to obtain Eq. 15.19. Use the solid angle in spherical coordinates (Appendix L).

Problem 13. Find the limit of Eq. 15.16 as $x \to \infty$.

Problem 14. Write Eq. 15.16 in the form

$$\frac{d\sigma_C}{d\Omega} = \frac{r_e^2}{2}\left(1 + \cos^2\theta\right) F_{KN},$$

where F_{KN} is the *Klein–Nishina factor*. Find an expression for F_{KN} in terms of θ and x. Show that as $x \to 0$, $F_{KN} \to 1$. Show that when $\theta = 0$, $F_{KN} = 1$.

Problem 15. Use the expansion $\ln(1+x) = x - x^2/2 + x^3/3$ to show that Eq. 15.18 approaches Eq. 15.19 as $x \to 0$.

Problem 16. Eq. 15.11 shows that the wavelength shift is independent of the wavelength of the incident photon. Calculate the fractional wavelength shift $(\lambda - \lambda_0)/\lambda_0$ for an infrared photon ($\lambda_0 = 10$ μm), an ultraviolet photon ($\lambda_0 = 100$ nm), a low-energy ("soft") x ray ($\lambda_0 = 1$ nm), and a high-energy x ray ($\lambda_0 = 0.01$ nm).

Problem 17. Suppose that attenuation is measured for 60-keV photons passing through water in such a way that photons scattered less than 5° still enter the detector. Estimate the incoherent Compton scattering cross section per electron for photons scattered through more than 5°.

Section 15.5

Problem 18. A beam of 59.5-keV photons from ^{241}Am scatters at 90° from some calcium atoms ($A = 40$).

(a) What is the energy of a Compton-scattered photon?
(b) What is the energy of a coherently scattered photon?
(c) What is the recoil energy of the atom in coherent scattering?

Section 15.6

Problem 19. Show that a single photon cannot produce an electron–positron pair in free space because energy and momentum cannot be simultaneously conserved.

Section 15.7

Problem 20. Most diagnostic x rays use photon energies in the range 20–100 keV. For carbon (Fig. 15.2), which mechanism is most important in this range: photoelectric effect, Compton scattering, coherent scattering, or pair production?

Problem 21. Use Fig. 15.7 to make the following estimates for 1-MeV photons. What is the mass attenuation coefficient for water? For aluminum? For lead? What is the linear attenuation coefficient in each case?

Problem 22. Use Fig. 15.7 to estimate the attenuation coefficient for 0.1-MeV photons on carbon and lead. Compare your results to values you obtain from the internet or the literature. Repeat for 1-MeV photons.

Problem 23. Consider photons of three energies: 0.01, 0.02, and 0.03 MeV. What fraction of the photons at each energy will be unattenuated after they pass through 0.1 mm of lead ($\rho = 11.35$ g cm^{-3})? Comment on the differences in your results.

Section 15.8

Problem 24. Use Fig. 15.7 to find the mass attenuation coefficient for 0.2-MeV photons in a polyethylene absorber. The Compton effect predominates. Polyethylene has the formula $(CH_2)_n$.

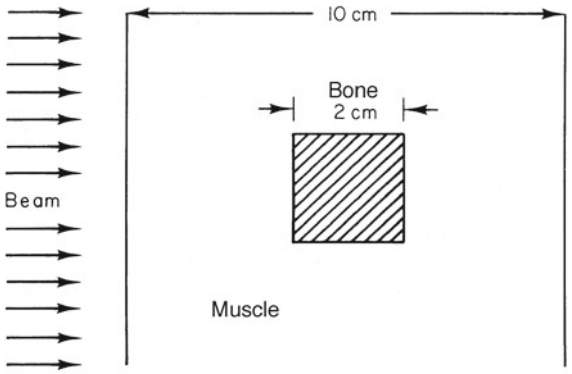
Problem 25. What will be the attenuation of 40-keV photons in muscle 10 cm thick? Repeat for 200-keV photons.

Problem 26. Assume that a patient can be modeled by a slab of muscle 20 cm thick of density 1 g cm^{-3}. What fraction of an incident photon beam will emerge without any interaction if the photons have an energy of 10 keV? 100 keV? 1 MeV? 10 MeV?

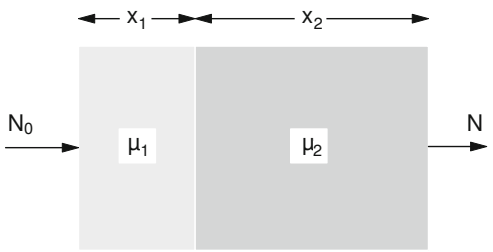
Problem 27. Muscle and bone are arranged as shown. Assume the density of muscle is 1.0 g cm^{-3} and the density of bone is 1.8 g cm^{-3}. The attenuation coefficients are

E	$(\mu/\rho)_{\text{muscle}}$ (cm^2 g^{-1})	$(\mu/\rho)_{\text{bone}}$ (cm^2 g^{-1})
60 keV	0.200	0.274
1 MeV	0.070	0.068

Compare the intensity of the emerging beam that has passed through bone and muscle and just muscle at the two energies.



Problem 28. A beam of monoenergetic photons travels through a sample made up of two different materials of unknown thickness x_1 and x_2, as shown below. The attenuation coefficients at two different energies, E_a and E_b, are accurately known. They are $\mu_1(a)$, $\mu_2(a)$, $\mu_1(b)$, and $\mu_2(b)$. One measures accurately the log of the ratio of the number of photons emerging from the sample to the number entering, $R = \ln(N_0/N)$, at each energy so that R_a and R_b are known. Find an expression for x_2 in terms of R_a, R_b, and the attenuation coefficients.



Section 15.9

Problem 29. The text showed that the mass atenuation coefficient for incoherent scattering is nearly independent of Z (assuming (Z/A) is constant). Show how the mass attenuation coefficient depends on Z for the photoelectric effect (use Eq. 15.8) and for pair production (assume $\kappa \propto Z^2$).

Problem 30. Use Table 15.1 to calculate the photon energies of the 13 spectral lines for tungsten that are shown in Fig. 15.13.

Problem 31. Use Table 15.1 and the selection rules in Eq. 15.32 to determine the allowed spectral lines for transitions in tungsten between M and N levels. Construct a drawing like that in Fig. 15.13 showing these transitions. Be sure to include the N_{VI} and N_{VII} levels in your drawing. Give these two levels slightly different energies so you can distinguish the transitions.

Problem 32. Use data from http://webelements.com/lead/orbital_properties.html to create a table like Table 15.1 and figures like Figs. 15.1 and 15.13, but for lead instead of tungsten.

Problem 33. You wish to use x-ray fluorescence to detect lead that has been deposited in a patient's bone. You shine 100-keV photons on the patient's bone and want to detect the 73-keV fluorescence photons that are produced. The incident photon fluence is $\Phi_0 = 10^{12}$ photons m^{-2}. There are 10^{14} lead atoms ($\approx$ 1 nmol) in the region illuminated by the incident beam. The photoelectric cross section is 1.76×10^{-25} m^2 atom^{-1}. The fluorescence yield is $W = 0.94$. Assume for simplicity that the fluorescence photons are emitted uniformly in all directions. The detector has a sensitive area 1×2 cm and is located 10 cm from the lead atoms. How many fluorescence photons are detected?

Section 15.10

Problem 34. A 5-keV photon strikes a calcium atom. The following events take place:

1. A K-shell photoelectron is ejected.
2. A K_α photon is emitted. This corresponds to the movement of a hole from the K shell to the L shell.
3. An electron in the M shell goes to the L shell and an M-shell electron is emitted.

Give the excitation energy of the atom, the total energy in the form of photons, and the total energy in the form of electron kinetic energy at each stage. Use the following data for calcium: $Z = 20$, $A = 40$, $B_K = 4000$ eV, $B_L = 300$ eV (ignore differences in subshells), $B_M = 40$ eV.

Problem 35. The following are the binding energies for hydrogen and oxygen.

	H	O
B_K	13.6 eV	532 eV
B_L		24 eV

(a) Determine f_τ for hydrogen from first principles.
(b) Use Eqs. 15.38 and 15.39 to estimate f_τ (K shell) for oxygen.

Problem 36. Use the Thomson scattering cross section, $d\sigma/d\Omega = (r_e^2/2)(1 + \cos^2\theta)$, the total cross section $\sigma_C = 8\pi r_e^2/3$, and the expression for the total energy of the recoil electron Eq. 15.15 to find an expression for f_C as $x \to 0$. Compare some values of $f_C \sigma_{C\,\text{incoh}}$ with the plot in Fig. 15.7.

Problem 37. (a) For 50-keV photons on calcium, estimate f_τ.
(b) For 100-keV photons on calcium, the photoelectric cross section is $\tau = 5.89 \times 10^{-28}$ m^2 atom^{-1}. Use $f_\tau = 1.0$. Estimate μ_{tr}/ρ. Use the following data for calcium if you need them: $Z = 20$, $A = 40$, $B_K = 4000$ eV, $B_L = 300$ eV, $B_M = 40$ eV.

Section 15.11

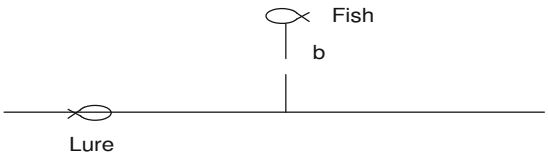
Problem 38. Prove that if a particle of mass M_1 and kinetic energy T collides head on with a particle of mass M_2 that is at rest, the energy transferred to the second particle is $4TM_2/M_1$ or $2M_2V^2$ in the limit $M_2 \ll M_1$. The maximum energy is transferred when the particles move apart along the line of motion of the incident particle.

Problem 39. The expression for $S_e = dT/dx$ has the SI units J m^{-1}. Therefore S_e/ρ in Eq. 15.56 has units J m^2 kg^{-1}.

(a) How must the coefficient in front of Eq. 15.56 be changed if T is in MeV? If x is in cm instead of m?
(b) What numerical factors must be introduced if N_A is in atoms per g mol and ρ is in g cm^{-3}?
(c) What is the average force on a 10-MeV proton in carbon? On a 100-keV proton? Use I=78 eV.
(d) What are the units of the stopping cross section (defined just below Eq. 15.46)?

Problem 40. The peak in the stopping power occurs roughly where the projectile velocity equals the velocity of the atomic electrons in the target. Find an expression for the velocity of an electron in the $n = 1$ Bohr orbit. Use Eq. 14.8, and the fact that the total energy is the sum of the kinetic and potential energies. Use the classical arguments and the fact that the electron is in a circular orbit to relate the kinetic and potential energies. The acceleration in a circular orbit is v^2/r.

Problem 41. A fishing lure is trolled behind a boat for a total distance D. Suppose that fish are distributed uniformly throughout the water at a concentration C fish m^{-3}, and that the probability of a fish striking the lure depends on b, the perpendicular distance from the path of the lure to the fish: $p = \exp(-b/b_0)$. Calculate the average number of fish caught.



Section 15.13

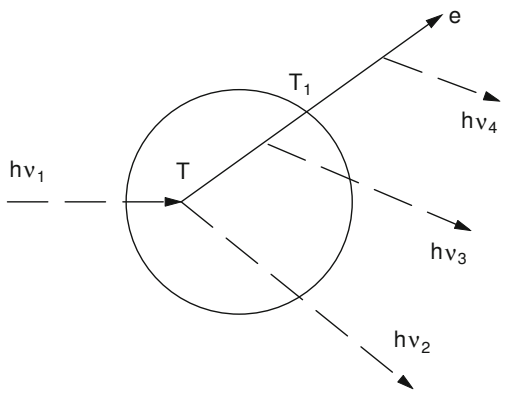
Problem 42. What is the range energy relationship for high-speed non-relativistic particles if the variation of L with T is neglected and Eq. 15.56 is the dominant term?

Problem 43. Estimate the maximum electron range, and hence the radius of the δ-ray cloud surrounding the track of a 5-MeV α particle. (The rest energy Mc^2 of an α particle is about 4 times 938 MeV.) The range of a low energy electron in cm is about $10^{-2}\beta^2$.

Section 15.15

Problem 44. Suppose that a photon of energy $h\nu$ enters a volume of material and produces an electron–positron pair. Both particles come to rest in the volume, and the positron annihilates with an electron that was already in the volume. Both annihilation photons leave the volume. Show that the formal definition of energy transfer agrees with the common-sense answer that it is the kinetic energy of the electron and positron, which is $h\nu - 2m_e c^2$. What is the energy imparted?

Problem 45. What are the energy transferred, the net energy transferred, and the energy imparted in the volume shown?



References

Ahlen SP (1980) Theoretical and experimental aspects of the energy loss of relativistic heavily ionizing particles. Rev Mod Phys 52:121–173

Arqueros F, Montesinos GD (2003) A simple algorithm for the transport of gamma rays in a medium. Am J Phys 71(1):38–45

Attix FH (1986) Introduction to radiological physics and radiation dosimetry. Wiley, New York

Bambynek W, Crasemann B, Fink RW, Freund HU, Mark H, Swift CD, Price RE, Rao PV (1972) X-ray fluorescence yields, Auger, and Coster–Kronig transition probabilities. Rev Mod Phys 44:716–813

Boone JM, Chavez AE (1996) Comparison of x-ray cross sections for diagnostic and therapeutic medical physics. Med Phys 23(12):1997–2005

Budd T, Marshall M (1983) Microdosimetric properties of electron tracks measured in a low-pressure chamber. Radiat Res 93:19–32

Carlsson GA, Carlsson CA, Berggren K-F, Ribberfors R (1982) Calculation of scattering cross sections for increased accuracy in diagnostic radiology. 1. Energy broadening of Compton-scattered photons. Med Phys 9(6):868–879

Hobbie RK (1992) MacDose: a simulation for understanding radiological physics. Comput Phys 6(4):355–359 (MacDose is available at the book web site https://files.oakland.edu/users/roth/web/hobbie.htm. It runs on a Macintosh computer using OS-9 or earlier)

Hobbie RK (2009) Photon interactions: a simulation study with MacDose. 26 minute video. https://itunes.apple.com/us/itunes-u/photon-interactions-simulation/id448438300?mt=10

Hubbell JH (1969) Photon cross sections, attenuation coefficients, and energy absorption coefficients from 10 keV to 100 GeV. NBS 29. US Govt. Printing Office, Washington DC

Hubbell JH (1982) Photon mass attenuation and energy absorption coefficients. Int J Appl Radiat Inst 33:1269–1290 http://dx.doi.org/10.1016/0020-708X(82)90248-4

Hubbell JH, Seltzer SM (1996) Tables of X-ray mass attenuation coefficients and mass energy-absorption coefficients 1 keV to 20 MeV for elements $Z = 1$ to 92 and 48 additional substances of dosimetric interest. National Institute of Standards and Technology. Report No. NISTIR 5632 Web Version. http://www.nist.gov/pml/data/xraycoef/index.cfm

Hubbell JH, Veigele WJ, Briggs EA, Brown RT, Cromer DT, Howerton RJ (1975) Atomic form factors, incoherent scattering functions and photon scattering cross sections. J Phys Chem Ref Data 4:471–538 (errata ibid. 6:615–616 (1977))

Hubbell JH, Gimm HA, Øverbø I (1980) Pair, triplet and total atomic cross sections (and mass attenuation coefficients) for 1 MeV–100 GeV photons in elements $Z = 1$ to 100. J Phys Chem Ref Data 9:1023–1147

Hubbell JH, Trehan PN, Singh N, Chand B, Mehta D, Garg ML, Garg RR, Singh S, Puri S (1994) A review, bibliography, and tabulation of K, L, and higher atomic shell x-ray fluorescence yields. J Phys Chem Ref Data 23(2):339–364

ICRU Report 16 (1970) Linear energy transfer. International Commission on Radiation Units and Measurements, Bethesda

ICRU Report 33 (1980) Radiation quantities and units. International Commission on Radiation Units and Measurements, Bethesda

ICRU Report 37 (1984) Stopping powers for electrons and positrons. International Commission on Radiation Units and Measurements, Bethesda. http://www.nist.gov/pml/data/radiation.cfm

ICRU Report 49 (1993) Stopping power and ranges for protons and alpha particles. International Commission on Radiation Units and Measurements, Bethesda. http://www.nist.gov/pml/data/star/index.cfm

Jackson JD (1999) Classical electrodynamics, 3rd edn. Wiley, New York

Jackson DF, Hawkes DJ (1981) X-ray attenuation coefficients of elements and mixtures. Phys Rep 70:169–233

Kissel LH, Pratt RH, Roy SC (1980) Rayleigh scattering by neutral atoms, 100 eV to 10 MeV. Phys Rev A 22:1970–2004

Powell CF, Fowler PH, Perkins DH (1959) The study of elementary particles by the photographic method. Pergamon, New York

Seltzer SM (1993) Calculation of photon mass energy-transfer and mass energy-absorption coefficients. Radiat Res 136:147–170

Tung CJ, Ashley JC, Ritchie RH (1979) Range of low-energy electrons in solids. IEEE Trans Nucl Sci NS-26:4874–4878

Ziegler JF, Manoyan JM (1988) The stopping of ions in compounds. Nucl Instrum Meth B B35:215–228

Ziegler JF, Biersack JP, Littmark U (1985) The stopping and range of ions in solids. Pergamon, New York

Medical Uses of X-Rays

16

X-rays are used to obtain diagnostic information and for cancer therapy. They are photons of electromagnetic radiation with higher energy than photons of visible light. Gamma rays are photons emitted by radioactive nuclei; except for their origin, they are identical to x-ray photons of the same energy. Section 16.1 describes the production of x-rays. Section 16.2 introduces some new quantities that are important for measuring how the absorbed photon energy relates to the response of a detector—which might be a film, an ionization chamber, or a chemical detector. Several detectors are introduced in Sect. 16.3: film, fluorescent screens, scintillation detectors, semiconductor detectors, thermoluminescent dosimeters (TLD), and digital detectors. Section 16.4 describes the diagnostic radiograph, and the following section discusses image quality, particularly the importance of noise in determining image quality. Section 16.6 provides a brief mention of angiography, and Sect. 16.7 discusses some of the special problems of mammography. Computed tomography with x-rays is discussed in Sect. 16.8. The final sections deal with the biological effects of x-rays, cancer therapy, dose, and the risk of radiation.

16.1 Production of X-Rays

When a beam of energetic electrons stops in a target x-rays are emitted. *Characteristic x-rays* have discrete photon energies and are produced after excitation of an atom by the electron beam. *Bremsstrahlung* (Sect. 15.11) is the continuous spectrum of photon energies produced when an electron is scattered by an atomic nucleus. Bremsstrahlung is responsible for most of the photons emitted by most x-ray tubes. The total bremsstrahlung radiation yield as a function of electron energy for various materials is shown in Fig. 16.1. High-Z materials are most efficient for producing x-rays. Tungsten ($Z = 74$) is often used as a target in x-ray tubes because it has a high radiation yield and can withstand high temperatures. For 100-keV electrons on tungsten, the radiation yield is about 1 %: most of the electron energy heats the target. We now consider these two processes in greater detail.

16.1.1 Characteristic X-Rays

Atomic energy levels are described in Sect. 14.3. The levels for tungsten are shown in Table 15.1 and Fig. 15.1. An electron bombarding a target can impart sufficient energy to a target electron to remove it from the atom, leaving an unoccupied energy level or *hole*. The deexcitation of the atom is described in Sect. 15.9. For a high-Z material with a hole in the K shell, the fluorescence yield is large (see Fig. 15.14). The hole is usually filled when an electron in a higher energy level drops down to the unoccupied level. As it does so, the atom emits a characteristic x-ray—a photon with energy equal to the difference in energies of the two levels. This leaves a new hole, which is then filled by an electron from a still higher level with the emission of another x-ray, or by an Auger cascade.

As a hole moving to larger values of n corresponds to a decrease of the total energy of the atom, it is customary to draw the energy-level diagram for holes instead of electrons, which turns the graph upside down, as in Fig. 16.2. The zero of energy is the neutral atom in its ground state. As this is a logarithmic scale, zero cannot be shown.

Creation of the hole requires energy to remove an electron. That energy is released when the hole is filled. Various possible transitions are indicated in Fig. 16.2. These transitions are consistent with these selection rules, which can be derived using quantum theory:

$$\Delta l = \pm 1, \qquad \Delta j = 0, \pm 1. \tag{16.1}$$

The transitions are labeled by the letters K, L, M, and so forth, depending on which shell the hole is in initially. Greek-letter subscripts distinguish the x-rays from transitions to different final states.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_16, © Springer International Publishing Switzerland 2015

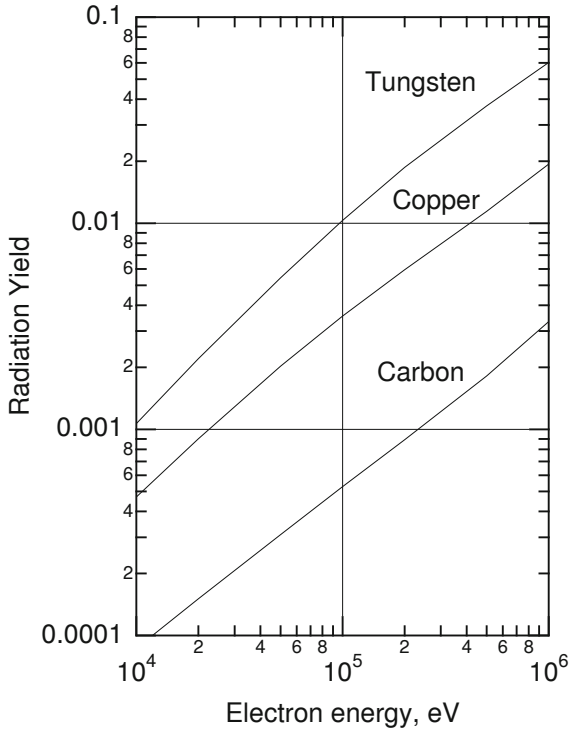


Fig. 16.1 Radiation yield vs. electron energy for carbon, copper, and tungsten. Plotted from data in ICRU Report 37 (1984). The radiation yield is the fraction of the electron's energy that is converted to photon energy; see Sect 15.13

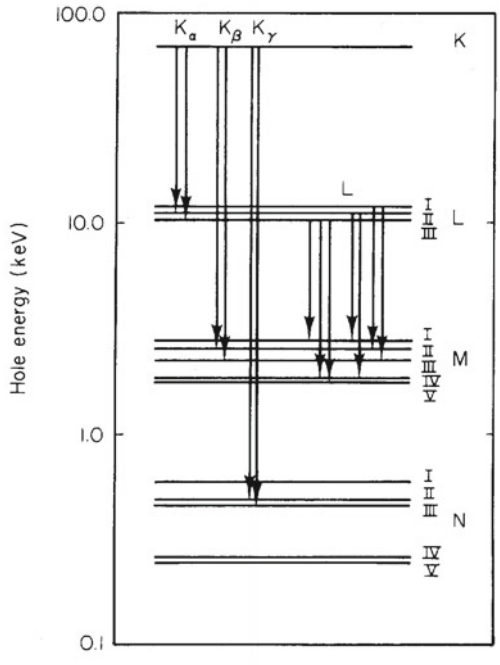


Fig. 16.2 Energy-level diagram for holes in tungsten and some of the x-ray transitions

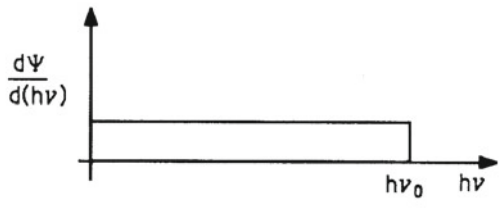


Fig. 16.3 The energy fluence spectrum of bremsstrahlung x-rays emitted when monoenergetic electrons strike a thin target

Analogous to the approximate formula of Eq. 15.2 is the following estimate of the energy of the K_α line (which depends on the screening for two values of n), which we have seen before as Eq. 15.33:

$$E_{K_\alpha} = \tfrac{3}{4}(13.6\,\text{eV})(Z-1)^2. \tag{16.2}$$

The factor $3/4$ is what one would have for hydrogen if $n_i = 2$ and $n_f = 1$ are substituted in the Bohr formula, Eq. 14.9. The screening also depends strongly on l.

16.1.2 Bremsstrahlung

The other mechanism for x-ray production is the acceleration of electrons in the Coulomb field of the nucleus, described in Sect. 15.11. Classically, a charged particle at rest creates an electric field that is inversely proportional to the square of the distance from the charge. When in motion with a constant velocity, it creates both an electric field and a magnetic field. When accelerated, additional electric and magnetic fields appear that fall off less rapidly—inversely with the first power of distance from the charge. This is classical electromagnetic radiation. Quantum mechanically, when a charged particle undergoes acceleration or deceleration, it emits photons. The radiation is called deceleration radiation, braking radiation, or *bremsstrahlung*. It has a continuous distribution of frequencies up to some maximum value.

The photon *energy* fluence spectrum of bremsstrahlung radiation from monoenergetic electrons passing through a thin target is constant from a maximum energy $h\nu_0$ down to zero, as shown in Fig. 16.3 (Attix 1986, p. 212). The maximum frequency is related to the kinetic energy of the electrons by $T = h\nu_0$, as one would expect from conservation of energy. (A photon of energy $h\nu_0$ is emitted when an electron loses all of its energy in a single collision).

For a thick target, we assume that all electrons at the same depth have the same energy (that is, we ignore straggling), and we ignore attenuation of photons coming out of the target. The spectrum is then the integral of a number of spectra

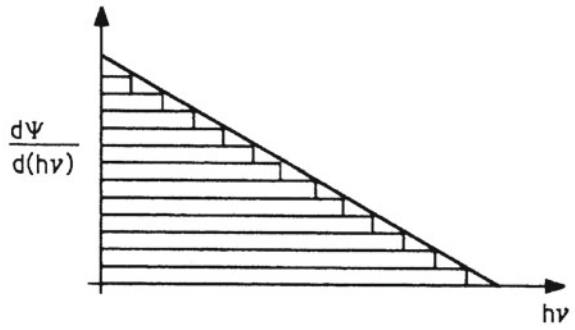


Fig. 16.4 The energy fluence spectrum of bremsstrahlung x-rays from a thick target, ignoring absorption of the photons in the target. The form is $d\Psi/d(h\nu) = CZ(h\nu_0 - h\nu)$

like that in Fig. 16.3. The thick-target spectrum is shown in Fig. 16.4. The spectral form is

$$\frac{d\Psi}{d(h\nu)} \equiv \frac{d\Psi}{dE} = CZ(h\nu_0 - h\nu) = CZ(T - E). \quad (16.3a)$$

The photon particle fluence spectrum is

$$\frac{d\Phi}{dE} = \frac{1}{h\nu}\frac{d\Psi}{dE} = CZ\left(\frac{h\nu_0}{h\nu} - 1\right). \quad (16.3b)$$

More of the low-energy photons that are generated within the target are attenuated as they escape, because of the much larger values of the attenuation coefficient at low energies (recall Figs. 15.10 and 15.11). This cuts off the low-energy end of the spectrum. If the electron energy is high enough, the discrete spectrum due to characteristic fluorescence is superimposed on the continuous bremsstrahlung spectrum. Both of these effects are shown in Fig. 16.5, which compares calculations and measurements of the particle fluence spectrum $d\Phi/dE$.

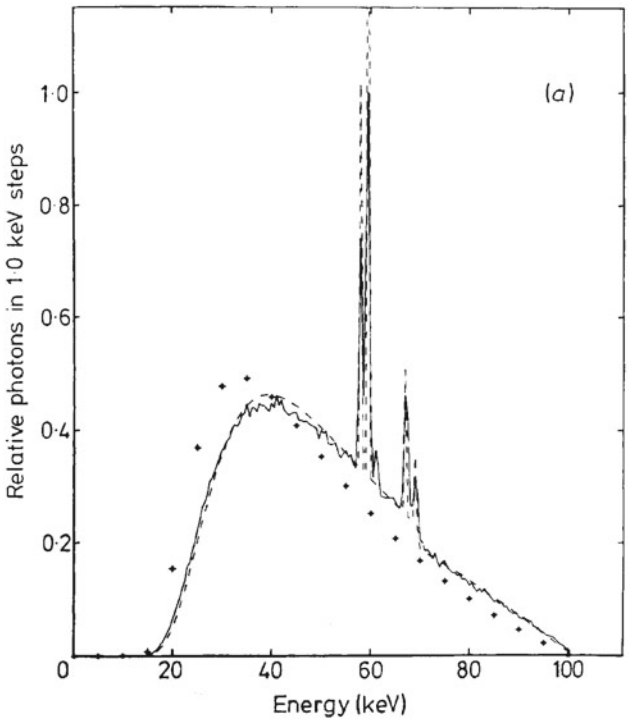


Fig. 16.5 Theoretical and measured photon particle spectra, $d\Phi/d(h\nu)$, for 100-keV electrons striking a thick tungsten target. The solid line represents measurements with a high-resolution semiconductor detector. The dashed line is the theory of Birch and Marshall (1979), which takes photon absorption into account. The crosses show an earlier theoretical model. (From Birch and Marshall 1979. Used by permission of the Institute of Physics)

16.2 Quantities to Describe Radiation Interactions: Radiation Chemical Yield, Mean Energy Per Ion Pair, and Exposure

Section 15.15 introduced the quantities energy transferred, energy imparted, kerma, and absorbed dose, which are used to describe radiation and its effects. This section introduces some additional quantities (ICRU Report 33 1980, Reprinted 1992) that are used to describe the interaction of the radiation with the detectors discussed in Sect. 16.3.

16.2.1 Radiation Chemical Yield

The *radiation chemical yield G* of a substance is the mean number of moles $\overline{n}$ of the substance produced, destroyed, or changed in some volume of matter, divided by the mean energy imparted to the matter:

$$G = \frac{\overline{n}}{\overline{E}}. \quad (16.4)$$

Its units are mol J^{-1}. (A related quantity expressed in non-SI units is the *G value*, expressed in molecules or moles per 100 eV of energy imparted.) The radiation chemical yield is particularly useful for describing chemical dosimeters. These are usually dilute aqueous solutions, so the radiation chemical yield of water is the important parameter.

16.2.2 Mean Energy per Ion Pair

Other detectors measure ionization produced in a gas by the radiation. The *mean energy expended in a gas per ion pair*

Table 16.1 Some representative values of the average energy per ion pair, W

Gas	W (eV per ion pair for electrons[a])
He	41.3
Ar	26.4
Xe	22.1
Air	33.97[b]
Semiconductors	**W (eV per electron-hole pair)**
Si	3.68
Ge	2.97

[a] From ICRU Report 31 (1979)
[b] ICRU Report 39 (1979) recommends 33.85 J C^{-1}. Attix (1986) uses 33.97 J C^{-1}. Note that 1 J C^{-1} is equivalent to 1 eV per singly-charged ion pair

formed, W, is

$$W = \frac{T_0}{\overline{N}_i}, \tag{16.5}$$

where T_0 is the initial kinetic energy of a charged particle and $\overline{N}_i$ is the mean number of ion pairs formed when T_0 is completely dissipated in the gas. The units are joules or electron volts per ion pair. The mean energy expended per ion pair is not equal to the ionization energy. To see why, consider three processes that can take place. The first is ionization, with $\overline{E}_i$ being the average energy of an ionized atom. Second, the collision may promote an atomic electron to an excited state without ionization. The average excitation energy is $\overline{E}_{\text{ex}}$. Finally, the charged particle may lose energy without producing ionization, a process called *subexcitation*. The average subexcitation energy is defined to be the energy lost by this process, E_{se}, divided by $\overline{N}_i$. Conservation of energy for this model leads to

$$T_0 = \overline{N}_i \overline{E}_i + \overline{N}_{\text{ex}} \overline{E}_{\text{ex}} + \overline{N}_i \overline{E}_{\text{se}}.$$

Dividing each term by $\overline{N}_i$ leads to an expression for W. In general, W is determined experimentally, because the terms in this equation are quite difficult to calculate. However, they have been calculated for helium.[1] The mean energy of an ionized helium atom is only 62 % of the value of W:

$$\underbrace{W}_{41.8\text{ eV}} = \underbrace{\overline{E}_i}_{\substack{25.9\text{ eV} \\ 62\,\%}} + \underbrace{\left(\overline{N}_{\text{ex}}/\overline{N}_i\right)\overline{E}_{\text{ex}}}_{\substack{0.4\times 20.8 = 8.3\text{ eV} \\ 20\%}} + \underbrace{\overline{E}_{\text{se}}}_{\substack{7.6\text{ eV} \\ 18\,\%}}$$

Values of W are tabulated in ICRU Report 31 (1979). There are variations of a few percent depending on whether the charged particle is an electron or an α particle. Table 16.1 provides a few representative values. Though defined for a gas, W is also applied to semiconductors as the average energy per electron–hole pair produced. Values of W for semiconductors are much smaller than for a gas.

[1] See Platzman (1961); also Attix (1986, pp. 339–343).

16.2.3 Exposure

The *exposure* X is defined only for photons and measures the energy fluence of the photon beam. It is the amount of ionization (total charge of one sign) produced per unit mass of dry air when all of the electrons and positrons liberated in a small mass of air are completely stopped in air:

$$X = \frac{dq}{dm}. \tag{16.6}$$

The units are coulomb per kilogram. Since the average amount of energy required to produce an ion pair is well defined, exposure is closely related to collision kerma in air. The definition of X does not include ionization arising from the absorption of bremsstrahlung emitted by the electrons, so there is a slight difference at high energies.[2] The relationship is

$$X = (K_c)_{\text{air}} \left(\frac{e}{W_{\text{air}}}\right) = \Psi \left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{air}} \left(\frac{e}{W_{\text{air}}}\right). \tag{16.7}$$

If charged-particle equilibrium exists, the dose in air is related to the beam energy fluence by Eqs. 15.71 and 15.74:

$$D_{\text{air}} = \left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{air}} \Psi.$$

The dose for the same energy fluence in some other medium is

$$D_{\text{med}} = \left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{med}} \Psi = \frac{(\mu_{\text{en}}/\rho)_{\text{med}}}{(\mu_{\text{en}}/\rho)_{\text{air}}} D_{\text{air}}. \tag{16.8}$$

The *roentgen* (R) is an old unit of exposure equivalent to the production of 2.58×10^{-4} C kg^{-1} in dry air; this corresponds to a dose of 8.69×10^{-3} Gy. (The relationship is developed in Problem 7).

16.3 Detectors

Detectors are used for recording an image and also for measuring the amount of radiation to which a patient is exposed. This section describes the most common kinds of detectors.

[2] There is also a problem at high energies because the range of the electrons is large. If they are to come to rest within the chamber, the size of the chamber becomes comparable to the photon attenuation coefficient.

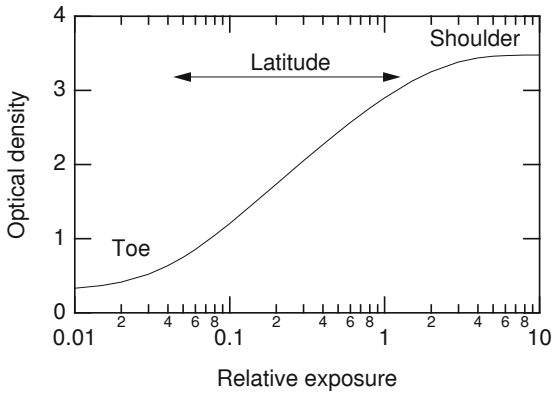


Fig. 16.6 Optical density vs. the logarithm of the relative exposure for a hypothetical x-ray film

16.3.1 Film and Screens

Film was the original x-ray detector used by Wilhelm Roentgen, the discoverer of x-rays. For years it was the most common detector for diagnostic radiology. In recent years it has been replaced by digital detectors. We describe it briefly for comparison with the newer techniques.

A typical x-ray film has a transparent base about 200 μm thick, coated on one or both sides with a sensitive emulsion containing a silver halide (usually silver bromide). We will not discuss the rather complicated sequence of steps by which the absorption of photons or energy loss by charged particles leads to a latent or developable image in the film. When the film is developed, the emulsion grains that have absorbed energy are reduced to black specks of metallic silver. The film is then fixed, a process in which the silver halide that was not reduced is removed from the emulsion. The result is a film that absorbs visible light where it was struck by ionizing radiation.

The fraction of incident light passing through the film after development is called the *transmittance*, T. The *optical density* or *density* is defined to be

$$\mathrm{OD} = \log_{10}(1/T). \tag{16.9}$$

A film that transmits 1 % of the incident viewing light has an optical density of 2.

The response of a film is described by plotting the optical density against the log of the exposure in air immediately in front of the film (or equivalently, the absorbed dose in the film emulsion). Since the optical density is the logarithm of the transmittance, this is a log–log plot of the reciprocal of the fraction of the visible light transmitted when viewing the processed film vs. the x-ray exposure before processing. A typical plot of film response is shown in Fig. 16.6. If the curve is linear, the transmittance is proportional to the exposure raised to some power:

$$T \propto X^{-\gamma}.$$

At very small exposures (the *toe*) the transmission is that of the base and "clear" emulsion. At very high exposures (the *shoulder*) all of the silver halide has been reduced to metallic silver, and the film has its maximum optical density. In between is a region that is almost linear (on a log–log scale). The ratio of maximum to minimum usable exposure is called the *latitude* of the film. The largest value of the exponent occurs at the inflection point and is called the *gamma* or *contrast* of the film. Both the exponent and the position of the curve along the log exposure axis depend on the development time, the temperature of the developing solution, and the energy of the x-ray beam. The *film speed* is the reciprocal of the exposure required for an optical density that is 1 greater than the base density.

A typical film has an emulsion containing AgBr. It requires a dose of 1.74×10^{-4} Gy (J kg^{-1}) in air just in front of the film to produce an optical density of 1. This might be where the body is not blocking the beam. The smaller dose to the film where there has been significant attenuation in the body gives a lighter region, as in the heart and bone shadows of Fig. 16.17.

The dose to the part of the body just in front of the film (the *exit dose* to the patient) can be written in several ways. For simplicity we assume monoenergetic photons. In terms of the energy fluence of the photon beam, the exit dose is

$$D_{\mathrm{body}} = \left(\frac{\mu_{\mathrm{en}}}{\rho}\right)_{\mathrm{body}} \Psi. \tag{16.10a}$$

In terms of the dose in air just in front of the film it is

$$D_{\mathrm{body}} = \frac{(\mu_{\mathrm{en}}/\rho)_{\mathrm{body}}}{(\mu_{\mathrm{en}}/\rho)_{\mathrm{air}}} D_{\mathrm{air}}, \tag{16.10b}$$

and in terms of the dose in the film it is

$$D_{\mathrm{body}} = \frac{(\mu_{\mathrm{en}}/\rho)_{\mathrm{body}}}{(\mu_{\mathrm{en}}/\rho)_{\mathrm{film}}} D_{\mathrm{film}}. \tag{16.10c}$$

For 50-keV photons we find from the tables at physics.nist.gov/PhysRefData/XrayMassCoef/ that $(\mu_{\mathrm{en}}/\rho)_{\mathrm{muscle}}/(\mu_{\mathrm{en}}/\rho)_{\mathrm{air}} = 0.004\,349/0.004\,098 = 1.061$. Therefore the exit dose would be $(1.74 \times 10^{-4})(1.061) = 1.85 \times 10^{-4}$ Gy. Because of attenuation, the entrance dose can be much larger.

The dose can be reduced by a factor of 50 or more if the film is sandwiched between two fluorescent *intensifying screens*. The x-ray photons have a low probability of interacting in the film. The screens have a greater probability of absorbing the x-ray photons and converting them to visible light, to which the film is more sensitive. For 50-keV

photons on typical emulsion, the value of μ_{en}/ρ is about 0.261 $m^2\ kg^{-1}$. A typical value of $\rho\Delta x$ for the film might be 0.02 $kg\ m^{-2}$. Therefore $\mu_{en}\Delta x = 0.0052$. The fraction of incident energy absorbed in the emulsion is $1 - e^{-0.0052} = 0.0052$.

A typical screen might consist of particles of terbium-doped gadolinium oxysulfide (Gd_2O_2S:Tb) suspended in a carrier about 150 μm thick (0.5–1.5 $kg\ m^{-2}$). This layer is backed by a thin reflective layer. Two such screens (one on each side of the film) with a total thickness of 1.2 $kg\ m^{-2}$ absorb 28 % of the 50-keV photons that pass through them (see Problem 11). The overall effect is to produce the same optical density when the energy fluence in the x-ray beam is reduced by a factor of 54—the ratio of the incident radiation absorbed in the screen and in the film in each case.[3] Typically, a sheet of film is placed in a light-tight *cassette* whose front and back walls are made of screen material.

Figure 16.6 shows a plot of optical density vs. the log of the exposure. The slope at any point on the curve is[4]

$$\gamma = \frac{d\log_{10}(1/T)}{d\log_{10}X} = \frac{d\ln(1/T)}{d\ln X} = -\frac{dT/T}{dX/X}$$
$$= -\frac{X}{T}\frac{dT}{dX} = -\frac{1}{G}g, \qquad (16.11)$$

where $G = T/X$ is the *large-signal transfer factor* and $g = dT/dX$ is the *incremental-signal transfer factor*. This will be used in our discussion of detecting signals in noise in the next section.

16.3.2 Scintillation Detectors

When x-ray photons interact with matter, some of their energy is transferred to electrons. These electrons interact in turn, and some of their energy can become ultraviolet or visible photons. A *scintillator* is a substance that produces these photons with high efficiency, yet is transparent to them. The photons are then transferred by an optical fiber or a lens system to a light detector such as a photomultiplier tube or a solid-state photodetector. Each x-ray photon produces an electrical current pulse at the detector output, called a *count*. When the number of counts is recorded vs. the pulse height

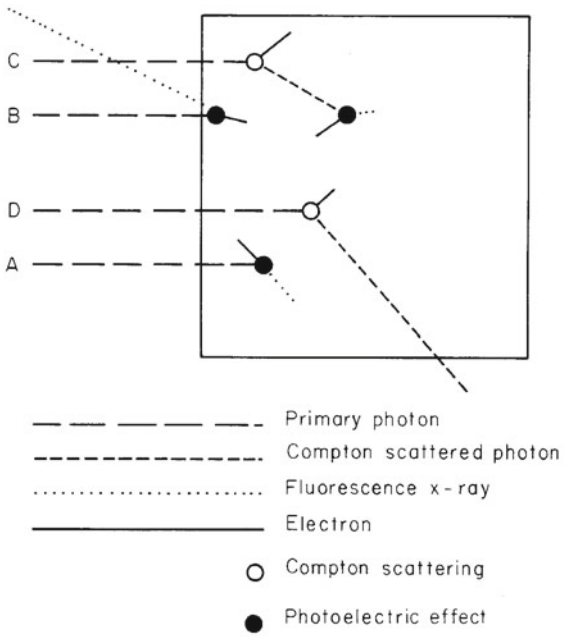


Fig. 16.7 Mechanisms by which some of the energy of a primary photon can escape from a detector. Photons *A* and *B* undergo photoelectric absorption. All of the energy from *A* is absorbed in the detector, while the fluorescence x-ray from *B* escapes. Photons *C* and *D* are Compton scattered. The scattered photon from *C* undergoes subsequent absorption, while that from *D* escapes

(total charge in the pulse, which is proportional to the energy deposited in the scintillator), the result is a *pulse-height spectrum*. For monoenergetic photons, the ideal pulse height spectrum would consist of a single peak: all pulses would have the same height. This is not realized in practice for two reasons: statistical variations in the scintillation process cause the line to be broadened, and the entire energy of the incident photon is not converted into electrons.

An atom that has been excited by photoelectric absorption can decay by the emission of a fluorescence photon. If this photon is subsequently absorbed in the scintillator, all of the original photon energy is converted to electron energy so rapidly that the visible light is all part of one pulse. The pulse height then corresponds to the full energy of the original photon. However, if the initial photoelectric absorption takes place close to the edge of the detector, the fluorescence photon can escape, and the pulse has a lower height than those in the primary peak. This can be seen in Fig. 16.7. Photons *A* and *B* interact by photoelectric effect. All the energy for photon *A* is deposited in the scintillator, while the *K* fluorescence photon from *B* escapes. The effect on a pulse height spectrum is shown in Fig. 16.8 for a scintillator of sodium iodide.

In Compton scattering, the energy of the recoil electron is transferred to the scintillator (unless the electron escapes).

[3] The fluorescent radiation has a wavelength of about 545 nm (green), and each absorbed high-energy photon has sufficient energy to produce about 14,000 fluorescence photons. However, the efficiency of production is only about 5 % so 700 photons are produced. Some of these escape or are absorbed. Each x-ray photon produces about 150 photons of visible and ultraviolet light that strike the emulsion—more than enough to blacken the film in the region where the x-ray photon was absorbed by the screen.

[4] An argument based on Eq. 2.14 can be used to show that $\log_{10} x = (1/2.303)\ln x = 0.43\ln x$.

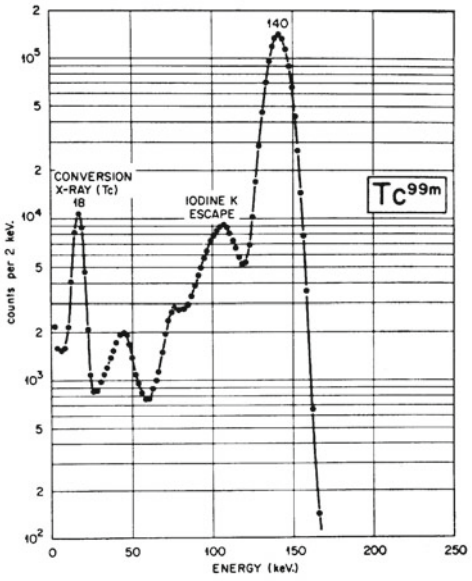


Fig. 16.8 Spectrum of pulse heights for 140-keV photons from isotope ^{99m}Tc incident on a sodium iodide scintillator. The 140-keV total energy peak is prominent, as is the peak at 110 keV corresponding to the escape of the K fluorescence x-ray from iodine. The Compton scatter continuum runs from 49 keV down to zero energy. The peak at 18 keV is from additional radiation from ^{99m}Tc (see Chap. 17; Reproduced from Wagner 1968, p. 162. Copyright 1968 by W. B. Saunders. Used by permission of Elsevier)

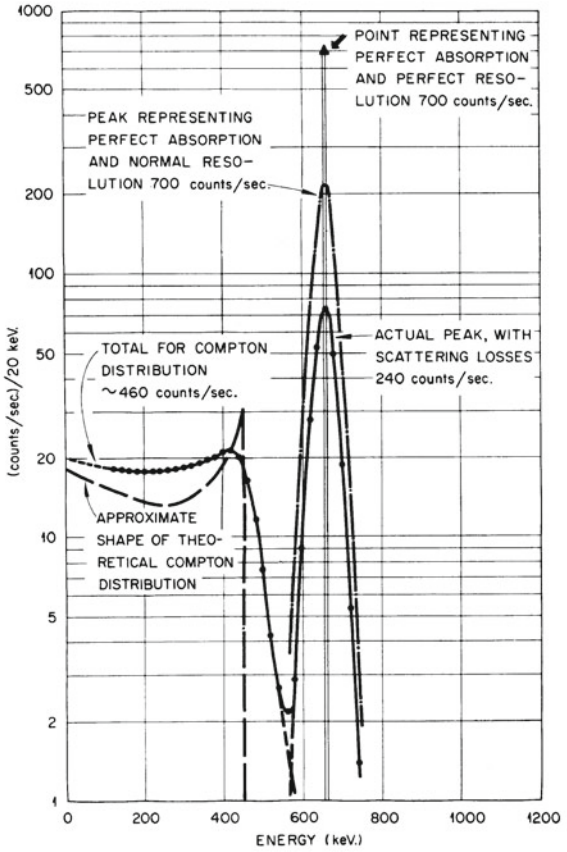


Fig. 16.9 The response of a sodium iodide detector to 662-keV photons from ^{137}Cs. Theoretical responses are shown for a detector that absorbs the energy of all photons and has perfect resolution, for a detector with perfect absorption and finite resolution, and for a detector in which Compton-scattered photons can escape. Experimental data are for a $1\frac{1}{2}$-in. by 1-in. NaI crystal. (Redrawn from Harris et al. 1969; Reproduced from Wagner 1968, p. 153. Copyright 1968 W. B. Saunders. Used by permission of Elsevier)

The scattered photon may escape from the detector, as in D of Fig. 16.7. (If it is subsequently absorbed, as in mechanism C of Fig. 16.7, the pulse height will have the peak value). The energy of the recoil electron is given by Eq. 15.15. The maximum electron energy occurs when the photon is scattered through $\theta = 180\,^\circ$. Then

$$T_{\max} = \frac{2h\nu_0 x}{1+2x} = \frac{(h\nu_0)^2}{h\nu_0 + m_e c^2/2}.$$

If the photon energy is in keV, this is

$$T_{\max} = \frac{(h\nu_0)^2}{h\nu_0 + 256}. \tag{16.12}$$

A pulse-height spectrum for "pure" Compton scattering of 662-keV photons (as emitted by ^{137}Cs) is shown in Fig. 16.9. The peak of the Compton continuum is at $T_{\max} = (662)^2/(662+256) = 477$ keV. The cases of perfect resolution with complete absorption and real resolution with complete absorption are shown, along with the theoretical Compton continuum with perfect resolution, and a real spectrum.

When the energy of the primary photons is so large that pair production is important, an additional escape mechanism must be considered. We know (Eq. 15.22) that

$$h\nu_0 = T_+ + T_- + (m_e c^2)_{e^-} + (m_e c^2)_{e^+}.$$

The positron will eventually combine with another electron to produce two annihilation radiation photons:

$$(m_e c^2)_{e^+} + (m_e c^2)_{\text{another } e^-} = 2E_\gamma.$$

The energy of each annihilation photon is 511 keV. The initial photon energy is finally distributed as

$$h\nu_0 = T_+ + T_- + \gamma(511) + \gamma(511).$$

If all this energy is absorbed in the detector, the pulse height corresponds to the full energy of the incident photon. One

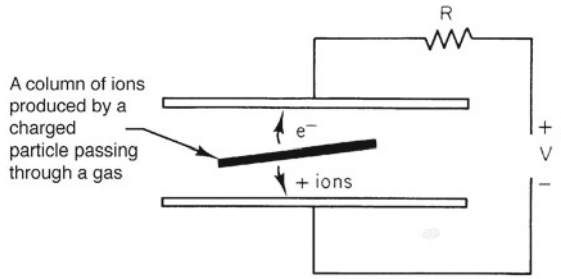


Fig. 16.10 Schematic of an ionization chamber or proportional counter. The ions discharge the capacitor, which is recharged between counts through resistor R

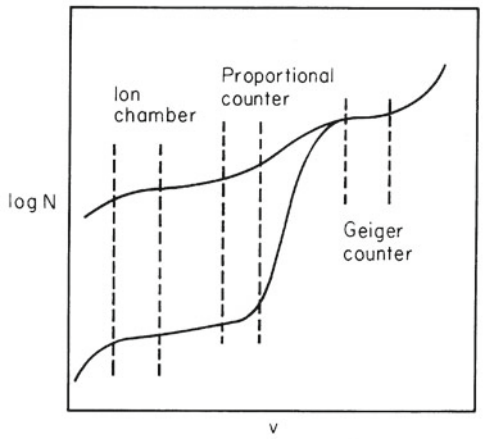


Fig. 16.11 The number of ions collected vs. collecting potential for two particles that deposit different amounts of energy in a gas detector. The voltage regions are indicated where the device operates as an ionization chamber, a proportional counter, and a Geiger counter

or both of the annihilation photons can escape, giving the one-photon escape peak and the two-photon escape peak.

Scintillation detectors vary greatly in size. Large ones may be tens of centimeters in diameter; others can be less than a millimeter. A large number of materials are used (van Eijk 2002).

16.3.3 Gas Detectors

Ionization in gas is the basis for three kinds of x-ray detectors: ionization chambers, proportional counters, and Geiger counters. A photon passing through a gas can produce photoelectric, Compton, or pair-production electrons. These then lose energy by electron collisions. Ion pairs are produced in the gas, the average number being proportional to the amount of energy lost in the gas. The average amount of energy required to produce an ion pair is W, as we saw in Sect. 16.2. Imagine that the ions are produced between the plates of a charged capacitor as shown in Fig. 16.10. The electrons are attracted to the positive plate and the positive ions travel to the negative plate. If all the electrons and ions are captured, the total charge collected on each plate has magnitude $q = Ne$, where e is the charge on the electron or ion and N is the number of ion pairs formed. If the capacitance is C, the change in voltage is $\delta v = q/C = Ne/C$. Such a device is called an *ionization chamber*. The cumulative discharge of the capacitor is measured in some pocket dosimeters; in other cases, the capacitor is slowly recharged through a large resistance R so that each photon detected generates a voltage pulse of height δv.

A certain minimum voltage between the two plates is necessary to ensure that all the ions produced are collected, corresponding to the ion chamber region of Fig. 16.11. The ionization chamber is the "workhorse" detector for accurately measuring radiation dose.

If the potential on the plates is raised further, the number of ions collected increases. Between collisions the electrons and ions are accelerated by the electric field, and they acquire enough kinetic energy to produce further ionizations when they collide with molecules of the gas, a process called *gas multiplication*. At moderate potentials, the multiplication factor is independent of the initial ionization, so the number of ions collected is larger than that in an ionization chamber but still is proportional to the initial number of ions. In this region of operation the device is called a *proportional counter*. Parallel-plate geometry is not used in a proportional counter. One electrode is a wire, and the other is a concentric cylinder.

At still higher values of the applied voltage, pulse size is independent of the initial number of ion pairs. In this mode of operation, the device is called a *Geiger counter*.

Any gas detector used to detect high-energy photons suffers from the fact that the gas is not very dense. At low energies the photoelectric cross-section is high and most photons interact. At higher energies, many photons pass through the gas and detector walls without interacting. A thin sheet of absorber in front of the gas detector can actually increase the counting rate. An example is shown in Fig. 16.12. The detector had an aluminum wall of thickness 0.3 kg m^{-2}. Electrons of 125 keV or more pass completely through the detector wall. The maximum energy of Compton electrons from the 1.1-MeV photons is 890 keV. Compton electrons produced in a thin layer of lead can pass through the aluminum and ionize the gas in the detector. Once the total thickness of lead and aluminum is sufficient to stop all the Compton electrons, the addition of more lead upstream does not increase the detector efficiency, and exponential attenuation is seen.

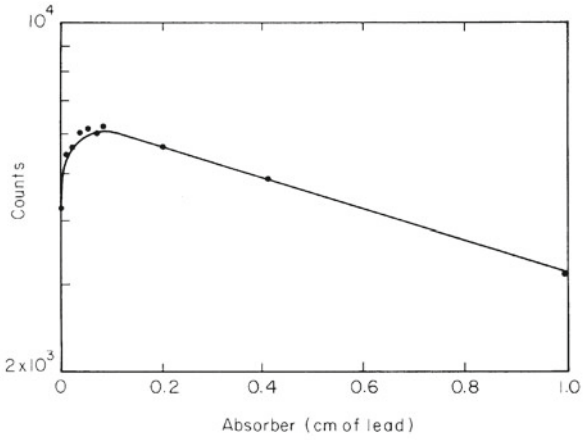


Fig. 16.12 Counting rate of a Geiger counter vs. the thickness of a lead absorber in front of the detector, showing the buildup of counting rate due to the conversion of photons to electrons in the lead by Compton scattering. These electrons pass through the thin wall of the counter and ionize the gas. The photons were from ^{60}Co and had an energy of 1.1 MeV

16.3.4 Semiconductor Detectors

A semiconductor detector is very much like an ionization chamber, except that a solid is used as the detecting medium. The "ion pair" is an electron that has received sufficient energy to be able to leave its atom and move freely within the semiconductor (but not enough energy to leave the semiconductor entirely) and the "hole" that the electron left behind. Electrons from neighboring atoms can fall into the hole, so the hole can move from atom to atom just like a positive charge. Details of the operation of semiconductor detectors can be found in Lutz (1999).

A semiconductor detector has two principal advantages over a gas ionization chamber. First, the amount of energy required to create an electron–hole pair is only about 3 eV, one tenth the value for a typical gas. This means that many more pairs are produced and the statistical accuracy is better. Second, the density of a solid is much greater than the density of a gas, so the probability that a photon interacts is larger. The cross-section for interaction increases with high Z, so detectors made of germanium ($Z = 32$) are better for photon detection than those made of silicon ($Z = 14$). Diode detectors are used for real-time dose measurement in patients receiving radiation therapy (AAPM Report 87 2005).

16.3.5 Thermoluminescent Dosimeters

Thermoluminescent phosphors consist of a small amount of dielectric material (0.1 g or less) that has been doped with impurities or has missing atoms in the crystal lattice to form metastable energy levels or traps. These impurities or defects are far from one another and are isolated in the lattice, so that electrons cannot move freely from one trap to another. When the phosphor is irradiated with ionizing radiation, some of the electrons are trapped in these metastable states. There are levels associated with the material at an energy E above the trap energy (the conduction band) which allow electrons to move throughout the phosphor. The probability that an electron escapes from the trap is proportional to a Boltzmann factor, $\exp(-E/k_BT)$. If E is large enough, the lifetime in the trapped state can be quite long—up to hundreds of years. Heating allows the electrons to escape to the higher levels, where they then fall back to the normal state with the emission of visible photons. Ordinary table salt (NaCl) exhibits this behavior. If it is irradiated and then sprinkled on an electric hot plate in a darkened room, one can see the flashes of light. The light emitted on heating is called *thermoluminescence*. In a *thermoluminescent dosimeter* (TLD), the light emitted is measured with a photomultiplier tube as the temperature is gradually increased. The total amount of light released is proportional to the energy imparted to the phosphor by the ionizing radiation.

Thermoluminescent dosimeter can measure an integrated dose from 10^{-5} to 10^{3} Gy. Great care must be taken both in the preparation and reading of the phosphor. Thermoluminescent dosimeter chips are widely used to measure radiation doses because they are small and have the approximate atomic number and atomic weight of tissue. They are often made of LiF. Detailed descriptions are found in Chap. 14 of Attix (1986), and Shani (1991, 2001). (The two editions of Shani complement one another.)

16.3.6 Chemical Dosimetry

When radiation interacts with water, *free radicals* are produced. A free radical, such as H or OH, is electrically neutral but has an unpaired electron. Free radicals promote other chemical reactions. Typically, a dilute indicator of some sort is added to the water. A common dosimeter is the Fricke ferrous sulfate dosimeter. A 1 mM $FeSO_4$ solution is irradiated. The radiation changes the iron from the ferrous (Fe^{2+}) to the ferric (Fe^{3+}) state with a G of about 1.6×10^{-6} mol J^{-1}. The concentration of ferric ion can be measured by absorption spectroscopy. Details are found in Chap. 14 of Attix (1986) and in Shani (1991, 2001). Magnetic resonance imaging (Chap. 18) is also used to measure the amount of ferric ion in the Fricke dosimeter since the relaxation times depend on the ion concentration. This has led to the gel dosimeter which allows a three-dimensional measurement of the dose distribution—useful for planning radiation treatments (Shani 2001, Chap. 9).

Another form of chemical dosimeter is *radiochromic film*. It consists of a thin layer of radiosensitive dye bonded to a

mylar base. The dye darkens with radiation. Radiochromic films are sensitive for doses of 1–500 Gy, making them useful for measuring doses in radiation therapy (Shani 2001, Chap. 5).

16.3.7 Digital Detectors

Digital x-ray detectors have replaced film in clinical radiography (Doi 2006; Armato and van Ginneken 2008; Cowen et al. 2008b; Körner et al. 2007; Uffmann and Schaefer-Prokop 2009). A digitally recorded image generally has up to 400 times the dynamic range (latitude) of film. A factor-of-2 error in film exposure,[5] which renders a conventional radiographic image almost useless, is easily tolerated by digital recording.[6] A digitally stored image allows easier retrieval, transmission, manipulation with computer algorithms, and duplication.

A number of techniques are used. In *Computed Radiography* (CR), the image is formed on a plate of phosphor crystals such as barium fluorobromide. Absorption of x-ray photons leaves the BaFBr crystals in a metastable state, like a TLD phosphor. Scanning by a thin laser beam in a horizontal and vertical raster pattern like a television image causes visible light to be emitted by the trapped electrons. The dynamic range of a storage phosphor can be as high as 10^4, compared to about 10^2 for radiographic film (Rowlands 2002).

In *Direct Radiography* (DR) the thin-film transistor (TFT) array technology used in flat-panel computer screens is used to make large detector arrays. The TFT arrays provide the spatial readout. They are combined either with an amorphous selenium photoconductor that converts the x-ray energy to charge (direct conversion), or with a structured scintillator such as a large array of CsI crystals. Each CsI crystal may be as small as 6 μm diameter by 500 μm long.

16.4 The Diagnostic Radiograph

Figure 16.13 shows the typical elements for making a diagnostic x-ray. An image recorded on film or a detector array is called a *radiograph*. The x-ray tube ideally serves as a point source of photons. The photons are filtered and collimated to illuminate only the portion of the patient of interest. Typically, about 10 % pass through the patient and strike the detector in a chest radiograph. In the abdomen the fraction is about 1 %. There may be an optional grid, as discussed below. We discuss each element below, and then discuss the quality of the image.

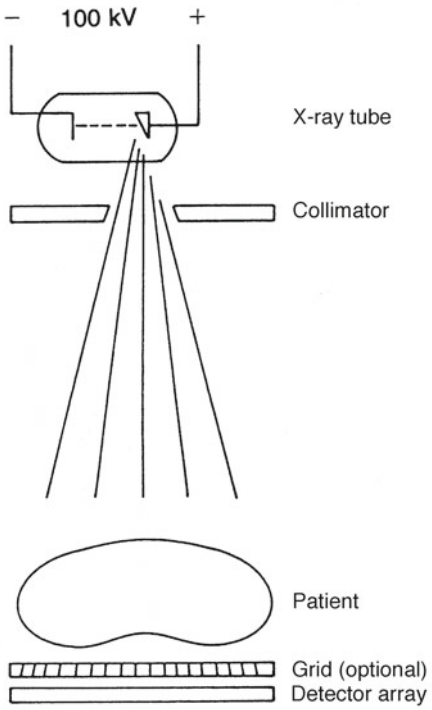


Fig. 16.13 Overall scheme for making a radiograph. Photons are produced when electrons strike the tungsten anode. The beam is collimated to prevent unnecessary dose to the patient. The patient is placed directly in front of the grid (if any), and the digital detector array, which can be film, a sandwich of film and intensifying screen, or a detector array

16.4.1 X-Ray Tube and Filter

Most routine radiography is done with photons in the range from 35 to 85 keV. (Mammography uses lower energy, and computed tomography is somewhat higher.) Figure 16.14 shows the loss of radiographic contrast as the energy of the incident photons increases and Compton scattering becomes more important.

The photons are typically produced by an x-ray tube running with a voltage between cathode and anode of about 100 kilovolt peak[7] (100 kVp). The anode is usually made

[5] Even though the film may have a linear response over a larger range, doubling the exposure usually makes the film too dense to read.

[6] Although a digital detector has greater dynamic range, proper exposure is still important. Too low an exposure introduces noise; an excessive exposure increases he dose to the patient unnecessarily.

[7] The word *peak* is included because the voltage from power supplies in older machines had considerable "ripple" caused by the alternating voltage from the power lines. Even in modern machines, the voltage

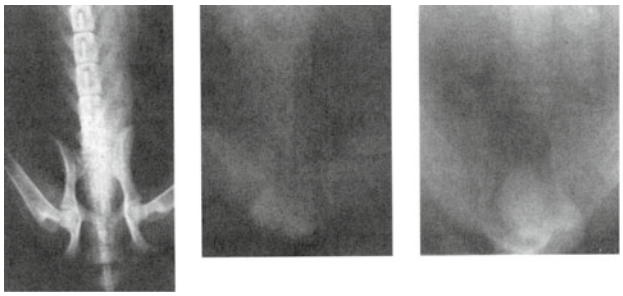


Fig. 16.14 Radiographs taken at 70 kVp, 250 kVp, and 1.25 MeV (^{60}Co), illustrating the loss of contrast for higher energy photons. (From Hendee and Ritenour (2002). Used by permission)

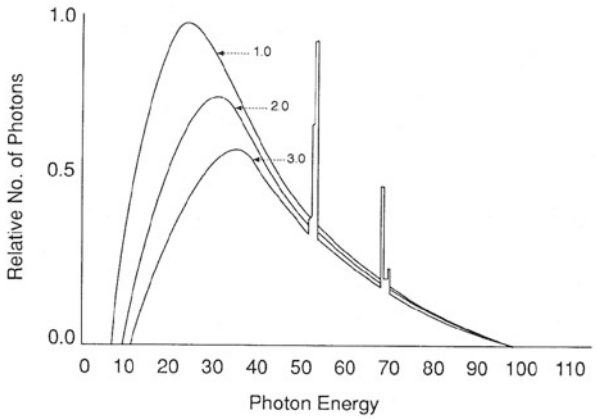


Fig. 16.15 The particle energy spectrum $d\Phi/dE$ from a tube operating at 100 kVp with 1, 2, and 3 mm of aluminum filtration. (From Hendee and Ritenour (2002). Used by permission)

of tungsten (which has a high radiation yield and withstands high temperatures) with a copper backing to conduct thermal energy away. The number of x-rays produced for a given voltage difference depends on the total number of electrons striking the anode, which is proportional to the product of the current and the duration of the exposure (mA s). The anode rotates to help keep it cool. Additional filtration removes low-energy photons that would not get through the body and would not contribute to the image. Figure 16.15 shows the effects of different thicknesses of aluminum on the particle fluence ($d\Phi/dE$) from a tube operating at 100 kVp. The average photon energy depends upon the filtration as well as the kVp, and is about 45 keV for 100 kVp and 2 mm of aluminum filtration (see Problem 18).

pulse applied to the tube may not have a purely rectangular waveform, and kVp may not uniquely determine the x-ray spectrum during the pulse. Modern kilovolt power supplies are described by Sobol (2002).

16.4.2 Collimation

The collimator is placed just after the x-ray tube. It has adjustable jaws, usually of lead, that limit the size of the beam striking the patient. Making the beam as small as possible reduces the total energy absorbed by the patient. It also reduces the amount of tissue producing Compton-scattered photons that strike the detector and reduce the image quality.

16.4.3 Attenuation in the Patient: Contrast Material

The purpose of a radiograph is to measure features of the internal anatomy of a patient through differences in the attenuation of rays passing through different parts of the body. The photon fluence falls with distance from the x-ray tube as $1/r^2$. It also falls because of attenuation along the path. (We ignore the fact that scattered photons may also strike the detector). We saw in Sect. 15.8 that the mass attenuation coefficient of a compound can be calculated as a weighted average of the elements in the compound:

$$\frac{\mu}{\rho} = \sum_i \left(\frac{\mu}{\rho}\right)_i w_i .$$

Table 16.2 lists various elements, their mass attenuation coefficients at 50 keV, and their composition in water, fat, muscle, and bone. Water and muscle are quite similar, fat has a somewhat smaller attenuation coefficient, and the attenuation of bone is significantly greater.

Figure 16.16 shows attenuation vs. ρx for the beam in Fig. 16.15 with 2 mm of Al filtration in water and in bone. Bone contains calcium, which has a relatively high atomic number, and the attenuation coefficient rises rapidly as the energy decreases. Also shown as dashed lines are the corresponding values of $\exp(-\mu_{\text{atten}} x)$ for the average photon energy in the incident beam, which is 50 keV. In each case the transmitted fraction initially falls more steeply than the dashed line because there is more attenuation of the low-energy photons. For thicker bone the slope of the curve is less than the dashed line because only the high-energy photons remain. This shift of the beam energy and curvature of the attenuation curves is called *beam hardening*.

These differences in attenuation make it easy to distinguish bone from soft tissue. It is also easy to distinguish lungs from other tissues because they contain air and have much lower density. Air-filled lung has a density of 180–320 kg m^{-3}, compared to about 1000 kg m^{-3} for water, muscle, or a solid tumor. Figure 16.17 shows a normal anterior–posterior (A–P) chest radiograph. You can see the exponential decay through layers of bone, the outline of the heart, the arch of the aorta, and the lacy network of

Table 16.2 Relative composition of various tissues and the attenuation coefficient for 50-keV photons

Element	μ_{atten}/ρ^b ($m^2\ kg^{-1}$)	Fractional mass composition[a]			
		Adipose tissue	Water	Skeletal muscle	Cortical bone, adult
H	0.0336	0.114	0.112	0.102	0.034
C	0.0187	0.598		0.143	0.155
N	0.0198	0.007		0.034	0.042
O	0.0213	0.278	0.888	0.710	0.435
Na	0.0280	0.001		0.001	0.001
Mg	0.0329				0.002
P	0.0492	0.001		0.002	0.103
S	0.0585	0.001		0.003	0.003
Cl	0.0648	0.001		0.001	
K	0.0868			0.004	
Ca	0.1020				0.225
μ_{atten}/ρ ($m^2\ kg^{-1}$)		0.0214	0.0227	0.0227	0.0424
ρ ($kg\ m^{-3}$)		970	1000	1050	1920
μ_{atten} (m^{-1})		20.8	22.7	23.8	81.5

[a] Fractional mass compositions are available at http://physics.nist.gov/PhysRefData/XrayMassCoef/tab2.html
[b] Values are from Hubbell and Seltzer (1996)

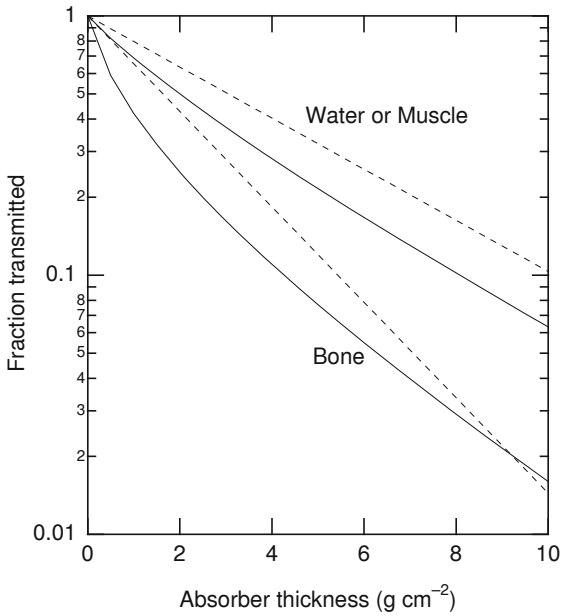


Fig. 16.16 Attenuation of photons in water or muscle and in bone for the spectrum of Fig. 16.15 (100 kVp, 2 mm aluminum filtration). The *dashed lines* are for the attenuation coefficients at 50 keV

blood vessels in the lungs. The patient in Fig. 16.19 has *pneumothorax*. Air has leaked into the pleural cavity and partially collapsed the lungs. You can see this collapse in the upper portion of each lung. Spontaneous pneumothorax can occur in any pulmonary disease that causes an alveolus (air sac) on the surface of the lung to rupture: most commonly emphysema, asthma, or tuberculosis. Pneumothorax can also be caused by perforating trauma to the chest wall. Spontaneous idiopathic (meaning cause unknown) pneumothorax occasionally occurs in relatively young people.

Abdominal structures are more difficult to visualize because except for gas in the intestine, everything has about the same density and atomic number. *Contrast agents* are introduced through the mouth, rectum, urethra, or bloodstream. One might think that the highest-Z materials would be best. However the energy of the K edge rises with increasing Z. If the K edge is above the energy of the x-rays in the beam, then only L absorption with a much lower cross-section takes place. The K edge for iodine is at 33 keV, while that for lead is at 88 keV. Between these two limits (and therefore in the range of x-ray energies usually used for diagnostic purposes), the mass attenuation coefficient of iodine is about twice that of lead. The two most popular contrast agents are barium ($Z = 56$, K edge at 37.4 keV) and iodine ($Z = 53$). Barium is swallowed or introduced into the colon. Iodine forms the basis for contrast agents used to study the cardiovascular system (angiography), gall bladder, brain, kidney, and urinary tract.

If the detector can discriminate photons of different energies, then one can measure photons on either side of an element's K edge, obtaining images that are easily distinguished from the image of background material (Schlomka et al. 2008).

Some pathologic conditions can be identified by the deposition of calcium salts. Such *dystrophic* (defective) calcification occurs in any form of tissue injury, particularly if there has been tissue necrosis (cell death). It is found in necrotizing tumors (particularly carcinomas), atherosclerotic blood vessels, areas of old abscess formation, tuberculous foci, and damaged heart valves, among others.

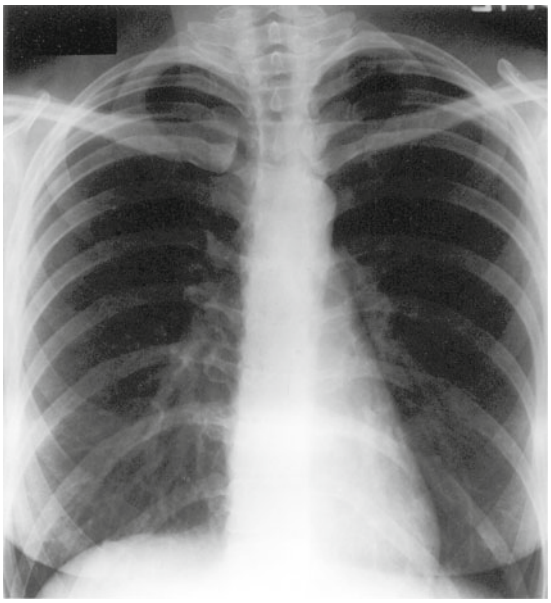


Fig. 16.17 Radiograph of a normal chest. Some of the features are identified in Fig. 16.18 and are described in the text. (Radiograph courtesy of D. Ketcham, M.D., Department of Diagnostic Radiology, University of Minnesota Medical School)

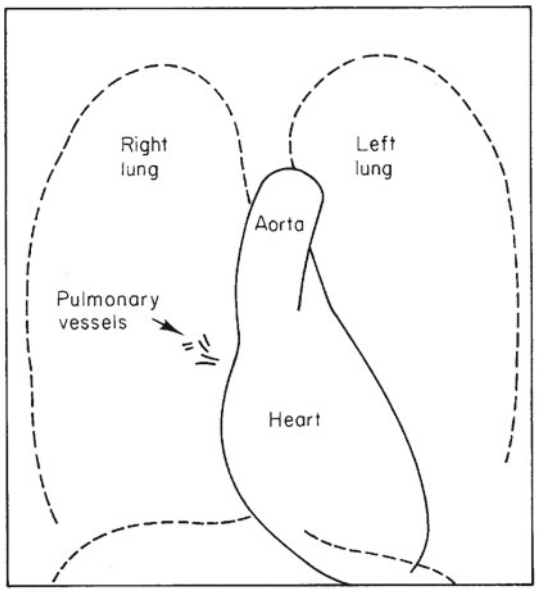


Fig. 16.18 Some of the features in the radiograph of a normal chest, Fig. 16.17

16.4.4 Antiscatter Grid

Since the radiograph assumes that photons either travel in a straight line from the point source in the x-ray tube to the detector or are absorbed, Compton-scattered photons that strike the detector reduce the contrast and contribute an overall background darkening. This effect can be reduced by placing an *antiscatter grid* (or radiographic grid, or "bucky" after its

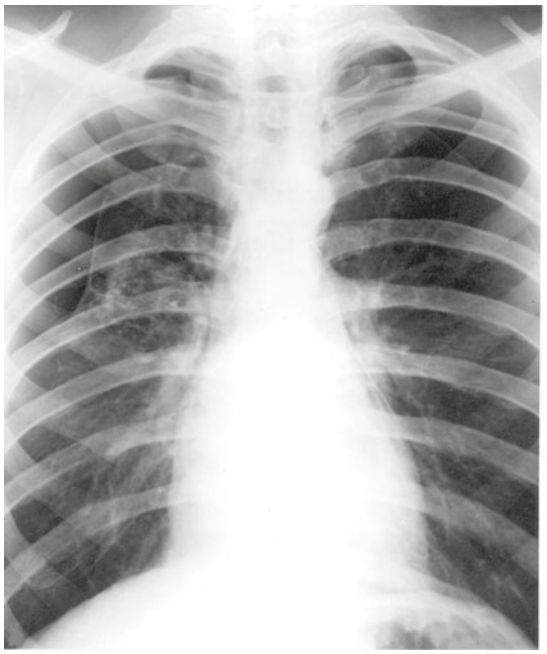


Fig. 16.19 Radiograph of a patient with pneumothorax. Air has escaped from the lungs and caused them to collapse partially. The features are indicated in Fig. 16.20. (Radiograph courtesy of D. Ketcham, M. D., Department of Diagnostic Radiology, University of Minnesota Medical School)

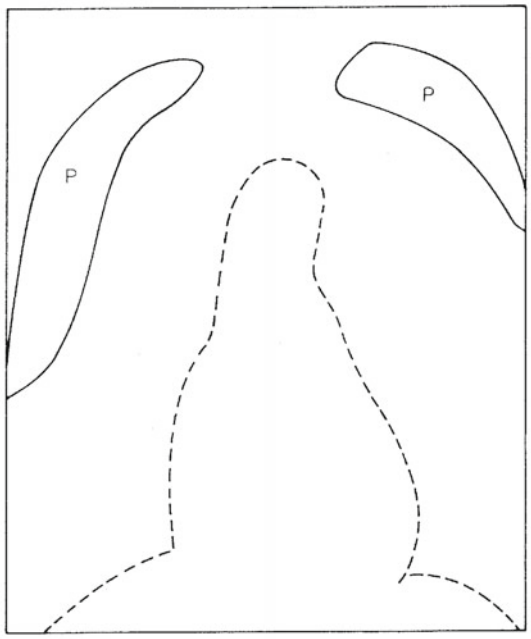


Fig. 16.20 Key to features in Fig. 16.19. The areas of pneumothorax are indicated by *P*. The one on the patient's left (the viewer's right) is difficult to see in the printed version; a radiographic film viewed by transmitted light has a much greater dynamic range

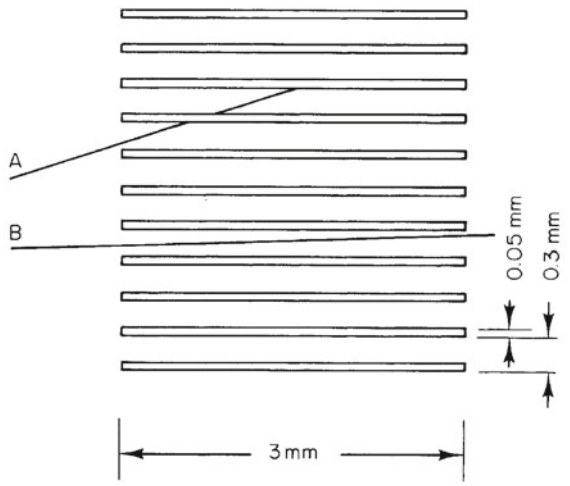


Fig. 16.21 Scale drawing of the elements of a typical grid. The thin lead strips absorb photons that have been scattered through more than a few degrees. As a result background fog due to scattering is reduced and the contrast is increased. Since the x-rays come from a point source, the elements of the grid are usually tilted toward the source and are not parallel over the entire detector surface

inventor, Gustav Bucky) just in front of the detector. Figure 16.21 shows how a grid works. The grid stops x-rays that are not traveling parallel to the sides of the grid strips. A typical grid might have 10–50 strips of lead per centimeter that are 3 mm high and 0.05 mm thick, embedded in plastic or aluminum. The strips can be either parallel or "focused," that is, slanted to aim at the point source on the anode of the x-ray tube. The grid can be either linear or crossed, with strips of lead running in both directions. A grid with a ratio of height to spacing of 12 improves the contrast by a factor of about 3.75, while increasing the exposure to the patient by a factor of about 4.25 to keep the detector dose about the same (Hendee and Ritenour 2002, p. 227). Sometimes the grid is moved during the exposure if the strips in the grid are thick enough to show up on the detector.

16.4.5 Detector

Detectors were described in Sect. 16.3. The film-screen detector is nearly obsolete. In computed radiography a photophosphor or photostimulable phosphor replaces the film-screen combination. The latent image on this phosphor is "read" by a scanning laser beam in a process called photostimulated luminescence. The resulting image is digitized (Rowlands 2002). Direct radiography uses a thin-film transistor detector array (p. 470) to directly produce a digitized image.

16.5 Image Quality

The quality of a radiographic image depends on three things: *resolution*, *contrast*, and *noise*. The resolution and contrast can be described by concepts introduced in Chap. 12 for a linear, shift-invariant system: the point-spread function and its Fourier transform, the optical transfer function, whose magnitude is the modulation transfer function.[8] The noise arises primarily from the fluctuations in the number of photons striking a given area of the detector—quantum noise—though granularity of the film or detector array is also important.

The transfer function for the entire system depends on many factors: the tube and spot size, filter, source–screen and source–patient distances, grid, detector, and scatter. If each of these subsystems operates in series, as in an audio system, one can successively convolve the point-spread functions or multiply together the (complex) optical transfer functions. It is also possible to have parallel[9] subsystems, each contributing to the final image, in which case the analysis is more complicated. An excellent review of the use of transfer-function analysis in radiographic imaging is the article by Metz and Doi (1979). The text by Macovski (1983) is at about the level of this book and presents many details of noise and convolution for radiographic, fluoroscopic, tomographic, nuclear medicine, and ultrasound images. The size of the spot where the electrons strike the anode of the x-ray tube is critical in determining the resolution of the final image, as discussed in detail by Wagner et al. (1974).

The *exposure contrast* is the change in exposure between two (usually adjacent) parts of the image divided by the average:

$$C_{\text{in}} = \frac{\Delta X}{X}. \tag{16.13}$$

This is similar to the modulation defined in Eq. 12.20. The *brightness contrast* is the analogous quantity for the light from the computer display or the light transmitted through the processed film:

$$C_{\text{out}} = \frac{\Delta T}{T}. \tag{16.14}$$

[8] The point-spread function of a detector is easily measured. A point source is created by passing the x-rays through a pinhole in a piece of lead placed directly on the detector. The resulting image is the point-spread function. We saw in Chap. 12 how this is related to the modulation transfer function. Standard techniques have been developed for measuring the modulation transfer function (MTF) (ICRU Report 41 1986; ICRU Report 54 1996).

[9] Examples of parallel subsystems are the two emulsion layers on double-coated film, and the effect of primary and scattered radiation on the formation of the image.

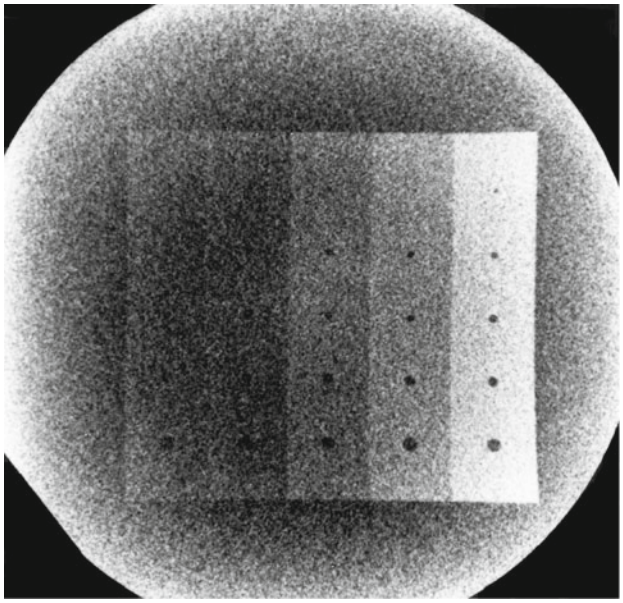


Fig. 16.22 An example of the relationship among exposure, image size, and detectability. A type 1100 aluminum phantom was imaged with digital fluorography. It was exposed to an 80-kVp x-ray beam with 4.5-mm Al filter. As the image becomes lighter, the thickness of the aluminum phantom increases in steps: 0.85, 1.3, 2.1, 3.2, and 5.2 mm. The holes are 1, 1.5, 2, 2.5, and 3 mm in diameter. As the attenuation in the aluminum increases, so does the signal, and the easier it is to detect the smaller holes. (Photograph courtesy of Richard Geise, Ph.D., Department of Radiology, University of Minnesota)

The exposure contrast and brightness contrast are proportional (Eq. 16.11):

$$C_{\text{out}} = \gamma C_{\text{in}}. \tag{16.15}$$

The *radiographic signal* is a small change in optical brightness in adjacent areas of the image. Changes in brightness below a certain value are not detectable by the viewer. This is apparent in Fig. 16.22, which shows signals with different contrasts and different sizes on a uniform background. The smaller the diameter of the signal region, the more difficult the signal is to detect. We will develop a simple model to explain why.

Suppose first that there is no signal, but that the detector is illuminated with a uniform beam of x-rays with a constant fluence. We make an exposure for a certain time and count the number of photons striking a sampling area of the detector, S. Though the average fluence is constant across the detector, the photons are randomly distributed. A somewhat different number of photons strike a nearby sampling area of the same size. This is a situation where the average number striking a sampling area of a given size is constant, the total number of photons is very large, and the probability that any one photon is absorbed in a given sampling area is small, so the situation is described by Poisson statistics (Appendix J). The mean number of photons striking a sampling area is ΦS and the standard deviation is $(\Phi S)^{1/2}$. Suppose that some fraction $f \leq 1$ of these photons actually interact with the detector. Then the mean number interacting is $f\Phi S$ and the standard deviation is[10] $(f\Phi S)^{1/2}$. Thus there are fluctuations in the brightness of the image across the uniformly exposed viewing region, just because of the Poisson statistics—quantum noise or shot noise—of the x-ray photons striking the detector.

The fluctuations in the number of photons striking area S can be related to fluctuations in the exposure of that area of the detector, and hence to the response of the detector. Since the exposure (measured in air just in front of the detector) is proportional to the photon fluence, $X = A\Phi$, $\overline{(X - \overline{X})^2} = A^2\overline{(\Phi - \overline{\Phi})^2}$ and $(\Delta X)_{\text{rms}} = A(\Delta\Phi)_{\text{rms}}$. We define the *noise exposure contrast* to be the standard deviation of the number of photons affecting the detector in area S divided by the average number affecting an area that size:[11]

$$C_{\text{noise in}} \equiv \frac{(f\Phi S)^{1/2}}{f\Phi S} = (f\Phi S)^{-1/2}. \tag{16.16}$$

The *noise brightness contrast* is then

$$C_{\text{noise out}} = \gamma(f\Phi S)^{-1/2}. \tag{16.17}$$

The fluctuations in the noise, measured by noise contrast, are inversely proportional to the square root of the area of the lesion to be detected.[12] This is seen in Fig. 16.22. The noise in the system is determined by measuring the charge collected in each pixel of the uniformly exposed detector. Variations with position can be described either in terms of its two-dimensional autocorrelation function or its Fourier transform, the *Wiener spectrum*. The radiographic noise consists of three components: *quantum mottle*, the statistical fluctuations in the number of photons absorbed in a small area (shot noise); *structure mottle* due to nonuniformities in the x-ray absorbing layer of the detector, and *graininess*, variation in the size and distribution of the transistors in the detector. Here we discuss only quantum mottle.

[10] This is very similar to the arguments about the fraction of photons absorbed by a visual pigment molecule in Eq. 14.68. Changes in the value of f in Fig. 14.43 shift the response curve along the axis.

[11] It is sometimes useful to write it as

$$C_{\text{noise in}} \equiv \frac{(f\Phi S)^{1/2}}{f\Phi S} = \frac{1}{f^{1/2}S^{1/2}}\frac{(\Delta\Phi)_{\text{rms}}}{\Phi} = \frac{1}{f^{1/2}S^{1/2}}\frac{A(\Delta X)_{\text{rms}}}{AX} = \frac{1}{f^{1/2}S^{1/2}}\frac{(\Delta X)_{\text{rms}}}{X}.$$

[12] An analogous phenomenon is seen when counting individual photons with a radiation detector at a fixed average rate. The number counted in a given time interval fluctuates, with the fractional fluctuation inversely proportional to the square root of the counting time.

Now introduce a signal, which is a small increase in the exposure or photon fluence: $\Delta X_{\text{signal}} = A\Delta\Phi_{\text{signal}}$. This gives a brightness contrast

$$C_{\text{signal out}} = \gamma \frac{\Delta X_{\text{signal}}}{X} = \gamma \frac{\Delta\Phi_{\text{signal}}}{\Phi}. \tag{16.18}$$

The ratio of the signal contrast to the noise contrast is called the *signal-to-noise ratio*:

$$\text{SNR} = \frac{C_{\text{signal out}}}{C_{\text{noise out}}} = \frac{\gamma\left(\Delta\Phi_{\text{signal}}/\Phi\right)}{\gamma\left(f\Phi S\right)^{-1/2}} = (fS)^{1/2}\frac{\Delta\Phi_{\text{signal}}}{\Phi^{1/2}}. \tag{16.19}$$

The signal will be detectable only if the signal brightness contrast exceeds the noise brightness contrast by a certain amount:[13]

$$\text{SNR} > k, \qquad (fS)^{1/2}\frac{\Delta\Phi_{\text{signal}}}{\Phi^{1/2}} > k. \tag{16.20}$$

The larger the value of the signal-to-noise ratio, the greater the probability of detecting the signal. Many experiments on detecting lesions in a noisy background have been done; they will not be discussed here.[14] Values of k that are used range from 2 to 5.

Let us apply the result in Eq. 16.20 to a simple model: a monoenergetic x-ray beam passing through a patient. The total thickness of the patient is L. The attenuation coefficient is μ. If an x-ray beam with fluence Φ_0 strikes the patient, the fluence of x-ray photons emerging is $\Phi_1 = \Phi_0 e^{-\mu L}$. Imagine a nearby region where for a distance x the attenuation coefficient is $\mu - \Delta\mu$. The x-ray fluence emerging along a line passing through this region is

$$\begin{aligned}\Phi_2 &= \Phi_0 e^{-\mu(L-x)-(\mu-\Delta\mu)x} = \Phi_0 e^{-\mu L} e^{\Delta\mu x} \\ &= \Phi_1 e^{\Delta\mu x} \\ &\approx \Phi_1(1 + x\,\Delta\mu).\end{aligned} \tag{16.21}$$

The exposure contrast is therefore $C_{\text{in}} = (\Delta\Phi)_{\text{signal}}/\Phi_1 \approx x\,\Delta\mu$. We combine this with Eq. 16.20 to obtain

$$(fS\Phi_1)^{1/2}(x\,\Delta\mu) > k, \tag{16.22}$$

where Φ_1 is the fluence leaving the patient or striking the detector. (These are the same if variations in $1/r^2$ can be neglected, where r is the distance from the tube to the patient or the tube to the detector). The signal-to-noise ratio increases as the square root of the photon fluence or exposure, the square root of the area to be detected, and the square root of f, the fraction of the photons striking the detector that are actually detected.

The fraction f in this Poisson model is equal to the *detective quantum efficiency* (DQE). It is easily visualized as the fraction of the photons striking the detector that actually affect it. The number of *noise equivalent quanta* (NEQ) in our model is $f\Phi_1$.[15]

We can apply Eq. 16.22 to determine the number of photons that must be transmitted through the patient for a given image size and given signal-to-noise ratio. We assume that $f = 1$. The required photon fluence emerging from the patient is (dropping the subscript on Φ_1)

$$\Phi S > \left(\frac{k}{x\,\Delta\mu}\right)^2. \tag{16.23}$$

If the lesion thickness is $x = 1$ cm and $\Delta\mu = 0.01\mu_{\text{water}} = (0.01)(22.7)\ \text{m}^{-1}$, then $x\,\Delta\mu = 0.00227$. For $k = 4$, the number of photons in the image area must be greater than 3×10^6. The exit dose to the patient is (assuming monoenergetic photons)

$$\begin{aligned}D_{\text{body}} &= \Psi\left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{body}} = (h\nu)\Phi\left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{body}} \\ &= \frac{(h\nu)(3\times 10^6)}{S}\left(\frac{\mu_{\text{en}}}{\rho}\right)_{\text{body}}.\end{aligned} \tag{16.24}$$

The dose increases as the area to be detected decreases. In order to detect an image 1 mm square using 50-keV photons, the exit dose in water would have to be at least 9.8×10^{-5} Gy.

16.6 Angiography and Digital Subtraction Angiography

One important problem in diagnostic radiology is to image portions of the vascular tree. *Angiography* can confirm the existence of and locate narrowing (stenosis), weakening and bulging of the vessel wall (aneurysm), congenital malformations of vessels, and the like. This is done by injecting

[13] There are statistical fluctuations in the signal as well as the noise. The variance of the difference between signal and noise will be the sum of the variances in the signal and in the noise. This has the effect of increasing the noise by a factor of $\sqrt{2}$, which can be absorbed in the value of k that is chosen. See Problem 23.

[14] The ability to detect the signal accurately is greater when the observer knows the nature of the signal and is only asked whether it is or is not present. That is, the ability of an observer to detect a signal is less in the more realistic situation where the observer does not know what the signal is or where it might be in the radiograph.

[15] This simple equality exists only because we are using a model with Poisson statistics. The DQE is defined more generally as the square of the signal-to-noise ratio of the detector output divided by the square of the signal-to-noise ratio of the detector input. The more general definitions of DQE and NEQ are discussed in Wagner (1983) and Wagner (1977).

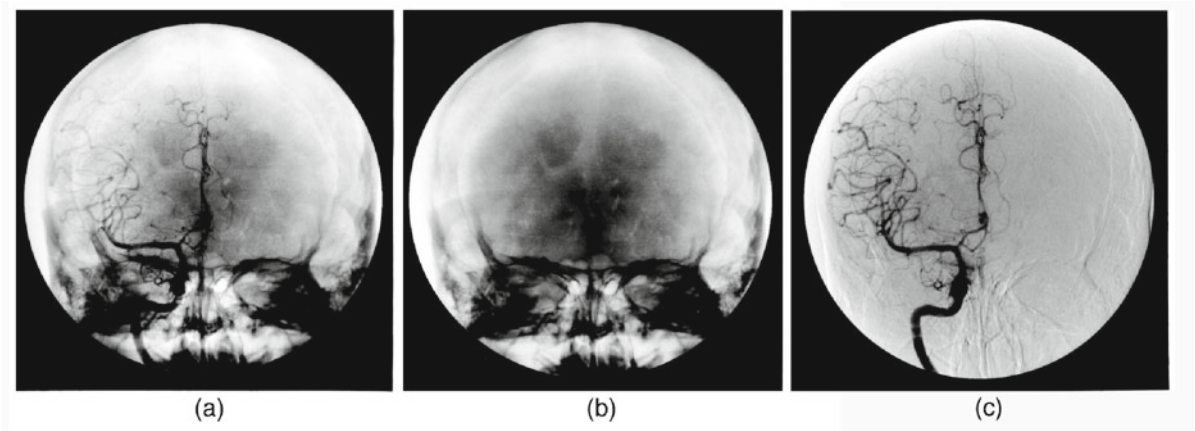


Fig. 16.23 Digital subtraction angiography. **a** Brain image with contrast material. **b** Image without contrast material. **c** The difference image. Anterior view of the right internal carotid artery. (Photograph courtesy of Richard Geise, Ph.D., Department of Radiology, University of Minnesota)

a contrast material containing iodine into an artery. If the images are recorded digitally, it is possible to subtract one without the contrast medium from one with contrast and see the vessels more clearly (Fig. 16.23).

In a typical angiographic study, 30–50 ml of contrast material is injected into an artery. For a vessel with a diameter of 8 mm, ρx of the contrast material is about 4 mg cm^{-2}.

16.7 Mammography

Mammography poses particular challenges for medical physicists. The resolution needed is extremely high (about 15 line pairs (lp) mm^{-1} compared to 5 lp mm^{-1} for a chest radiograph).[16] The radiologist may use a magnifying glass to inspect the image. The contrast in a breast image is inherently low. Fat and glandular tissue must be distinguished by the slight differences in their attenuation coefficients (see Problem 25). The dose must be made as small as possible.

These challenges have been met. Spatial frequencies of 14–16 lp mm^{-1} are routinely obtained. Noise limits the minimum size of a detectable object to $>$ 0.3 mm for microcalcifications or a few millimeters for soft tissue. Digital mammography is providing even higher resolution (Pisano and Yaffe 2005). The typical mammographic dose per view has been reduced from about 50 mGy in the 1960s and 4.1 mGy in the 1970s to 0.4 mGy in 2008.[17] One technique that has led to these improvements is the molybdenum target x-ray tube, operating at 25–28 kVp. Figure 16.24 shows the photon fluence from such a tube, with the 17-keV K_α and

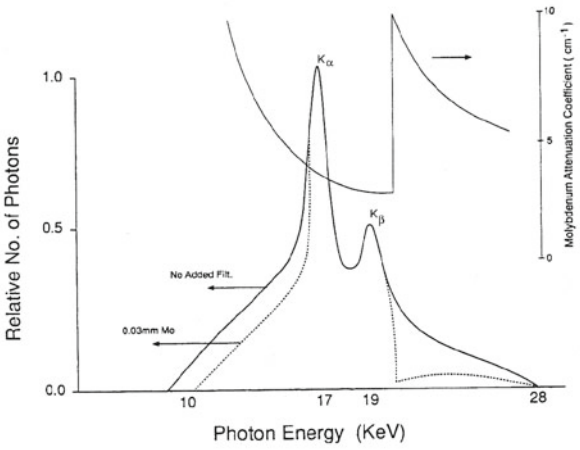


Fig. 16.24 The x-ray spectrum from a molybdenum anode tube used for mammography, with and without filtration by a molybdenum foil. (From Hendee and Ritenour 2002. Used by permission)

19-keV K_β lines being quite prominent. Filtration of the beam with the same material, molybdenum, further sharpens the spectrum. The K edge of molybdenum occurs at an energy just above the K_β line, removing photons for energies over about 20 keV. The dashed lines show the spectrum when a molybdenum filter is used. These photons interact primarily by the photoelectric effect, so attenuation depends strongly on atomic number. There are few Compton-scattered photons to degrade the image.

16.8 Computed Tomography

Radiographs provide only an integrated value of the attenuation coefficient. That is, if $N_0(y, z)$ monoenergetic x-ray photons traverse the body along a line in the x direction after

[16] Line pairs (abbreviated lp) are analogous to the period of a square wave.

[17] See NCRP Report 100 (1989) for early data; Mettler et al. (2008) for 2008 data.

entering the body at coordinates (y, z), the number emerging without interaction is $N(y, z) = N_0(y, z)e^{-\alpha(y,z)}$, where

$$\alpha(y, z) = \int \mu(x, y, z)\, dx.$$

The radiograph measures $N(y, z)$ or $\alpha(y, z)$. The desired information is $\mu(x, y, z)$. The radiographic image is often difficult to interpret because of this integration along x. For example, it may be difficult to visualize the kidneys because of the overlying intestines.

Several types of computed tomography (*tomos* means slice) have been developed in the last few decades. They include *transmission computed tomography* (CT), *single-photon emission computed tomography* (SPECT), and *positron emission tomography* (PET). They all involve reconstructing, for fixed z, a map of some function $f(x, y)$ from a set of projections $F(\theta, x)$, as described in Sect. 12.4 and 12.5. For CT the function f is the attenuation coefficient $\mu(x, y)$. For SPECT and PET it is the concentration of a radioactive tracer within the body, as will be described in Chap. 17.

The history of the development of computed tomography is quite interesting (Kalender 2011). The Nobel Prize in Physiology or Medicine was shared in 1979 by a physicist, Allan Cormack, and an engineer, Godfrey Hounsfield. Cormack had developed a theory for reconstruction and done experiments with a cylindrically symmetric object that were described in two papers in the *Journal of Applied Physics* in 1963 and 1964. Hounsfield, working independently, built the first clinical machine, which was installed in 1971. It was described in 1973 in the *British Journal of Radiology*. The Nobel Prize acceptance speeches (Cormack 1980; Hounsfield 1980) are interesting to read. A neurologist, William Oldendorf, had been working independently on the problem but did not share in the Nobel Prize (See DiChiro and Brooks 1979; and Broad 1980).

Early machines had an x-ray tube and detector that moved in precise alignment on opposite sides of the patient to make each pass. The size of these machines allowed only heads to be scanned. After one pass, the gantry containing the tube and detector was rotated 1° and the next pass was taken. After data for 180 passes were recorded, the image was reconstructed. A complete scan took about 4 min.

Figure 16.25 shows the evolution of the detector and source configurations. The third generation configuration is the most popular. All of the electrical connections are made through slip rings. This allows continuous rotation of the gantry and scanning in a spiral as the patient moves through the machine. Interpolation in the direction of the axis of rotation (the z axis) is used to perform the reconstruction for a particular value of z. This is called *spiral CT* or *helical CT*. Kalender (2011) discusses the physical performance of CT machines, particularly the various forms of spiral machines.

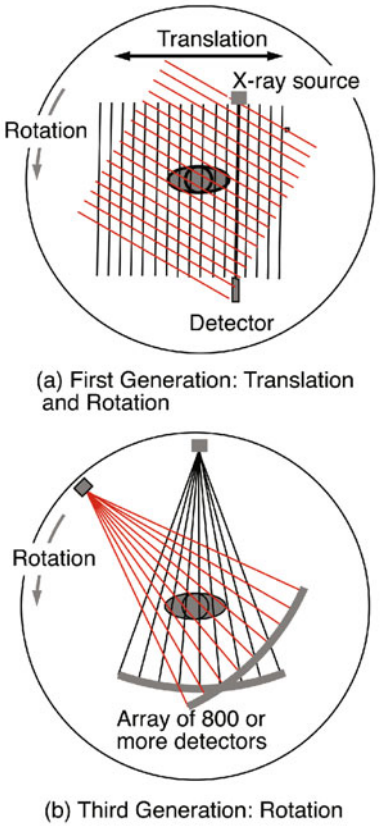


Fig. 16.25 Scanning techniques used in CT scanners. **a** In the earliest scanners the x-ray tube and detector moved in synchronization on either side of the subject. Then their translation path was rotated one degree. **b** Nearly all machines now use the "third generation" configuration: a fan beam and multidetector array rotate continuously around the patient

There may be a single row of detectors or multiple detector rows parallel to the z axis. Table 16.3 shows how scanners have improved since they were first introduced. Spiral CT provides $\mu(x, y, z)$, and the images can be displayed in three dimensions.

Figure 16.26 is an abdominal scan showing a benign tumor in the liver. Computer analysis of $\mu(x, y, z)$ data can be used to display 3-dimensional images of an organ (Fig. 16.27). The surface of an organ is defined by a change in μ.

It is often desirable to measure the attenuation coefficient with an accuracy of ±0.5 %. For water at 60 keV, $\mu = 20\ \text{m}^{-1}$, so μ must be measured with an accuracy of $\delta\mu = 0.1\ \text{m}^{-1}$. (A beam of 120 kVp with 2–3 mm of aluminum filtration has about this average photon energy). It is customary to report the fractional difference between μ and μ_{water}. The *Hounsfield unit* is

$$H = 1000\frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}}}. \qquad (16.25)$$

The desired accuracy is ±5 Hounsfield units.

Table 16.3 The evolution of typical values for high-performance CT machines. (Adapted from Kalender 2011, p. 41)

Feature	1972	1980	1990	2000	2010
Minimum scan time	300 s	5–10 s	1–2 s	0.33–0.5 s	0.27–0.35 s
Data per 360 ° scan	57.6 kB	0.2–1 MB	1–2 MB	5–20 MB	0.1–1 GB
Data per spiral scan			12–24 MB	0.1–1 GB	1–100 GB
Image matrix	80 × 80	256 × 256	512 × 512	512 × 512	512 × 512
Power (kW)	2	10	40	60–100	80–120
Slice thickness (mm)	13	2–10	1–10	0.5–1	0.4–0.6
Spatial resolution (Line pair cm^{-1})	3	8–12	10–15	12–16	12–25

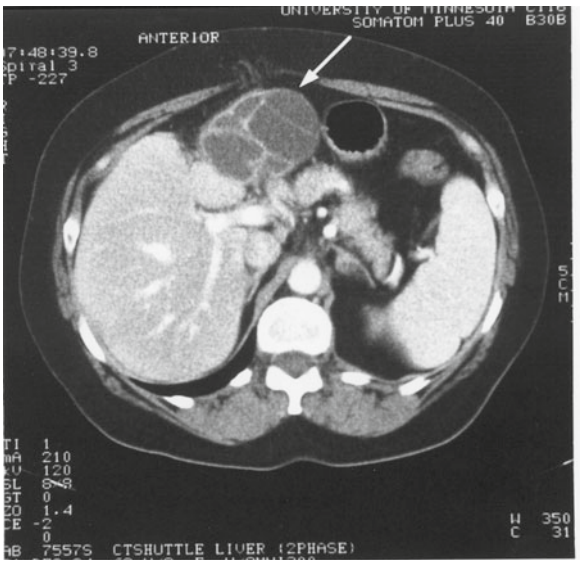


Fig. 16.26 A spiral CT scan of the abdomen. The arrow points to a biliary cystadenoma, a benign tumor of the liver. (Scan courtesy of E. Russell Ritenour, Ph.D., Department of Radiology, University of Minnesota Medical School)

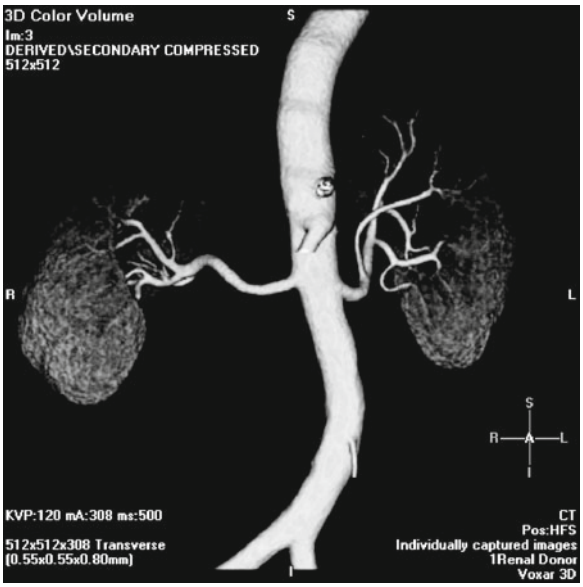


Fig. 16.27 A 3-dimensional rendering of the aorta, renal arteries and kidneys. (Scan courtesy of E. Russell Ritenour, Ph.D., Department of Radiology, University of Minnesota Medical School)

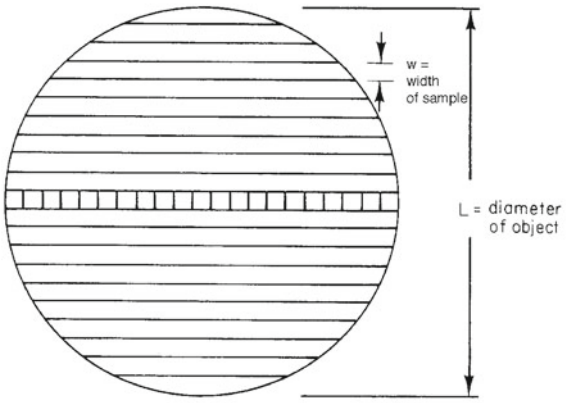


Fig. 16.28 A circular object that is to be analyzed. The diameter of the object is L; the width of each sample in the scan is w. It is desired to resolve voxels in each sample which have a length w on each side, as shown for the center diameter

There is a fundamental relationship between the dose to the patient and the resolution. We derive it here for a first-generation machine. Suppose we are reconstructing the image of an object with a circular cross-section as shown in Fig. 16.28. The object is to be resolved into cubic volume elements or *voxels* of length w on each side parallel to the x, y and z axes. The length of each voxel along the z axis perpendicular to the scan is the slice thickness. The diameter of the object is L. For simplicity, we make the analysis assuming a first-generation machine, with rectilinear passes repeated at m different angles between 0 and 180 °. The width of each sample in a scanner pass is w. The number of samples in each pass is

$$n = \frac{L}{w}, \tag{16.26}$$

and the number of voxels in the object is approximately the area of the circular object divided by the area of the voxel in the plane of the slice: $\pi L^2/4w^2$ or $\pi n^2/4$. To determine $\pi n^2/4$ independent values of μ requires at least that many independent measurements. Since n measurements are made in each pass, we need m passes where $mn = \pi n^2/4$ or

$$m = \frac{\pi n}{4}. \tag{16.27}$$

With more passes the situation is overdetermined; with fewer it is underdetermined. If the values of μ are overdetermined, convolved back projection (Chap. 12) or a similar procedure can be used to assign the values of μ.

Now consider the attenuation of photons along a diameter of the object in one pass. In Sect. 16.5 we developed a relationship between the photon fluence in the beam needed to measure the attenuation with some desired accuracy (Eq. 16.21). The same arguments can be applied in this case:

$$\delta\Phi = \Phi_1 w\, \delta\mu.$$

The photons arrive at the detector, which we assume for simplicity to have 100 % efficiency, at a constant average rate, so they are Poisson distributed. The standard deviation in the number of counts is $(\Phi_1 S)^{1/2} = (\Phi_1 w^2)^{1/2}$. To detect the difference between the two samples, $w^2\delta\Phi$ must exceed this by the minimum signal-to-noise ratio, k. This gives the minimum photon fluence at the detector:

$$\Phi_1 > \frac{k^2}{w^4(\delta\mu)^2}. \tag{16.28}$$

It can be shown (Brooks and DiChiro 1976a) that these counts can be divided among all the passes. Since the dose is proportional to Φ_1, this equation shows a fundamental relationship between dose and resolution. Decreasing w by a factor of 2 requires a 16-fold increase in dose, while improving $\delta\mu$ by a factor of 2 requires a dose that is 4 times as large. For further discussion of this equation, see Kalender (2011) pp. 169–170. We discuss CT dose in Sect. 16.12. Reducing the dose is a matter of great current interest (Tack and Gevenois 2007).

16.9 Biological Effects of Radiation

Radiation at sufficiently high doses can kill cells, tumors, organs, or entire animals. Radiation, along with surgery and chemotherapy, is a mainstay of cancer treatment. Radiation can also cause mutations. *Radiobiology*, the study of how radiation affects cells and organs, has provided major improvements in our understanding of cell death and damage. This understanding has modified and improved our approach to radiation therapy. This section provides a brief introduction to radiobiology, but it ignores many important details. For these details see Hall and Giaccia (2012). The discussion starts with some cell-culture (*in vitro*) results, presents the most frequently used model for radiation damage, and then moves to *in vivo* tissue irradiation and the eradication of tumors.

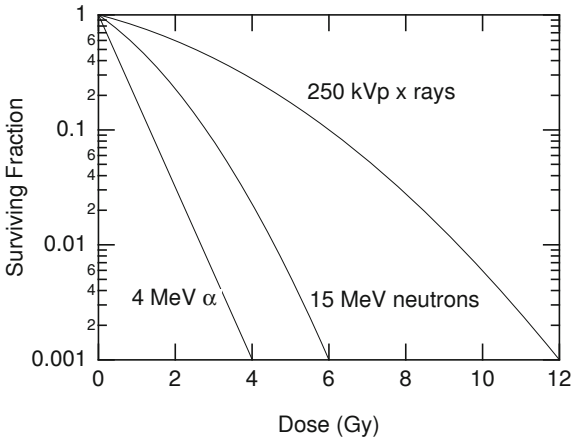


Fig. 16.29 Typical survival curves for cell culture experiments, for 4-MeV α particles, 15-MeV neutrons, and 250-kVp x-rays. These are representations of typical experimental data

There are two types of effects. *Deterministic* or *tissue reactions* occur immediately (early effects) and include skin reddening (erythema) and cataracts. Late effects are stochastic and include cancer and mental retardation for fetal irradiation exceeding 0.3 Gy. We discuss only stochastic effects here.

16.9.1 Cell-Culture Experiments

Cell-culture studies are the simplest conceptually. A known number of cells are harvested from a stock culture and placed on nutrient medium in plastic dishes. The dishes are then irradiated with a variety of doses including zero as a control. After a fixed incubation period the cells that survived irradiation have grown into visible colonies that are stained and counted. Measurements for many absorbed doses give *survival curves* such as those in Fig. 16.29. These curves are difficult to measure for more than two or three decades, because of the small number of colonies that remain.

Failure to survive means either that the cell was killed or that it can no longer reproduce. If the cells die attempting the next or a later cell division (*mitosis*) it is called *mitotic death*. Some cells die by *apoptosis:* a mechanism whereby the cell initiates its own programmed death, going through a well-defined series of morphologic events that culminate in fragmentation of the DNA (Hall and Giaccia 2012). Experiments with microscopic beams of radiation and short-range particles aimed at different parts of the cell have demonstrated that damage to the cell's DNA is only one factor in the cell's response to the radiation. Nonetheless, the differences in survival curve shapes that we discuss here are still important.

The shape of the survival curve depends on the linear energy transfer (LET) of the charged particles. For the α particles in Fig. 16.29 the LET is about 160 keV μm^{-1}, for neutrons it is about 12 keV μm^{-1}, and for the electrons from the 250-kVp x-rays it is about 2 keV μm^{-1}. The α particles and neutrons are called high-LET radiation; the electrons are low-LET radiation.

High-LET radiation produces so many ion pairs along its path that it exerts a direct action on the cellular DNA. Low-LET radiation can also ionize, but it usually acts indirectly. It ionizes water (primarily) according to the chemical reaction

$$H_2O \rightarrow H_2O^+ + e^-.$$

The H_2O^+ ion decays with a lifetime of about 10^{-10} s to the hydroxyl free radical:

$$H_2O^+ + H_2O \rightarrow H_3O^+ + OH.$$

This then produces hydrogen peroxide and other free radicals that cause the damage by disrupting chemical bonds in the DNA.

16.9.2 Chromosome Damage

Cellular DNA is organized into *chromosomes*. In order to understand radiation damage to DNA, we must recognize that there are four *phases* in the cell division cycle:

- M Cell division. This stage includes both division of the nucleus (*mitosis*) and of the cytoplasm (*cytokinesis*). This phase may last 1 or 2 h.
- G_1 The first "gap" phase. The cell is synthesizing many proteins. The duration of G_1 determines how frequently the cells divide. It varies widely by kind of tissue, from a few hours to 200 h.
- S Synthesis. A new copy of all the DNA is being made. This lasts about 8 h.
- G_2 The second "gap" phase, lasting about 4 h.

Figure 16.30 shows, at different magnifications, a strand of DNA, various intermediate structures that we will not discuss, and a chromosome as seen during the M phase of the cell cycle. The size goes from 2 nm for the DNA double helix to 1400 nm for the chromosome. In addition to cell survival curves one can directly measure chromosome damage. There is strong evidence that radiation, directly or indirectly, breaks a DNA strand. If only one strand is broken, there are efficient mechanisms that repair it over the course of a few hours using the other strand as a template. If both strands

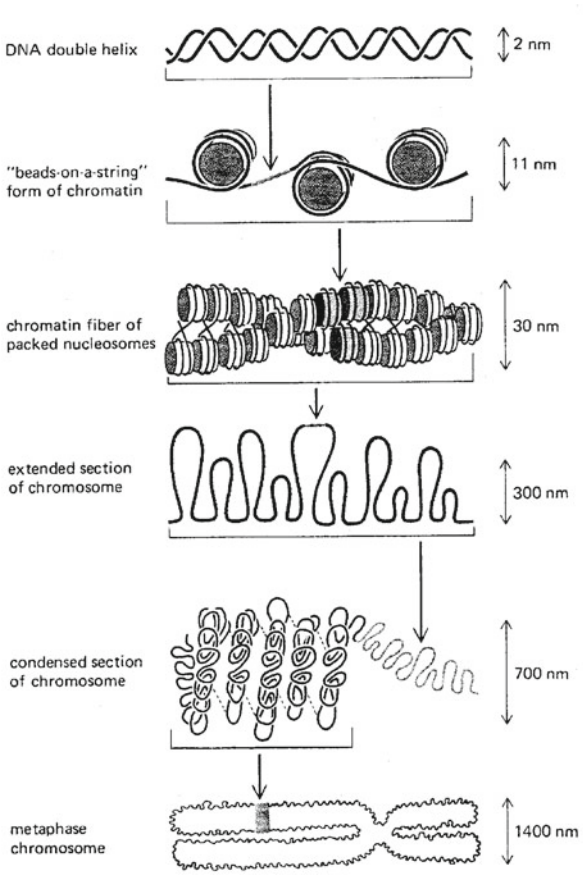


Fig. 16.30 A schematic diagram of how the DNA is packed to give a chromosome, shown at metaphase of the cell cycle. (Republished with permission of Taylor and Francis Group from Alberts et al. (1999, p. 230). Permission conveyed through Copyright Clearance Center, Inc)

are broken, permanent damage results, and the next cell division produces an abnormal chromosome.[18] Several forms of abnormal chromosomes are known, depending on where along the strand the damage occurred and how the damaged pieces connected or failed to connect to other chromosome fragments. Many of these chromosomal abnormalities are lethal: the cell either fails to complete its next mitosis, or it fails within the next few divisions. Other abnormalities allow the cell to continue to divide, but they may contribute to a multistep process that sometimes leads to cancer many cell generations later.

Even though radiation damage can occur at any time in the cell cycle (albeit with different sensitivity),[19] one looks for chromosome damage during the next M phase, when the

[18] This is a simplification. It is possible for a double strand break to repair properly. See Hall and Giaccia (2012, p. 18).

[19] In general, cells exhibit the greatest sensitivity in M and G_2.

DNA is in the form of visible chromosomes as in the bottom example in Fig. 16.30. If the broken fragments have rejoined in the original configuration, no abnormality is seen when the chromosomes are examined. If the fragments fail to join, the chromosome has a "deletion." If the broken ends rejoin other broken ends, the chromosome appears grossly distorted.

A sequence of processes leads to cellular inactivation. Ionization is followed by initial DNA damage. Most of this is repaired, but it can be repaired incorrectly. No repair or faulty repair results in DNA lesions that are then manifest as chromosome aberrations, which may be nonlethal, may cause mutations, or may lead to cell death. The numbers quoted here are from the review by Steel (1996). A cell dose of 1 Gy leads to the production of about 2×10^5 ion pairs per cell nucleus, of which about 2000 are produced in the cell's DNA. It has been estimated that the amount of DNA damage immediately after radiation can be quite large: 1000 single-strand breaks and 40 double-strand breaks per Gy. Yet survival curves for different cell types show between 0.3 and 10 lethal lesions per gray of absorbed dose. Thus the amount of repair that takes place is quite large, and the model introduced below is an oversimplification.

A number of chemicals enhance or inhibit the radiation damage. Some chemical reactions can "fix" (render permanent) the DNA damage, making it irreparable; others can scavenge and deactivate free radicals. One of the most important chemicals is oxygen, which promotes the formation of free radicals and hence cell damage. Cells with a poor oxygen supply are more resistant to radiation than those with a normal supply.

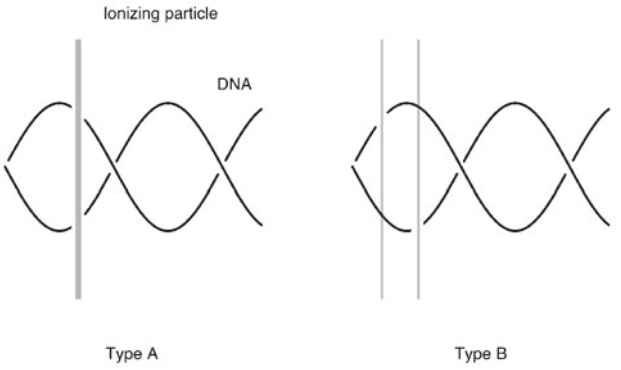


Fig. 16.31 The two postulated types of DNA damage from ionizing radiation for our simple model to explain the linear-quadratic cell culture survival curve. In type-A damage a single ionizing particle breaks both strands. Two ionizing particles are required for type-B damage, one breaking each strand

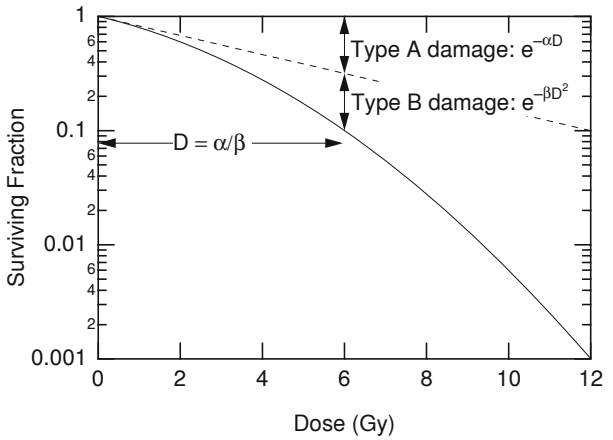


Fig. 16.32 A survival curve, showing the linear exponent for type-A damage and the quadratic exponent for type-B damage

16.9.3 The Linear-Quadratic Model

The *linear-quadratic model* is often used to describe cell survival curves. We will extended it to very small survival rates that cannot possibly be confirmed experimentally. We use it as a simplified model for DNA damage from ionizing radiation that recognizes two types of damage, shown in Fig. 16.31. In type-A damage a single ionizing particle breaks both strands of the DNA, and the chromosome is broken into two fragments. In type-B damage, a single particle breaks only one strand. If another particle breaks the other strand "close enough" to the first break before repair has taken place, then the chromosome suffers a complete break.

The probability of type-A damage is proportional to the dose. The average number of cells with type-A damage after dose D is $m = \alpha D = D/D_0$, and the probability of no damage is the Poisson probability $P(0; m) = e^{-m} = e^{-\alpha D}$. This is the dashed line in Fig. 16.32, which is redrawn from Fig. 16.29. For radiations with higher LET the proportionality constant is greater, as seen in Fig. 16.29.

In type-B damage one strand is damaged by one ionizing particle and the other by another ionizing particle. The probability of fragmenting the DNA molecule is therefore proportional to the square of the dose. The average number of molecules with type-B damage is βD^2, and the survival curve for type-B damage alone is $e^{-\beta D^2}$, also shown in Fig. 16.32. This leads to the linear-quadratic model for cell survival:

$$P_{\text{survival}} = e^{-\alpha D - \beta D^2}. \tag{16.29}$$

The dose at which mortality from each mechanism is the same is α/β, as shown in Fig. 16.32.

An extension of the cell survival experiments is the *fractionation curve* shown in Fig. 16.33. After a given dose, cells from the culture were harvested and used to inoculate new cultures. After a few hours they were irradiated again. The survival curve plotted against total dose starts anew from the

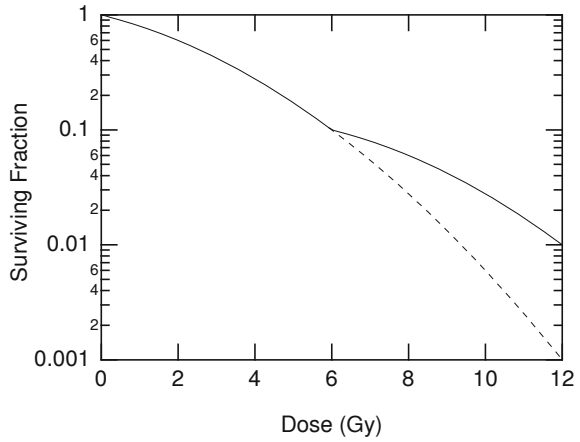


Fig. 16.33 If the dose for low-LET radiations is divided into fractions, with a few hours between fractions, all of the single-strand breaks have been repaired, and survival follows the same curve as for the original fraction

point corresponding to the first irradiation. The initial dose of 6 Gy caused both type-A and type-B damage. Before the second dose, the cells with single-strand damage had been repaired, and when the second dose was given, it acted on undamaged cells, so that only type-A damage occurred for small additional doses.

16.9.4 The Bystander Effect

Ionization damage is not the entire story. The *bystander effect* in radiobiology refers to the "induction of biological effects in cells that are not directly traversed by a charged particle, but are in close proximity to cells that are" (Hall 2003; Hall and Giaccia 2012).

One experiment showing the bystander effect involves irradiating cells in culture and transferring some of the culture medium to unirradiated cells, which then respond as if they had been irradiated. The effect is absent if the irradiated medium contains no cells. The irradiated cells secreted some chemical into the medium that affected the unirradiated cells. In one such experiment, apoptosis was induced in the unirradiated culture by quite low doses to the irradiated cells. The dose response curve was nearly flat.

Another type of experiment used microbeams of α particles to irradiate specific cells in a culture, and then measured the response of neighboring cells which had not been irradiated. The survival of cells not irradiated decreased as their neighbors were hit with more α particles. It is thought that some chemical produced in the irradiated cells migrated into the unirradiated cells through gap junctions connecting the cytoplasm of neighboring cells. Similar experiments are done with radioactive nuclides that emit very-short range Auger electrons (see Chap. 17). The nuclides are attached to molecules that are selectively taken up by the cell nucleus or cytoplasm or that bond to the cell's DNA (Kassis 2004).

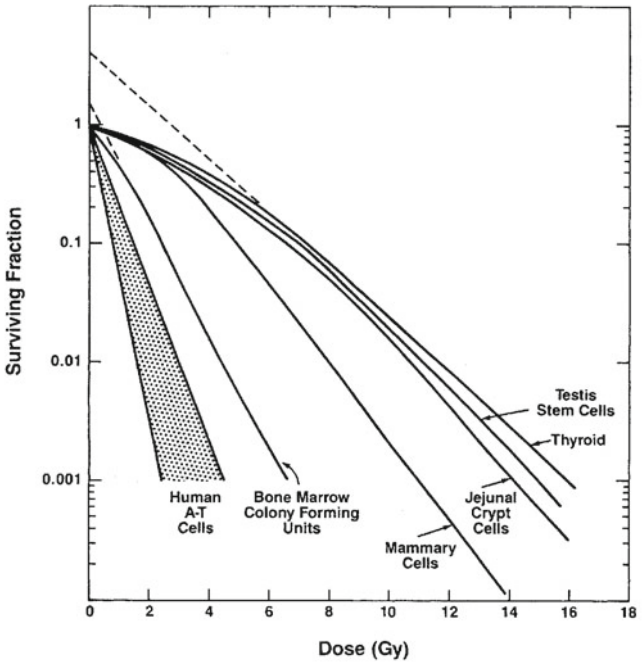


Fig. 16.34 Survival curves for assays of human cells. There is a wide range in initial sensitivity, but not too much difference in final slope. The shaded area labeled "human A-T cells" is for cells from a disease, ataxia-tangliectasia, where repair mechanisms are lacking. (Reproduced with permission from Hall 2002, p. 328)

16.9.5 Tissue Irradiation

There is considerable variation in the shape of the survival curves for human cells (Fig. 16.34). The shaded area labeled "human A-T cells" is for cells from patients with a genetic disease, *ataxia-tangliectasia*, where repair mechanisms are lacking and breakage of a single strand of DNA leads to cell death.

The radiation damage to the DNA is not apparent until the cell tries to divide. At that point, the chromosomes are either so badly damaged that the cell fails to divide or the damage survives in later generations as a mutation. Some tissues respond to radiation quite quickly; others show no effect for a long time. This is due almost entirely to the duration of the G_1 phase or the overall time between cells divisions. Tissues are divided roughly into two groups: *early-responding* and *late-responding*. Early-responding tissues include most cancers, skin, the small and large intestine, and the testes. Late-responding tissues include spinal cord, the kidney, lung, and urinary bladder.

The central problem of radiation oncology is how much dose to give a patient, over what length of time, in order to

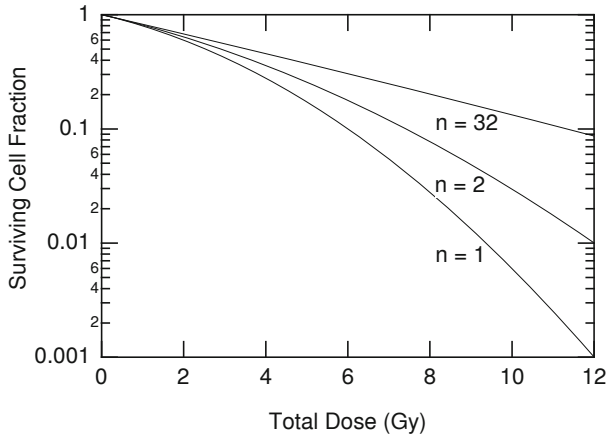


Fig. 16.35 The fraction of cells surviving a total radiation dose when the dose is divided into 1, 2, and 32 fractions, showing how the curve approaches $e^{-\alpha D}$ as the number of fractions becomes large

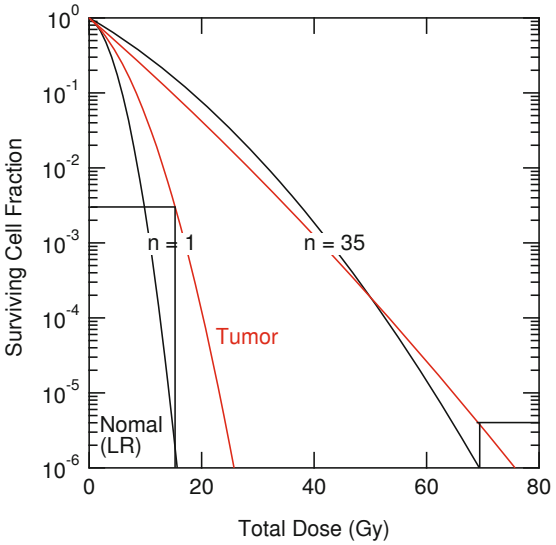


Fig. 16.36 Cell survival curves for late-responding normal tissue (*LR*) and for a hypothetical tumor (in red), showing the improvement obtained by dividing the dose into fractions. With a single fraction, the tumor survives much better than the normal tissue. With 35 fractions, this discrepancy has been reduced. The details are discussed in the text

have the greatest probability of killing the tumor while doing the least possible damage to surrounding normal tissue. While the dose is sometimes given all at once (over several minutes), it is usually given in *fractions* five days a week for four to six weeks. Some recent treatment plans, primarily for brachytherapy (see Sect. 17.11), use fractions given every few hours.

What total dose (or dose per fraction) should be given in how many fractions, with what time between fractions? We can gain some insight by using the linear-quadratic model. Let the dose per fraction be D_f, the number of fractions be n, and the total dose be $D = nD_f$. We plot survival vs. total dose for different numbers of fractions. We assume that the time between fractions allows for full repair of sublethal damage (single-strand breaks). The probability of a cell surviving after n fractions have been delivered is

$$P_s = P_{\text{survival}} = S = \left(e^{-\alpha D_f - \beta D_f^2}\right)^n = e^{-\alpha D - \beta D^2/n}. \tag{16.30}$$

As the number of fractions becomes very large for a given total dose, the survival curve approaches $e^{-\alpha D}$. This can be seen in Fig. 16.35, which plots survival vs. total dose delivered in different numbers of fractions. With many fractions the dose per fraction is very small, all the single-strand breaks are repaired, and almost no type-B cell deaths take place.

Early-responding tissue and tumors have been found to have an α/β ratio of about 10 Gy. The survival curve is primarily due to type-A damage. Late-responding tissues have an α/β ratio of 2–3 Gy. There is considerable variation in these numbers.

Some of the problems of radiation therapy and the benefits of fractionation can be seen if we consider a strictly hypothetical example (a toy model) in which $\alpha = 0.15\ \text{Gy}^{-1}$ for the tumor and $0.1\ \text{Gy}^{-1}$ for the surrounding tissue. The tumor is early responding with $\alpha/\beta = 10$ Gy, and the surrounding tissue is late responding with $\alpha/\beta = 2$ Gy. Figure 16.36 shows the cell-survival curves for 1 and 35 fractions. The tumor survival in each case is shown in red.

To see the benefit of fractionation, suppose that the patient can tolerate a dose at which only 10^{-6} of the cells of the surrounding tissue survive, represented by the horizontal line on the graph. (This is not realistic!) For a single fraction, this corresponds to a total dose of about 15 Gy, which, applied to the red line, shows that the surviving fraction of tumor cells is about 3×10^{-3}. For 35 fractions the normal tissue can tolerate about 70 Gy, yielding 4×10^{-6} as the fraction of tumor cells surviving.

Suppose next that it is possible to confine the radiation beam so that the dose to normal tissue is only about 0.6 times that to the tumor. This means that the tissue dose in Eq. 16.30 is multiplied by 0.6. The result is shown in Fig. 16.37 for 35 fractions. The tumor dose can now be as high as 115 Gy for the same effect on surrounding tissues, leading to a tumor survival of only 10^{-10}. We will see how beam shaping is accomplished in the next section.

These calculations are solely to illustrate the basic principles, and the doses are not realistic. Clinically useful calculations must take several additional factors into account: the actual values of α for the tissue and tumor under consideration, the effect of cell growth after irradiation, the effect of the first dose on synchronizing the cycles of the remaining cells, and the oxygen level in the tumor cells. (The greater

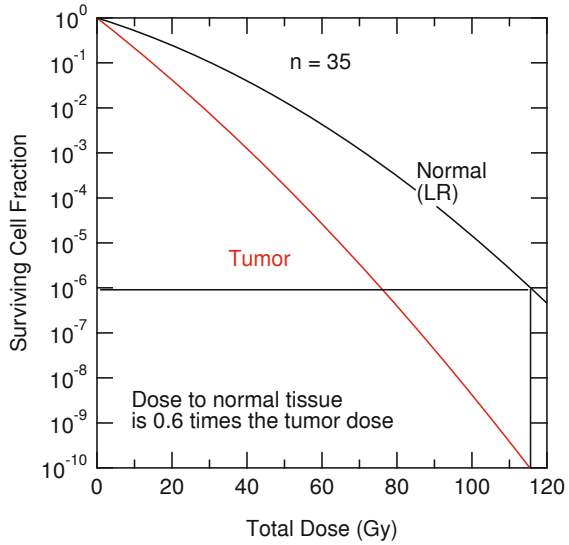


Fig. 16.37 Survival curves for the same cells as in the previous figure, with the dose to the surrounding tissue reduced to 0.6 times that to the tumor. Now the probability of tumor survival at high doses is about 0.0001 times that for the surrounding normal tissue. This shows the importance of confining the radiation to the tumor as much as possible

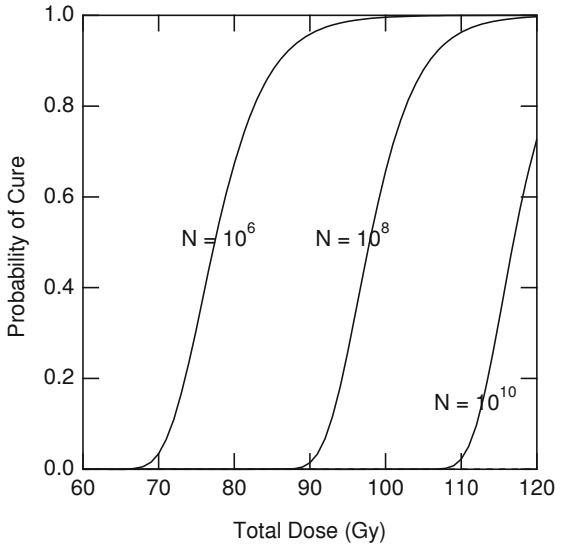


Fig. 16.38 The probability of eradicating the tumor (no surviving tumor cells) as a function of dose for tumors containing different numbers of cells

the oxygen concentration the more sensitive the cells are, particularly for low-LET radiation. Rapidly growing tumors often outstrip their blood supply, receive less oxygen, and are less radio-sensitive.) Fractionation is reviewed in Orton (1997) and in Hall and Giaccia (2012). It is also necessary to take into account the fact that neither the tumor nor the surrounding normal tissue receives a uniform dose of radiation.

16.9.6 A Model for Tumor Eradication

The target theory model can be applied to a collection of cells to give us insight into the central problem of radiation therapy: *tumor eradication*. Suppose that a tumor consists of N cells with identical properties. The cells are uniformly irradiated with dose D. If a collection of identical tumors were irradiated, the number of cells surviving in each tumor would fluctuate. The probability that a single cell survives is $p_s(D)$, which might be given by Eq. 16.30. If this number is small and N is large, the number surviving follows a Poisson distribution. The average number surviving is $m = Np_s(D)$. The probability of a cure is the probability that no tumor cells survive:

$$P_{\text{cure}} = e^{-m} = e^{-Np_s(D)}. \tag{16.31}$$

This can be evaluated using your cell-survival model of choice.

Figure 16.38 shows a tumor eradication curve based on the 35-fraction curve in Fig. 16.36. The larger the tumor, the greater the dose required for cure. Figure 16.39 shows a plot of the probability of tumor cure and the probability of unacceptable damage to the surrounding tissue. For this example, at least 60 Gy are required in order to have a good probability of cure; once the dose is higher than 63 Gy, the damage to normal tissue is unacceptable.

16.10 Radiation Therapy

The treatment of cancer must deal with two issues: eradication of the primary tumor (*local control*), and eradication of *metastases*, which may be in nearby tissue or may be at distant sites in the body. In many cases radiation therapy, either alone or combined with surgery, is the best technique for local control. Two oncologists have provided a review of the benefits and problems of radiation therapy, addressed specifically to the medical physics community (Schulz and Kagan 2002). They point out that many cancer deaths are due to metastatic disease, so improved local control does not necessarily provide a corresponding improvement in survival. Ratliff (2009) provides a survey of literature about radiation therapy.

Which method of treatment is best can change dramatically as new treatments are developed. For example, a combination of radiation therapy and chemotherapy was once used to treat Hodgkin's disease; chemotherapy has been improved to the point where radiation is no longer necessary (DeVita 2003).

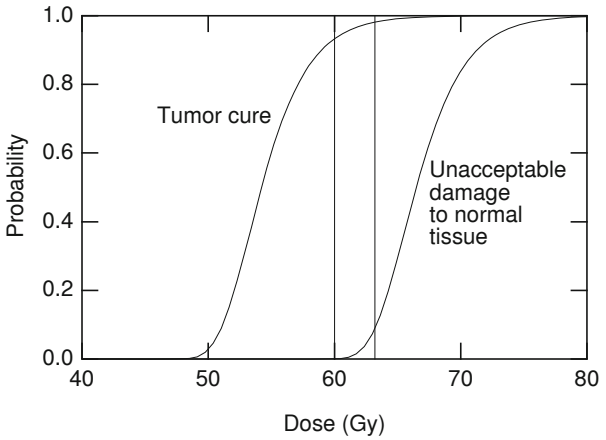


Fig. 16.39 The probability of curing the tumor and the probability of unacceptable damage to normal tissue vs. dose



Fig. 16.40 X-ray therapy was used to treat a carcinoma of the nose. *A* shows the original lesion; *B* is the result one year later. The patient remained asymptomatic 5 years after treatment. (Reprinted from William and James 1989, with permission from Elsevier)

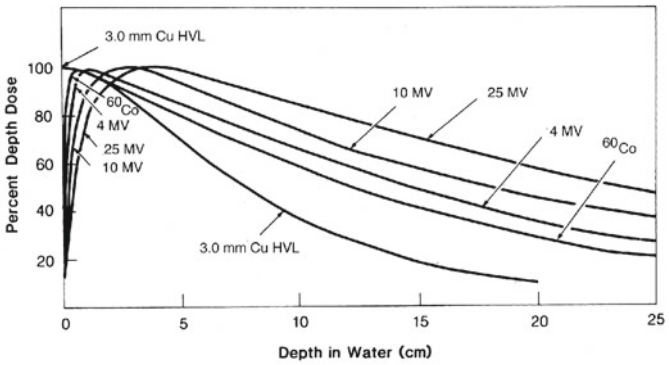


Fig. 16.41 The dose vs. depth for x-ray beams of different quality (energy) on the central axis of the beam. The source–surface distance (SSD) is 100 cm and the field size is 10×10 cm. The curve "3.0 mm Cu HVL" is for a photon beam that is reduced to half intensity by a copper filter 3.0 mm thick. The radioactive element ^{60}Co emits two gamma ray photons (1.17 and 1.33 MeV.) The labels 4, 10 and 25 MV refer to the energy of the electron beam striking the x-ray tube anode. (From Khan 2003, p.163. ©2003 Lippincott Williams & Wilkins)

16.10.1 Classical Radiation Therapy

Doses for diagnostic radiology vary from about 10^{-4} to 10^{-2} Gy. Doses of 20–80 Gy are required to treat cancer. A great deal of physics is involved in planning the treatment for each patient [See Khan (2010) or Goitein (2008)]. There is a choice of radiation beams: photons of various energies, electrons, neutrons, protons, or α particles. Photons and electrons are routinely available; the other sources require special facilities. The number of proton facilities is growing rapidly. Only a few of the beam issues will be raised here. Some of the dose measurement issues are discussed in the next section.

An example of the effectiveness of radiation therapy is shown in Fig. 16.40 The patient developed a carcinoma of the nose and refused surgery. Radiation with a total dose of 50 Gy was used, and the results one year later are shown. It is ironic that the carcinoma probably developed because the patient was treated with x-rays for acne many years earlier.

We have already seen the importance of reducing dose to tissue surrounding the tumor. Optimizing the dose determines the kind of radiation to be used and its energy, as well as the details of beam filtration and collimation and how it is aimed at the patients body. For now, we discuss a photon beam. Attenuation and the $1/r^2$ decrease of photon fluence help spare tissue downstream in the beam. Since μ_{atten} decreases with increasing photon energy up to a few MeV, higher-energy photons penetrate more deeply and must be used for treating deeper lesions. There is also dose buildup with depth over distances comparable to the range of the Compton-scattered electrons. Both of these effects are shown in Fig. 16.41.

The beam is *collimated* to spare normal tissue. Originally, the collimator consisted of four lead jaws that provided a rectangular opening with adjustable length and width. A wedge was sometimes placed in the beam to vary the intensity across the collimated radiation field.

Figure 16.42 shows *isodose contours* for various beams. In addition to the differences with depth seen in Fig. 16.41, there are significant differences in the sharpness of the dose distribution across the beam. The extent of the lesion to be radiated must be carefully determined with radiographs or CT scans.

If the tumor is not near the surface, the ratio of tumor dose to normal tissue dose can be increased by irradiating the patient from several directions. Figure 16.43 shows how the relative dose to a deep tumor can be increased by irradiating with two *fields* on opposite sides of the patient. In Fig. 16.44

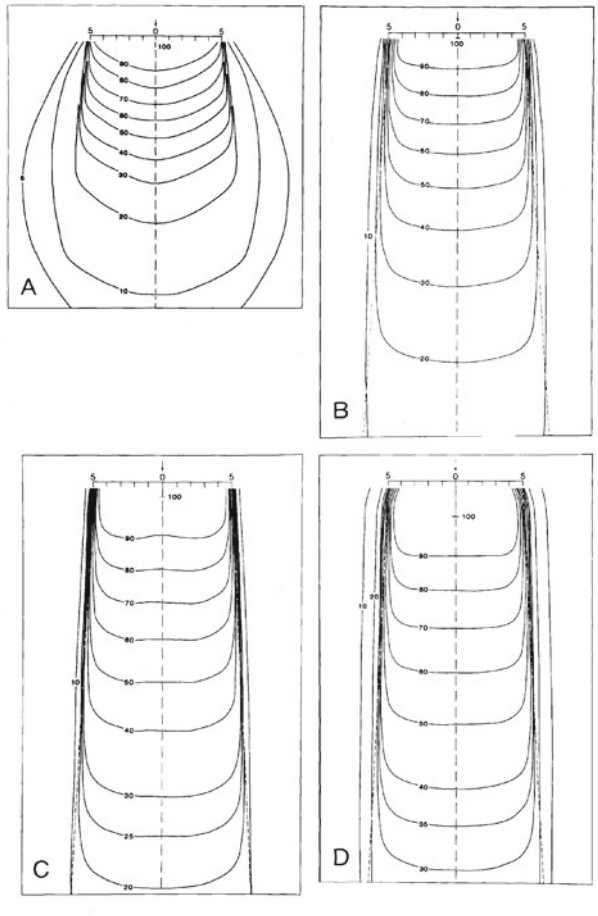


Fig. 16.42 Isodose distributions for radiation under different conditions, all collimated to 10×10 cm. **a** Radiation from an x-ray tube with 200 kVp, 0.5 m from the surface. **b** Photons from the radioactive isotope ^{60}Co, 0.8 m from the surface. **c** 4-MV photons, 1 m from the surface. **d** 10-MV photons, 1 m from the surface. (From Khan (2003, p. 204). ©2003 Lippincott Williams & Wilkins)

three and four fields are used. The angles of the fields can be changed by rotating the patient couch as well as the gantry holding the photon source and collimator.

Rectangular fields do not match the shape of the tumor. To overcome this problem a *multileaf collimator* replaces the four original jaws on the therapy machine. A typical multileaf collimator has up to 100 pairs of tungsten alloy leaves, each a few mm wide, which can be independently adjusted to provide a pattern like that in Fig. 16.45. This might be used for up to nine fields from different directions.

16.10.2 Modern X-Ray Therapy

The goal of radiation therapy is to provide as large a dose as possible to the tumor while sparing adjacent normal tissue.

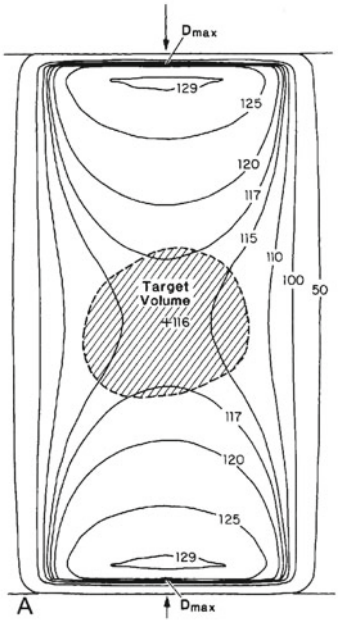


Fig. 16.43 Isodose distribution when the patient is irradiated equally from opposite sides. (From Khan (2003, p. 210). ©2003 Lippincott Williams & Wilkins)

The normal tissue may be quite close to the tumor. *Three-dimensional conformal radiation therapy* uses 3-dimensional information about the target volume. This is difficult, because even with 3-dimentional display of CT, MRI or ultrasound images, it may be impossible to see the edges of the tumor. Nevertheless, the *beam's-eye view* that can be computed from 3-d image data can be very useful in planning the treatment. For a discussion of conformal radiation therapy, see Khan (2010), Chap. 19.

In classical radiotherapy, the beam was either of uniform fluence across the field, or it was shaped by an attenuating wedge placed in the field. *Intensity-modulated radiation therapy* (IMRT) is achieved by stepping the collimator leaves during exposure so that the fluence varies from square to square in Fig. 16.45 (Goitein 2008; Khan 2010, Ch. 20)

It was originally hoped that CT reconstruction techniques could be used to determine the collimator settings at different angles. This does not work because it is impossible to make the filtered radiation field negative, as the CT reconstruction would demand. IMRT with conventional treatment planning improves the dose pattern (Goitein (2008); Yu et al. (2008)), providing better sparing of adjacent normal tissue and allowing a boost in dose to the tumor.

One problem in radiation therapy is movement of organs when the patient breathes. Four-dimensional CT records data at a fixed point in the respiratory cycle. The radiotherapy beam is turned on only at the same point in the cycle (Khan 2010, Ch. 25).

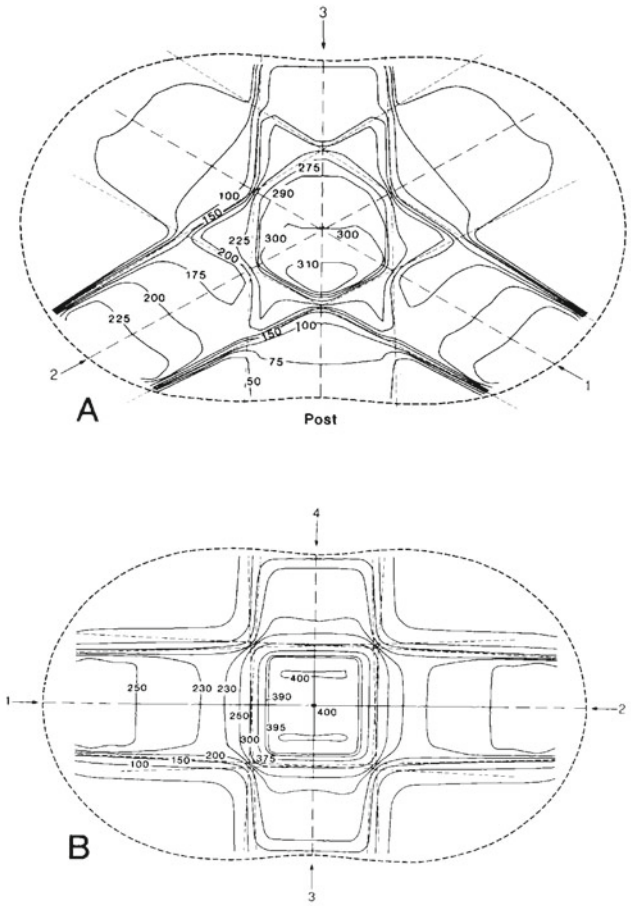


Fig. 16.44 Isodose distribution for (**a**) three and (**b**) four radiation fields, each designed to give a relative dose of 100 at the center of the tumor. (From Khan 2003, p. 215. © 2003 Lippincott Williams & Wilkins)

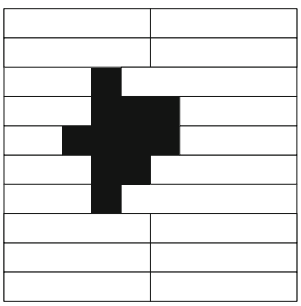


Fig. 16.45 A multileaf collimator (MLC). The tungsten leaves are shown in white; the opening is black

16.10.3 Charged Particles and Neutrons

Electrons, typically between 6 and 20 MeV, are also used for therapy (Hogstrom and Almond (2006)). Because of the range–energy relationship, the field falls nearly to zero in a few centimeters. Electrons are used primarily for skin and lip cancer, head and neck cancer, and irradiation of lymph nodes near the surface. Figure 16.46 shows the dose vs. depth as a percent of the maximum dose for electron beams of several different energies.

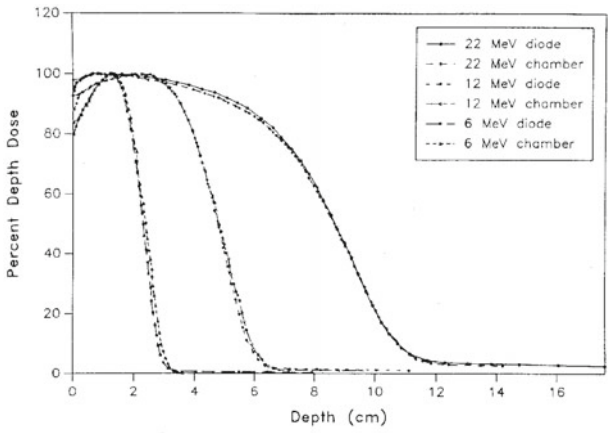


Fig. 16.46 Depth–dose curves for electrons of different energies, measured with a solid-state detector (diode) and an ionization chamber. Both the range and the straggling increase with increasing energy. (From F. M. Khan 1986). AAPM Monograph 15.

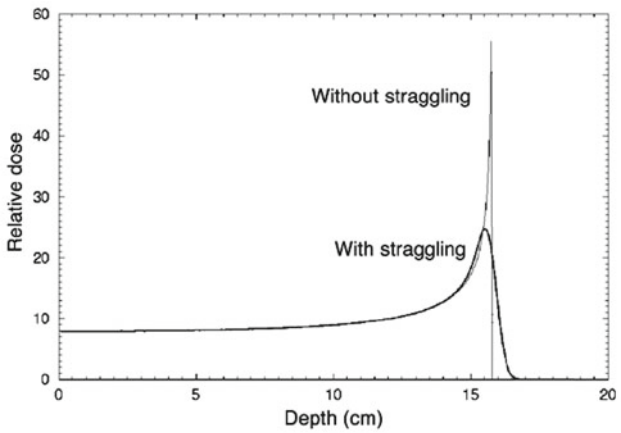


Fig. 16.47 Energy loss vs. depth for a 150 MeV proton beam in water, with and without straggling. The Bragg peak enhances the energy deposition at the end of the proton range. (Copyright © 2005 W. D. Newhauser, M. D. Anderson Cancer Center. Used by permission)

Protons are also used to treat tumors (Khan 2010, Ch. 26; Goitein 2008). Their advantage is the increase of stopping power at low energies. It is possible to make them come to rest in the tissue to be destroyed, with an enhanced dose relative to intervening tissue and almost no dose distally ("downstream") as shown by the *Bragg peak* in Fig. 16.47. Placing an absorber in the proton beam before it strikes the patient moves the Bragg peak closer to the surface. Various

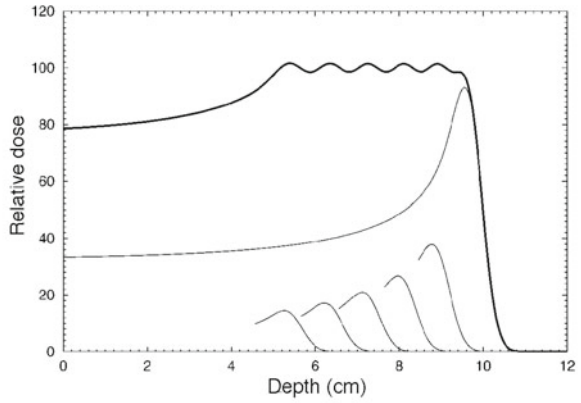


Fig. 16.48 Irradiating the patient through a number of absorbers of different thickness spreads out the region of maximum dose. (Copyright © 2005 by W. D. Newhauser, M. D. Anderson Cancer Center. Used by permission)

techniques, such as rotating a variable-thickness absorber in the beam, are used to shape the field by spreading out the Bragg peak (Fig. 16.48). The edges of proton fields are sharper than for x-rays and electrons (Delaney and Kooy 2008). This can provide better tissue sparing, but it also means that alignments must be more precise. Another technique is to extract the protons from the accelerator at the desired energy and use magnets to sweep the resulting beam across the desired region of the patient (Goitein 2008).

Sparing tissue reduces side effects immediately after treatment. It also reduces the incidence of radiation-induced second cancers many years later. These are particularly important in pediatric patients as the initial treatment proves more successful and the patients survive longer. Miralbell et al. (2002) estimated the incidence rate for secondary cancers in certain pediatric cancers and found a reduced incidence for proton therapy compared to both conventional x-ray and IMRT.

Proton therapy is used in a number of diseases. Delaney and Kooy (2008) provide an extensive review. Some institutions are experimenting with intensity-modulated proton therapy (IMPT) (Xu et al. 2008).

Fast neutrons are used for therapy (Duncan 1994). The dose is due to charged particles: protons, α particles (^{4}He nuclei), or recoil nuclei of oxygen and carbon that result from interactions of the neutrons with the target tissue. All of these have high LET, and the oxygen effect is less than for low-LET radiation. Fast neutron therapy shows promise in some salivary gland cancers (Douglas et al. 2003).

Boron neutron capture therapy (BNCT) is based on a nuclear reaction which occurs when the stable isotope ^{10}B is irradiated with neutrons, leading to the nuclear reaction (in the notation of Chap. 17)

$$^{10}_{5}\mathrm{B} + ^{1}_{0}n \rightarrow ^{4}_{2}\alpha + ^{7}_{3}\mathrm{Li} \qquad \text{or} \qquad ^{10}_{5}\mathrm{B}(n,\alpha)^{7}_{3}\mathrm{Li}.$$

Both the alpha particle and lithium are heavily ionizing and travel only about one cell diameter. BNCT has been tried since the 1950s; success requires boron-containing drugs that accumulate in the tumor. The field has been reviewed by Barth (2003).

Brachytherapy (brachy means short) involves the implantation of radioactive isotopes in a tumor and will be discussed in Chap. 17.

16.11 Dose Measurement

It is important to measure radiation doses accurately for radiation therapy in order to compare the effectiveness of different treatment protocols and to ensure that the desired protocol is indeed being followed. Accuracies of 2 % are expected. An extensive literature about relating the dose in the measuring instrument to the dose in surrounding tissue exists.[20] Here we describe one of the techniques that is used.

A basic problem in dosimetry is that the measuring instrument has different properties than the medium in which it is immersed. Imagine, for example, that a gas-filled ionization chamber is placed in water. If the radiation field were very large and uniform, one could in principle use an ionization chamber whose dimensions are large compared to the range of secondary electrons, and the interaction of the radiation field with the chamber gas would be the dominant effect. This is not practical. At the other extreme, we imagine an ionization chamber that is so small that it does not alter the radiation field of the water. That is, its dimensions must be small compared to the range of the charged particles created in the water and passing through it.

We saw in Sect. 15.16 that the absorbed dose in a parallel beam of charged particles with particle fluence Φ is (Eq. 15.75)

$$D = \frac{S_e}{\rho}\Phi.$$

Usually the beam consists of particles with different kinetic energies T. Let Φ_T be the energy spectrum:

$$\Phi = \int_0^{T_{\max}} \Phi_T \, dT.$$

Then the dose is the integral of the number of particles with energy T times the mass stopping power for particles of that

[20] See Attix (1986), Chap. 10ff or Khan (2010), Chap. 8.

energy:

$$D = \int_0^{T_{\max}} \Phi_T \frac{S_e}{\rho}\, dT. \tag{16.32}$$

We can define an average mass collision stopping power:

$$\frac{\overline{S}_e}{\rho} = \frac{1}{\Phi} \int_0^{T_{\max}} \Phi_T \frac{S_e}{\rho}\, dT \tag{16.33}$$

so that

$$D = \Phi \frac{\overline{S}_e}{\rho}. \tag{16.34}$$

Let us apply this to the situation where a small detector ("gas") is introduced in a medium ("water") in which we want to know the dose. The charged particle fluence is not altered by the detector because it is small compared to the range of the charged particles. Applying Eq. 16.34 in both media, we obtain

$$\frac{D_w}{D_g} = \frac{(\overline{S}_e/\rho)_w}{(\overline{S}_e/\rho)_g} \equiv (\overline{S}_e/\rho)_g^w. \tag{16.35}$$

This is the *Bragg–Gray relationship* for the absorbed dose in the cavity. It is standard in the literature to denote the dimensionless ratio of the stopping powers in the two media by $(\overline{S}_e/\rho)_g^w$ [or, in some books, $(\overline{S}_c/\rho)_g^w$].

This equation is often used with ionization chambers. The charge created in an ionization chamber of mass m is the charge per ion pair e times the number of ion pairs formed in mass m. The number of ion pairs is the energy deposited, mD_g, divided by the average energy required to produce an ion pair, W:

$$q = e\frac{mD_g}{W}. \tag{16.36}$$

Combining this with Eq. 16.35 gives the dose in the medium in terms of the charge created:

$$D_w = \frac{q}{m}\left(\frac{W}{e}\right)_g (\overline{S}_e/\rho)_g^w. \tag{16.37}$$

The charge q created is usually greater than the charge collected in the ion chamber because of recombination of ions and electrons before collection. The collection efficiency and the chamber mass are deduced from calibration of the chamber. Once the chamber has been calibrated, the factor $(\overline{S}_e/\rho)_g^w$ accounts for placing the chamber in different media.

16.12 The Risk of Radiation

Exposure to radiation may or may not cause a noticeable effect. Effects can include a change, which may not be harmful; damage to cells, which may not necessarily be deleterious to the individual; or harm, which is clinically observable in the subject or possibly a descendant (though current data suggest that genetic changes are rare). It may take years before the harm is observed. The International Commission on Radiological Protection in ICRP (1991) defines the *detriment* to an individual who receives a dose of radiation. It is a rather complex combination of the probability of harm, the severity of the harm, and the time of onset after exposure. It will be discussed more below.

In this section we focus on the increased probability of induction of cancer from an exposure to radiation. We have considerably more information about human exposure to ionizing radiation than we have for any other known or suspected carcinogen (Boice 1996). Several studies at moderate doses show that radiation is a relatively weak carcinogen, though this is not the public view.

We have already seen that the biological effect of radiation depends on the absorbed dose, the LET, the nature of the tissue that is irradiated, and the dose rate. It also depends on the age of the subject. This makes it very difficult to estimate the detriment. Ideally, we would multiply the dose to each organ or target in the body by the probability of a detriment to that target from that kind of radiation now and in the future, and sum over all the organs in the body. This is impossible: we do not know enough. We must simplify the problem while taking some of these differences into account.

16.12.1 Equivalent and Effective Dose

16.12.1.1 Equivalent Dose

Our first simplification assumes that the LET dependence is the same for all target organs. ICRP defines the *radiation weighting factor* W_R for each radiation type R striking the body. It depends on the radiation type and energy and is independent of organ or tissue type. The radiation weighting factor for x-rays is 1. W_R is determined "with guidance" from an earlier quantity, the *relative biological effectiveness* of the radiation (RBE). The weighting factor for each radiation W_R is multiplied by the average dose to the target organ or tissue $D_{R,T}$ and summed to give the *equivalent dose*[21] to

[21] The nomenclature here is quite confusing. ICRP used to define the *dose equivalent*, also denoted by H, as QD, where Q was called the *quality factor* of the radiation. The radiation weighting factor is very similar, and essentially numerically equivalent, to the earlier quality factor, Q. Values of Q recommended by Nuclear Regulatory Commission (NRC) are 1 for photons and electrons, 10 for neutrons of unknown energy and high-energy protons, and 20 for α particles, multiply charged ions, fission fragments, and heavy particles of unknown charge. The ICRP has its own recommendations, that differ slightly for protons and neutrons. See McCollough and Schueler (2000).

Table 16.4 Contribution of some organs to the whole-body radiation detriment. (From ICRP 1991, Table B-20)

Organ (T)	Cancer probability, per Sv	Severe genetic probability, per Sv	Corrected for life lost and nonfatal cancers, per Sv	Tissue weighting factor (W_T)
Bladder	30×10^{-4}		29.4×10^{-4}	0.04
Breast	20×10^{-4}		36.4×10^{-4}	0.05
Stomach	110×10^{-4}		100×10^{-4}	0.14
Gonads		100×10^{-4}	133×10^{-4}	0.18
Total	500×10^{-4}		752×10^{-4}	1.00

the target organ, H_T:

$$H_T = \sum_R W_R D_{R,T}. \tag{16.38}$$

The unit of H_T is the *sievert* (Sv).[22]

16.12.1.2 Detriment and Effective Dose

The detriment is a measure of the harm from an exposure to radiation. It might be a genetic effect (relatively rare) or the development of cancer some years later. If cancer, it might be fatal, shortening life span, or it might cause discomfort and inconvenience but not death. We want to estimate the detriment when a certain equivalent dose has been delivered to some target organs. We assume that the probability of developing cancer in a target organ depends on the dose to that organ and not on the dose to any other part of the body. We also assume that the probabilities are small, so that if several organs have received a radiation dose, the probability of developing cancer is the sum of the probabilities for each organ.

Most of our information about the detriment comes from extensive studies of atomic-bomb survivors, for whom the entire body received a fairly uniform equivalent dose. These survivors have now been followed for almost 70 years. Other studies include patients who have been followed for decades after receiving radiation therapy. ICRP (1991) estimates the radiation detriment using these data and taking into account the probability of a fatal cancer attributable to the radiation, the weighted probability of an attributable nonfatal cancer, the weighted probability of severe hereditary effects, and the relative decrease in lifespan. Table 16.4 shows four of the 14 entries in Table B-20 of ICRP (1991). The details of the various corrections are not shown; the point is to show how each organ contributes to the total detriment.

If a uniform dose is given to the entire body, some organs are more sensitive to the radiation than others. The *effective dose*[23] E is a sum over all irradiated organs:

$$E = \sum_T W_T H_T = \sum_{R,T} W_T W_R D_{R,T}. \tag{16.39}$$

The *tissue weighting factor* W_T is the radiation detriment for organ T from a whole body irradiation as a fraction of the total radiation detriment. By definition, the sum of W_T over all organs equals unity. The last column of Table 16.4 shows the W_T assigned to each target organ in ICRP (1991). See also the review by McCollough and Schueler (2000). Slightly different values of W_T are found in ICRP (2007), Table B.2.

Table 16.5 Major contributions to the effective dose from a typical CT head scan

Organ	W_T	H_T (mSv)	$W_T H_T$ (mSv)
Brain	0.025	36	0.90
Bone marrow (red)	0.12	3	0.36
Thyroid	0.05	5.5	0.28
Bone surface	0.01	14	0.14
All other organs			0.10
Effective dose			1.8

As an example, consider a typical CT head scan, which provides a significant equivalent dose to the brain, bone marrow, thyroid, and bone surface, as shown in Table 16.5.[24] The effective dose is 1.8 mSv. The probability of developing a radiation-induced cancer is $500 \times 10^{-4} \times 1.8 \times 10^{-3} = 9 \times 10^{-5}$. If the whole body were to receive an equivalent dose of 36 mSv, the probability of a radiation-induced cancer would be $500 \times 10^{-4} \times 36 \times 10^{-3} = 1.8 \times 10^{-3}$.

16.12.2 Comparison With Natural Background

One way to express risk is to compare medical doses to the natural background. We are continuously exposed to radiation from natural sources. These include cosmic radiation, which varies with altitude and latitude; rock, sand, brick, and

[22] Both the sievert and the gray are J kg^{-1}. Different names are used to emphasize the fact that they are quite different quantities. One is physical, and the other includes biological effects. An older unit for H is the rem. 100 rem = 1 Sv.

[23] An older, related quantity is the effective dose equivalent, $H_E = W_T Q D$.

[24] Values of H_T were provided by C. McCollough.

Table 16.6 Typical radiation doses from natural sources

Radiation source	Detail	Effective dose rate to target organ (mSv year^{-1})	US population average effective dose rate in 2006 (mSv year^{-1})[a]
Cosmic radiation	New York city	0.30	0.33
	Denver (1.6 km)	0.50	
	La Paz, Bolivia (3.65 km)	1.8	
	Flying at 40,000 ft	7×10^{-3} mSv hr^{-1}	
Terrestrial (radioactive minerals)			0.21
	Over fresh water	0	
	Over sea water	0.2	
	Sandy soil	0.1–0.25	
	Granite	1.3–1.6	
In the body			0.29
Inhalation of radon			2.28
Total			3.11

[a] NCRP Report 160 (2009) Table 1.1

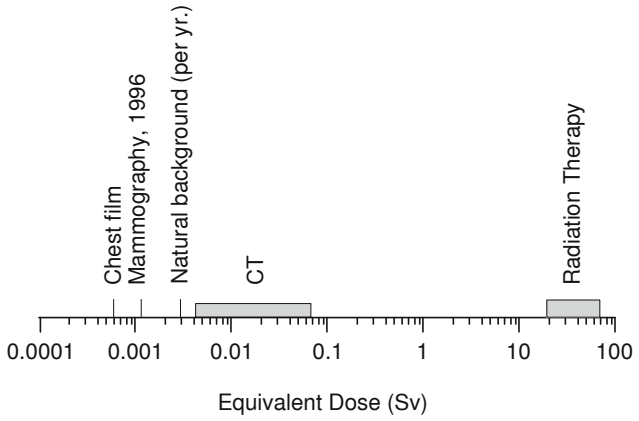


Fig. 16.49 Various doses on a logarithmic scale. Natural background is per year; other doses are per exposure

Table 16.7 Typical radiation equivalent doses for the population of the USA. (From AAPM Report 96 2008, Table 2)

Procedure	Equivalent dose (mSv)
Chest X-ray (Anterior Posterior)	0.1–0.2
Lumbar spine	0.5–1.5
Mammogram	0.3–0.6
Barium enema	3–6
Nuclear medicine cardiac	13–40
Head CT	1–2
Chest CT	5–7
Abdomen CT	5–7
Coronary CT angiography	5–15

concrete containing varying amounts of radioactive minerals; the naturally occurring radionuclides in our bodies such as ^{14}C and ^{40}K; and radioactive progeny from radon gas from the earth.[25] In a typical adult, there are about 4×10^7 radioactive disintegrations per hour from all internal sources. Table 16.6 and Fig. 16.49 summarize the various sources of radiation exposure. The radon entry in Table 16.6 is based on a W_R of 20 for α particles from radon progeny, the value used by NCRP.[26] There is considerable uncertainty in this determination: W_R could be as low as 3, in which case radon would contribute much less to the natural background.

Diagnostic procedures give doses that are in general comparable to the average annual background dose, as can be seen in Table 16.7. Mettler et al. (2008) give a more extensive set of doses. The higher CT doses correspond to pediatric CT; see Fig. 16.52. One can explain to a patient that a chest x-ray is equivalent to about 1 week of natural background, and a mammogram is equivalent to a month or two. A conventional fluoroscopic study of the lower digestive system is equivalent to about a year of natural background. ICRU Report No. 74 discusses patient dosimetry for medical imaging. There is a wide range of doses for a given procedure (Mettler 2008). Patients are having more and more radiologic examinations. Several steps are now taken to reduce the dose due to CT procedures. The cross-section of the body is elliptical; the x-ray tube current can be reduced when the path through the body is shorter. The overall tube current can be reduced as long as the photon-noise-limited resolution is good enough to identify the anatomy of interest.

[25] Radon is chemically inert gas that escapes from the earth. Since it is chemically inert, we breathe it in and out. When it decays in the air (the decay scheme is described in Sect. 17.12), the decay products attach themselves to dust particles in the air. When we breathe these dust particles, some become attached to the lining of the lungs, irradiating adjacent cells as they undergo further decay.

[26] The dose to the lungs from radon progeny is about 1 mGy yr^{-1}. This is multiplied by $W_r = 20$ and $W_T = 0.12$ (lungs) to arrive at an effective dose of 2.4 mSv yr^{-1}.

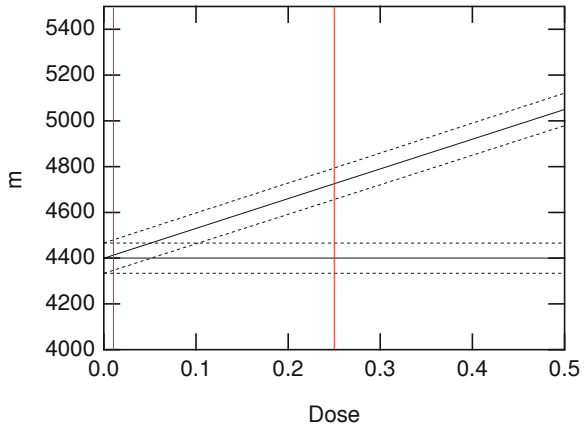


Fig. 16.50 Plot of the average number of cases from a population N for a linear no-threshold and a constant model. The *dashed lines* represent the mean ± 1 standard deviation

16.12.3 Calculating Risk

Assessing the risk of radiation is complicated, since a radiation-induced cancer, for example, may not appear for many years. It is therefore necessary to specify how many years one watches a population after exposure, age at exposure, and current age. One also has to specify whether the risk is of acquiring the disease or of dying from it. Whatever criteria we use, we can define a risk $r(H)$ that depends on the equivalent dose. We then define the *excess absolute risk* as

$$EAR = r(H) - r(0) \tag{16.40}$$

and the *excess relative risk* as

$$ERR = \frac{r(H) - r(0)}{r(0)} = \frac{r(H)}{r(0)} - 1 = \frac{EAR}{r(0)}. \tag{16.41}$$

The units of r and excess absolute risk can vary. The risk might be per person per year, or it might be for a certain number of years or for a lifetime exposure. The excess relative risk has the advantage of being dimensionless. It is frequently reported, even though it can be difficult to understand intuitively. (Plots of EAR vs. dose for breast cancer in Japanese women and women in the USA have nearly the same slope. However, $r(0)$ is smaller for Japanese women, leading to a higher ERR).

Consider a rare disease, and suppose that the probability of acquiring the disease over a lifetime is 2×10^{-3}, while in a population that has received a particular dose of something (which might be radiation, or a chemical, or a particular behavior) the probability is 5×10^{-3}. Then the excess absolute risk is 3×10^{-3}, while the excess relative risk is 150 %. A person hearing that the relative risk has increased by 1.5 times might be unduly alarmed, not realizing that there are only three additional cases in 1000 people.

Statistical fluctuations can make it quite difficult to measure excess risk. Suppose that we want to determine whether r increases linearly with dose. Measurements at lower doses to determine if the response is linear are difficult to make, requiring large numbers of subjects, as the following simplified example shows. Suppose that we have two measurements of the probability of acquiring cancer: one at zero dose, which gives $r(0)$, the "spontaneous" probability due to nonradiation causes, and one at a fairly large dose (say 0.25 Sv, represented by the vertical red line on the right in Fig. 16.50). At some lower dose we want to make a measurement to distinguish between a linear increase of probability with dose and a probability that remains at the "spontaneous" value because we are below some threshold dose for carcinogenicity. The probability $p = r(H)$ of acquiring cancer is small, and the total population N is large. This means that if the experiment could be repeated several times on identical populations, the number of persons acquiring cancer, n, would be Poisson distributed with mean number $m = Np = Nr(H)$ and standard deviation $\sigma = \sqrt{m}$. Figure 16.50 plots m vs. dose for some value of N, with dotted lines to indicate $m \pm \sigma$. A measurement at the lower dose indicated by the vertical red line on the left will not distinguish between the two curves. The only way to reduce the width between the dotted lines at $m \pm \sigma$ would be to use a larger population N.

To give a quantitative but overly simplified example, suppose that $r(H) = e + \alpha H$, with $e = 0.044$ and $\alpha = 0.013\ \mathrm{Sv}^{-1}$. At a dose of 0.25 Sv, $r = 0.047$. For 10^5 persons, the constant curve (expected in the absence of radiation or below threshold) gives $m = 4400 \pm 66$, while the linear curve gives $m = 4730 \pm 69$. The two curves are distinguishable. At a dose of 0.01 Sv, m for the constant curve is still 4400 ± 66, while for the linear case it is 4410 ± 66. It is impossible to distinguish between the linear and constant curves.

16.12.3.1 The Linear No-Threshold Model and Collective Dose

In dealing with radiation to the population at large, or to populations of radiation workers, the policy of the various regulatory agencies has been to adopt the *linear no-threshold* (LNT) model to extrapolate from what is known about the excess risk of cancer at moderately high doses and high dose rates, to low doses, including those below natural background.

If the excess probability of acquiring a particular disease is αH in a population N, the average number of extra persons with the disease is

$$m = \alpha N H. \tag{16.42}$$

The product NH, expressed in person Sv, is called the *collective dose*. It is widely used in radiation protection, but it is meaningful only if the LNT assumption is correct. Small doses to very large populations can give fairly large values of

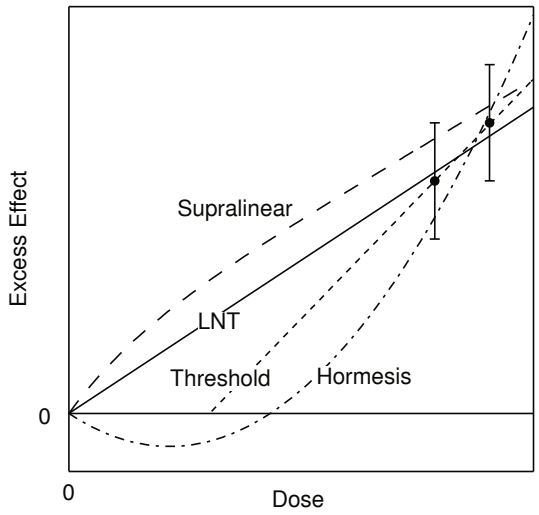


Fig. 16.51 Possible responses to various doses. The two lowest-dose measurements are shown. With zero dose there is no excess effect. The curves are discussed in the text

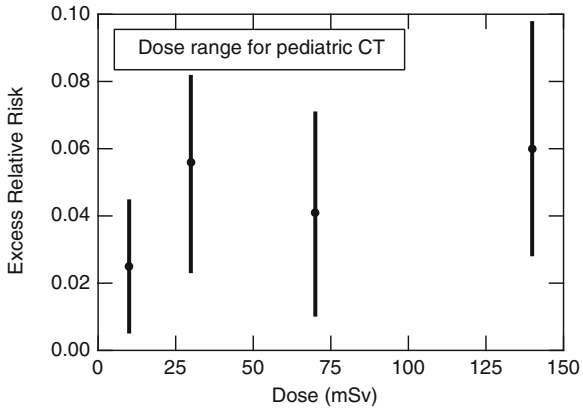


Fig. 16.52 Excess relative risk for atomic bomb survivors who were exposed to a dose of 150 mSv or less and followed for 55 years show a small, but statistically significant increase in cancer incidence. The range of doses from pediatric CT is also shown. (Redrawn from Hall (2002, pp. 225–227). With kind permission of Springer Science and Business Media)

m, assuming that the value of α determined at large doses is valid at small doses.

It has been suggested that there may in some cases be a threshold for radiation-induced damage. If there is a threshold, then the LNT model gives an overestimate. The latest reports of expert panels continue to recommend the LNT model (NCRP Report 136 2001; Upton 2003; BEIR Report VII 2005), but their recommendation is still questioned (Higson 2004; Tubiana et al. 2009). The debate about the LNT model continues (Doss et al. 2014).

To help put the risk in perspective, consider the following example from BEIR (2005), p. 15. Among 100 people, about 42 will be diagnosed with cancer during their lifetime in the absence of any excess radiation. If they had all received a dose of 0.1 Gy (100 mSv for low-LET radiation), there could be one additional cancer in the group.

Even if the LNT model is correct, it can lead to regulatory decisions that are not reasonable. For example, Brooks (2003) cites a study in which the process of cleaning up several Department of Energy sites resulted in more fatal worker accidents than the number of lives that were calculated to have been saved, based on the LNT model.

The Health Physics Society (2010) issued a position statement that radiogenic health effects have not been consistently demonstrated below 100 mSv. They recommend that estimates of risk should be limited to individuals who receive a dose of 50 mSv in one year or 100 mSv in a lifetime.

16.12.3.2 Other Models

Figure 16.51 plots the excess effect vs. dose, showing four possibilities for how some effect might depend on dose. By definition, there is no excess effect when the dose is zero. The two data points represent the lowest doses at which the effect has been measured. The LNT line is a linear interpolation to zero from these points. Lines are also shown for three other possibilities: (1) a *threshold* below which there is no excess effect, (2) a *supralinear* response, which is higher than predicted by the LNT model, and (3) *hormesis*. In hormesis there is a limited range in which the excess effect is negative—a reduction in the effect. Hormesis has been seen in the response of some organisms to chemicals and in some cases to radiation. Two issues of *Critical Reviews in Toxicology*, Vol. 31 No. 4–5 and Vol. 33, No. 3–4, have been devoted to reviews of hormetic effects in all fields.

Some investigators feel that there is evidence for a threshold dose, and that the LNT model overestimates the risk (Kathren 1996; Kondo 1993; Cohen 2002). Mossman (2001) argues against hormesis but agrees that the LNT model has led to "enormous problems in radiation protection practice" and unwarranted fears about radiation.

On the other hand, annual screening CTs (Brenner and Elliston 2004) and CTs to children (Hall 2002) lead to doses that are large enough so that there is a measured excess risk of developing cancer in an individual; no LNT extrapolation to lower doses is needed. This is shown in Fig. 16.52. A pediatric CT study can lead to a dose in the range of 5–100 mSv. This can be compared with data from the extensive study of 35,000 atomic bomb survivors who have been followed for 70 years. There is a small but statistically significant excess risk of developing cancer.

A concerted effort is underway to reduce the dose from a CT procedure to less than 1 mSv (McCollough et al. 2012).

16.12.4 Radon

The question of a hormetic effect or a threshold effect has received a great deal of attention for the case of radon, where remediation at fairly low radon levels is now recommended. Radon is produced naturally in many types of rock. It is a noble gas, but its radioactive decay products can become lodged in the lung. An excess of lung cancer has been well documented in uranium miners, who have been exposed to fairly high radon concentrations as well as high dust levels and tobacco smoke. Radon at lower concentrations seeps from the soil into buildings and contributes a large fraction of the exposure to the general population. Radon concentrations in the air are measured in the number of radioactive decays per second per cubic meter of air. One *becquerel* (Bq) is one decay per second.

Figure 16.53 shows a study by B. L. Cohen (1995) that plots annual age-adjusted lung-cancer mortality rates in 1601 counties in the USA vs. the average radon concentration measured in that county. The radon concentration is expressed as r/r_0, where r_0 is 37 Bq m^{-3} (1.0 pCi l^{-1} in old units, which will be discussed in the next chapter). The upper two panels are for males, and the lower two are for females. The two panels on the right are corrected for the effects of smoking, using the radon-and-smoking model from BEIR Report V (1990). The dashed lines labeled *Theory* are based on the LNT model. The mortality rate falls with increasing radon concentration, though other studies have shown that it rises at radon concentrations higher than shown here.

Epidemiological studies are difficult and can only be suggestive. A number of authors have criticized Cohen's study for dealing with county-wide averages, and Cohen has defended his results.[27] Cohen argues that his data are valid below about 150 Bq m^{-3}. Lubin (1999) compares an LNT fit and Cohen's model to several other radon studies, shown in Fig. 16.54. The error bars are much larger than in Cohen's figure because the populations are smaller. Lubin argues that this is irrelevant because Cohen has systematic errors. Cohen's data points are not inconsistent with those shown by Lubin. Recall from Table 16.6 that the average annual dose from radon is 2.28 mSv. ICRP (2007) uses the conversion that 600 Bq m^{-3} of radon in a dwelling corresponds to an annual dose of 10 mSv per year. This means that r_0 corresponds to 0.6 mSv per year.

Even if the LNT model for radon is correct, some of our remediation efforts are misdirected. Ayotte et al. (1998) used the LNT model to assess the lung cancer risk from radon in Québec. They predicted a total of 109 deaths from lung

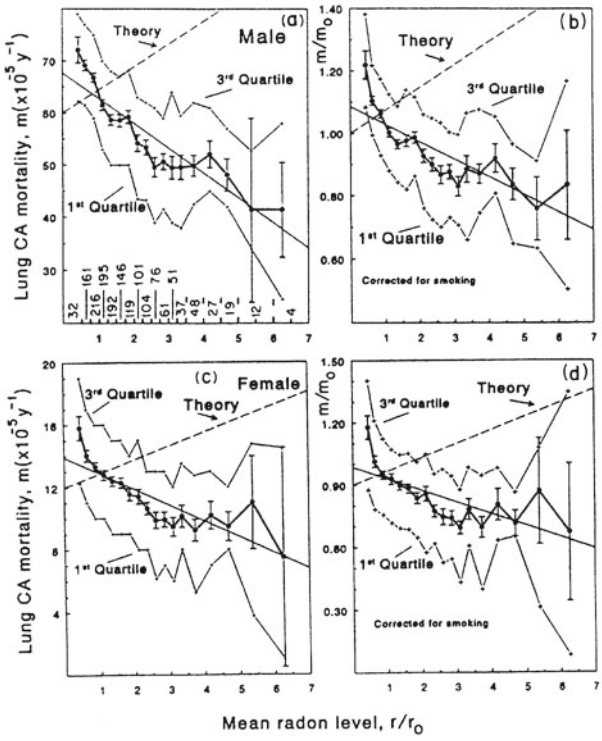


Fig. 16.53 Lung-cancer mortality rates vs. mean radon level in 1601 US counties. Graphs **a** and **b** are for males; (textbfc and **d** are for females. Graphs **b** and **d** have been corrected for smoking levels. Error bars show the standard deviation of the mean. The meaning of radon level is discussed in the text. (From Cohen 1995, pp. 157–174. Used by permission of the Health Physics Society)

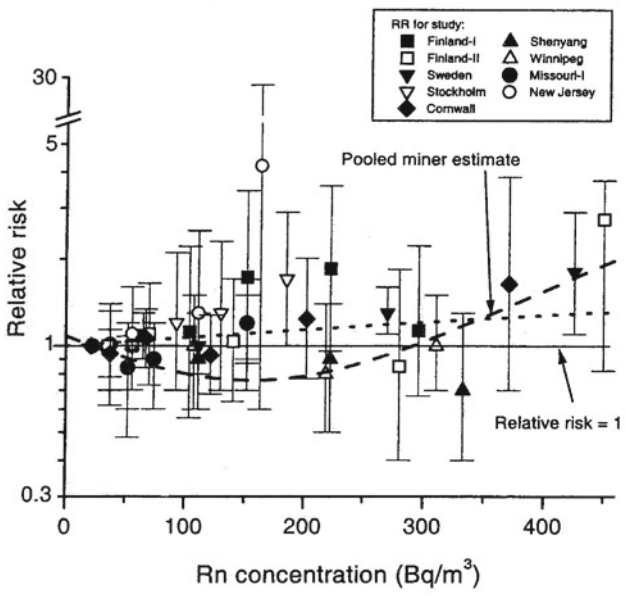


Fig. 16.54 Relative risk (on a log scale) vs. radon concentration. The data points are for several studies, not including Cohen's. The horizontal line shows a relative risk of 1. The dotted line is a linear extrapolation from the miner study. The dashed quadratic line is Cohen's model. (From Lubin 1999, pp. 330–332. Used by permission)

[27] For example, see Lubin (1998a); Cohen (1998); Lubin (1998b); Cohen (1999); Lubin (1999), BEIR VI (1999) and Cohen (2007).

cancer in a population of 60,000. Mitigating radon in all residences with concentrations of 200 Bq m^{-3} or more would reduce this number from 109 to 105. The same number of lives would be saved by reducing smoking by 0.04 %.

Symbols Used in Chapter 16

Symbol	Use	Units	First used page
b	Thickness of slice being scanned	m	479
c	Velocity of light	m s^{-1}	467
e	Electron charge	C	464
f	Fraction of photons that interact or detective quantum efficiency		475
f	General function to be represented		478
g	Incremental signal transfer function	kg C^{-1}	466
h	Planck's constant	J s	462
j	Total angular momentum quantum number		461
k	Minimum signal-to-noise ratio		476
k_B	Boltzmann factor	J K^{-1}	469
l	Orbital angular momentum quantum number		461
m	Mass	kg	464
m_e	Mass of electron	kg	467
m	Mean number		475
n	Principal quantum number		461
n	Number of slices in a scan		479
n	Number of moles of a substance	mol	463
n	Number of fractions		484
p	Probability		485
q	Charge	C	464
r	Distance	m	471
r	Risk or probability	varies	493
r, r_0	Radon concentrations	Bq m^{-3}	495
v	Voltage difference	V	468
w	Width of picture element	m	479
w_i	Mass fraction of ith constituent		471
x	Photon energy/electron rest mass energy		467
x, y, z	Coordinates	m	465
A	Proportionality constant	C m^2 kg^{-1}	475
C	Constant	J^{-1} m^{-2}	463
C	Capacitance	F	468
C_{in}	Exposure contrast		474
C_{out}	Brightness contrast		474
D	Absorbed dose	Gy (J kg^{-1})	464
D_0	Reciprocal of proportionality constant α	Gy	482
$D_{R,T}$	Absorbed dose of radiation type R to target organ T	Gy	491
E	Energy	J or eV	462
E	Effective dose to an organ	Sv (J kg^{-1})	491
F	Projection (integral) of f along some direction		478
G	Radiation chemical yield	mol J^{-1}	463
G	Large signal transfer factor	kg C^{-1}	466
H	Hounsfield CT unit		478
H	Dose equivalent	Sv (J kg^{-1})	491
H_E	Effective dose equivalent	Sv	491
H_T	Equivalent dose	Sv (J kg^{-1})	491
K_c	Collision kerma	J kg^{-1}	464
L	Length of object	m	476
N	Number		464
OD	Optical density		465
P	Probability		484
Q	Quality factor		491
R	Resistance	Ohm (Ω)	468
S	Area	m^2	476
S	Surviving fraction		484
S_e(or S_c)	Collision stopping power	J m^{-1}	490
T	Kinetic energy	J	462
T_0	Initial kinetic energy	J	464
T	Optical transmission		465
T	Temperature	K	469
W	Mean energy expended per ion pair formed	J or eV	464
W_R	Radiation weight factor		491
W_T	Tissue weighting factor		491
X	Exposure	C kg^{-1}	464
Z	Atomic number		461
α	Integral of attenuation		478
α	Dose proportionality constant	Gy^{-1}	482
α	Excess risk proportionality constant	Gy^{-1}	493
β	Squared dose proportionality constant	Gy^{-2}	482
γ	Film contrast		465
μ, μ_{atten}	Attenuation coefficient	m^{-1}	471
μ_{en}	Energy absorption coefficient	m^{-1}	464
ν, ν_0	Frequency	Hz	462
ρ	Density	kg m^{-3}	464
σ	Standard deviation		493
Φ	Particle fluence	m^{-2}	463
Φ_T	Particle fluence per unit energy interval	m^{-2} J^{-1} or m^{-2} eV^{-1}	490
Ψ	Energy fluence	J m^{-2}	463

Problems

Section 16.1

Problem 1. Use Eqs. 15.2 and 16.2 to answer the following questions. Then compare your answers to values given in tables, such as those in the *Handbook of Chemistry and Physics*. What is the minimum energy of electrons striking a copper target that will cause the K x-ray lines to appear? What is the approximate energy of the K_α line? Repeat for iodine, molybdenum, and tungsten.

Problem 2. When tungsten is used for the anode of an x-ray tube, the characteristic tungsten K_α line has a wavelength of 2.1×10^{-11} m. Yet a voltage of 69, 525 V must be applied to the tube before the line appears. Explain the discrepancy in terms of an energy-level diagram for tungsten.

Problem 3. Henry Moseley first assigned atomic numbers to elements by discovering that the square root of the frequency of the K_α photon is linearly related to Z. Solve Eq. 16.2 for Z and show that this is true. Plot Z vs. the square root of the frequency and compare it to data you look up.

Problem 4. Equation 16.3b, indicating the number of photons of energy $h\nu$ produced by bremsstrahlung, is known as *Kramer's law*, and is plotted as crosses in Fig. 16.5 (except for the drop at low energies caused by attenuation that is not included in Kramer's law).

(a) Sketch a plot of $d\Phi/dE$ versus energy ($0 < h\nu < h\nu_0$) using Eq. 16.3b.
(b) Use Eq. 16.3b, integrate $d\Phi/dE$ over energy from 0 to $h\nu_0$, and show that Kramer's law predicts that the number of photons goes to infinity if attenuation is not taken into account.
(c) Integrate Eq. 16.3a from 0 to $h\nu_0$ and show that the energy of the bremsstrahlung radiation predicted by Kramer's law is finite, even if the number of photons is infinite. Explain how this is possible. Derive an expression for the total bremsstrahlung energy.

Problem 5.

(a) The energy fluence spectrum for a thin target $d\Psi/d(h\nu)$ in Fig. 16.3 is constant (call it C') for $h\nu < h\nu_0$ and zero for higher energies. Calculate the photon particle fluence rate $d\Phi/d(h\nu)$ and plot it vs. $h\nu$.
(b) Use the chain rule to express the photon particle fluence rate $d\Phi/d\lambda$ for a thin target as a function of wavelength λ and plot it.
(c) Express Eq. 16.3a, giving the energy fluence rate $d\Psi/d(h\nu)$ for a thick target as a function of photon frequency $h\nu$, as an equation for $d\Psi/d\lambda$ as a function of wavelength λ, and plot it.
(d) Repeat the analysis in part (c) for Eq. 16.3b, giving the photon fluence rate $d\Phi/d\lambda$ for a thick target. Plot it.

Section 16.2

Problem 6. A beam of 0.08-MeV photons passes through a body of thickness L. Assume that the body is all muscle with $\rho = 1.0 \times 10^3$ kg m^{-3}. The energy fluence of the beam is Ψ J m^{-2}.

(a) What is the skin dose where the beam enters the body?
(b) Assume the beam is attenuated by an amount $e^{-\mu L}$ as it passes through the body. Calculate the average dose as a function of the fluence, the body thickness, and μ.
(c) What is the limiting value of the average dose as $\mu L \to 0$?
(d) What is the limiting value of the average dose as $\mu L \to \infty$? Does the result make sense? Is it useful?

Problem 7. The obsolete unit, the roentgen (R), is defined as 2.08×10^9 ion pairs produced in 0.001293 g of dry air. (This is 1 cm^3 of dry air at standard temperature and pressure.) Show that if the average energy required to produce an ion pair in air is 33.7 eV (an old value), then 1 R corresponds to an absorbed dose of 8.69×10^{-3} Gy and that 1 R is equivalent to 2.58×10^{-4} C kg^{-1}.

Problem 8. During the 1930s and 1940s it was popular to have an x-ray fluoroscope unit in shoe stores to show children and their parents that shoes were properly fit. These marvellous units were operated by people who had no concept of radiation safety and aimed a beam of x-rays upward through the feet and right at the reproductive organs of the children! A typical unit had an x-ray tube operating at 50 kVp with a current of 5 mA.

(a) What is the radiation yield for 50-keV electrons on tungsten? How much photon energy is produced with a 5-mA beam in a 30-s exposure?
(b) Assume that the x-rays are radiated uniformly in all directions (this is not a good assumption) and that the x-rays are all at an energy of 30 keV. (This is a very poor assumption.) Use the appropriate values for striated muscle to estimate the dose to the gonads if they are at a distance of 50 cm from the x-ray tube. Your answer will be an overestimate. Actual doses to the feet were typically 0.014–0.16 Gy. Doses to the gonads would be less because of $1/r^2$. Two of the early articles pointing out the danger are Hempelmann (1949) and Williams (1949).

Section 16.3

Problem 9. Rewrite Eq. 16.9 in terms of exponential decay of the viewing light and relate the optical density to the attenuation coefficient and thickness of the emulsion.

Problem 10. Derive the useful rule of thumb $\Delta(\text{OD}) = 0.43\gamma\, \Delta X/X$.

Problem 11. The atomic cross-sections for the materials in a gadolinium oxysulfide screen for 50-keV photons are

Element	Cross-section per atom (m^2)	A
Gd	1.00×10^{-25}	157
S	3.11×10^{-27}	32
O	5.66×10^{-28}	16

(a) What is the cross-section per target molecule of GdO_2S?
(b) How many target molecules per unit area are there in a thickness ρdx of material?
(c) What is the probability that a photon interacts in traversing 1.2 kg m^{-2} of GdO_2S?

Problem 12. The film speed is often defined as the reciprocal of the exposure (in roentgens) required to give an optical density that is 1 greater than the base density. Assume that in Fig. 16.6 a relative exposure of 1 corresponds to 10^{-5} C kg^{-1}. Calculate the film speed.

Problem 13. A dose of 1.74×10^{-4} Gy was estimated for part of the body just in front of an unscreened x-ray film. Suppose that a screen permits the dose to be reduced by a factor of 20. Calculate the skin dose on the other side of the body (the entrance skin dose) assuming 50-keV photons and a body thickness of 0.2 m. Ignore buildup, and assume that only unattenuated photons are detected.

Problem 14. Find an expression for photon fluence per unit absorbed dose in a beam of monoenergetic photons. Then find the photon fluence for 50-keV photons that causes a dose of 10^{-5} Gy in muscle.

Problem 15. A dose of 100 Gy might cause noticeable radiation damage in a sodium iodide crystal. How long would a beam of 100-keV photons have to continuously and uniformly strike a crystal of 1-cm^2 area at the rate of 10^4 photon s^{-1}, in order to produce this absorbed dose? For NaI, $\mu_{en}/\rho = 0.1158$ m^2 kg^{-1}.

Problem 16. Another method to measure the absorbed dose is by calorimetry. Show that if all the energy imparted warms the sample, the temperature rise is 2.39×10^{-4} °C per Gy.

Section 16.4

Problem 17. Plot μ for lead, iodine, and barium from 10 to 200 keV.

Problem 18. Use a spreadsheet to make the following calculations. Consider a photon beam with 100 kVp.

(a) Use Eq. 16.3b to calculate the photon fluence from a thick target at 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 keV.
(b) The specific gravity of aluminum is 2.7. Make a table of the photon fluence at these energies emerging from 2 and 3 mm of aluminum. Compare the features of this table to Fig. 16.15.
(c) Use trapezoidal integration to show that the average photon energy is 44 keV after 2-mm filtration and 47 keV after 3-mm filtration.
(d) Repeat for 120 kVp and show that the average energies after the same filtrations are 52 and 55 keV.

Problem 19. To get a qualitative understanding of Fig. 16.15, assume the photon particle fluence is given by Eq. 16.3b multiplied by a factor $exp(-BL/(h\nu)^3)$, where B is a constant, L is the thickness of the aluminum filtration (in cm) and the $1/(h\nu)^3$ dependence on the photon energy (in keV) arises from the photoelectric cross-section energy dependence, Eq. 15.8.

(a) What are the units of B?
(b) Use some simple numerical method to estimate B from Fig. 16.15. One method might be to calculate the maximum of the photon fluence curve and adjust B so the maximum occurs at the correct photon energy.
(c) For the value of B you found in part (b), plot the three relative photon fluence curves as a function of photon energy, as shown in Fig. 16.15. Normalize the curves so the peak of the 0.1-cm filtration curve is equal to 1.

Problem 20. X-ray beams have a spectrum of photon energies. It would be very laborious to measure the spectrum every time we want to check the quality of the beam. In addition to kVp, one simple measurement that is used to check beam quality (related to the energy spectrum) is the *half-value layer* HVL–the thickness of a specified absorber (often Cu or Al) that reduces the intensity of the beam to one-half.

(a) For a monoenergetic beam, relate HVL to the attenuation coefficient. What is the HVL if the attenuation coefficient is 0.46 mm^{-1}?
(b) For a monoenergetic beam, how does the quarter-value layer QVL relate to HVL?
(c) Suppose a beam has equal numbers of photons at two different energies. The attenuation coefficients at these energies are 0.46 mm^{-1} and 0.6 mm^{-1}. Find the HVL and QVL for this beam. You may need to plot a graph or use a computer algebra program.

Problem 21. The half value layer (HVL) is often used to characterize an x-ray beam. It is the thickness of a specified absorber that attenuates the beam to one-half the original value. Figure 16.41 refers to a beam with a 3.0 mm Cu HVL. What is the value of the attenuation coefficient? What monoenergetic x-ray beam does this correspond to?

Problem 22. Assume an antiscatter grid is made of lead sheets 3-mm long with a spacing between sheets of 0.3 mm. Ignore the thickness of the sheets. If all photons hitting the sheets are absorbed, what is the largest angle from the incident beam direction that a photon can be scattered and still emerge?

Section 16.5

Problem 23. Suppose that two measurements are made: one of the combination of signal and noise, $y = s + n$, and one of just the noise n. One wishes to determine $s = y - n$.

(a) Find $s - \overline{s}$ in terms of y, $\overline{y}$, n, and $\overline{n}$.
(b) Show that if y and n are uncorrelated, $\overline{(s-\overline{s})^2} = \overline{(y-\overline{y})^2} + \overline{(n-\overline{n})^2}$ and state the mathematical condition for being uncorrelated.
(c) If y and n are Poisson distributed, under what conditions is the $\sqrt{2}$ factor of Footnote 13 needed?

Section 16.7

Problem 24. A molybdenum target is used in special x-ray tubes for mammography. The electron energy levels in Mo

are as follows:

K	20 000 eV	L_{I}	2886 eV	M_{I}	505 eV
		L_{II}	2625 eV	M_{II}	410 eV
		L_{III}	2520 eV	M_{III}	392 eV
				M_{IV}	230 eV
				M_{V}	227 eV

What is the energy of the K_α line(s)? The K_β line(s) (defined in Fig. 16.2)?

Problem 25. As a simple model for mammography, consider two different tissues: a mixture of 2/3 fat and 1/3 water, with a composition by weight of 12 % hydrogen, 52 % carbon and 36 % oxygen; and glandular tissue, composed of 11 % hydrogen, 33 % carbon, and 56 % oxygen. The density of the fat and water combination is 940 kg m^{-3}, and the density of glandular tissue is 1020 kg m^{-3}. What is the attenuation in 1 mm of the fat-water combination and in 1 mm of glandular tissue for 50-keV photons? For 30-keV photons?

Section 16.8

Problem 26. It is often said that the number of photons that must be detected in order to measure a difference in fluence with a certain resolution can be calculated from $N = (\Delta\Phi/\Phi)^{-2}$. (For example, if we want to detect a change in Φ of 1 % we would need to count 10^4 photons.) Use Eq. 16.20 to make this statement more quantitative. Discuss the accuracy of the statement.

Problem 27. Spiral CT uses interpolation to calculate the projections at a fixed value of z before reconstruction. This has an effect on the noise. Let σ_0 be the noise standard deviation in the raw projection data and σ be the noise in the interpolated data. The interpolated signal, α, is the weighted sum of two values: $\alpha = w\alpha_1 + (1-w)\alpha_2$.

(a) Show that the variance in α is $\sigma^2 = w^2\sigma_0^2 + (1-w)^2\sigma_0^2$. Plot σ/σ_0 vs. w.

(b) Averaging over a 360° scan involves integrating uniformly over all weights:

$$\sigma^2 = \int_0^1 \left[w^2\sigma_0^2 + (1-w)^2\sigma_0^2\right] dw.$$

Find the ratio σ/σ_0.

Problem 28. An experimental technique to measure cerebral blood perfusion is to have the patient inhale xenon, a noble gas with $Z = 54$, $A = 131$ (Suess et al. 1995). The solubility of xenon is different in red cells than in plasma. The equation used is

$$(\text{arterial enhancement}) = \frac{5.15\theta_{\mathrm{Xe}}}{(\mu/\rho)_w/(\mu/\rho)_{\mathrm{Xe}}} C_{\mathrm{Xe}}(t),$$

where the arterial enhancement is in Hounsfield units, C_{Xe} is the concentration of xenon in the lungs (end tidal volume), and

$$\theta_{\mathrm{Xe}} = (0.011)(\mathrm{Hct}) + 0.10.$$

Hct is the *hematocrit*: the fraction of the blood volume occupied by red cells. Discuss why the equation has this form.

Section 16.9

Problem 29. Use Equations 16.30 and 16.31 to derive an expression for the probability of eradicating a tumor (no surviving tumor cells) as a function of dose for tumors containing different numbers of cells. Verify that your expression reproduces Fig. 16.38.

Section 16.10

Problem 30. Geiger's rule is an approximation to the range-energy relationship:

$$R = AE^p.$$

For protons in water $A = 0.0022$ when R is in cm and E is in MeV. The exponent $p = 1.77$. This is a good approximation for $E < 200$ MeV. Use Geiger's approximation to find dE/dx as a function of R for 100 MeV protons. Make a plot to show the Bragg peak when straggling is ignored.

Problem 31. Assume the stopping power of a particle, $S = -dT/dx$, as a function of kinetic energy, T, is $S = C/T$.

(a) What are the units of C? From Fig. 15.17, estimate for protons the range of kinetic energies over which $S = C/T$ is appropriate.

(b) If the initial kinetic energy at $x = 0$ is T_0, find $T(x)$.

(c) Determine the range R of the particle as a function of C and T_0. For protons in water, estimate C from Fig. 15.26.

(d) Plot $S(x)$ vs. x. Compare the shape of the curve to Fig. 16.47. Does this plot contain a Bragg peak?

(e) Discuss the implications of the shape of $S(x)$ for radiation treatment using this particle.

Section 16.11

Problem 32. Calculate $(\overline{S}_e/\rho)_g^w$ in argon for 0.1-, 1.0- and 10-MeV electrons. The values of S_e/ρ for argon at these energies are 2.918, 1.376, and 1.678 cm^2 g^{-1}.

Problem 33. An ion chamber contains 10 cm^3 of air at standard temperature and pressure. Find q vs. D for 0.5-MeV electrons.

Section 16.12

Problem 34. Suppose that the probability p per year of some event (death, mutations, cancer, etc.) consists of a spontaneous component S and a component proportional to the dose of something else, D: $p = S + AD$. The dose may be radiation, chemicals, sunlight, etc. Investigations of women given mammograms showed that if p is the probability of acquiring breast cancer, $S = 1.91 \times 10^{-3}$ and $A = 4 \times 10^{-4}$ Gy^{-1}. How many women had to be studied to distinguish between $A = 0$ and the value above if $D = 2$ Gy? If $D = 10^{-2}$ Gy?

References

AAPM Report 87 (2005) Diode in vivo dosimetry for patients receiving external beam radiation therapy. American Association of Physicists in Medicine, College Park. Report of Task Group 62 of the Radiation Therapy Committee

AAPM Report 96 (2007) The measurement, reporting and management of radiation dose in CT. American Association of Physicists in Medicine, College Park. Report of Task Group 23 of the Diagnostic Imaging Council CT Committee

Alberts B et al (2002) Molecular biology of the cell, 4th edn. Garland, New York, p 230

Armato SG, van Ginneken B (2008) Anniversary paper: image processing and manipulation through the pages of Medical Physics. Med Phys 35:4488–4500

Attix FH (1986) Introduction to radiological physics and radiation dosimetry. Wiley, New York

Ayotte P, Lévesque B, Gauvin D, McGregor RG, Martel R, Gingras S, Walker WB, Létourneau E G (1998) Indoor exposure to ^{222}Rn: a public health perspective. *Health Phys* 75(3):297–302

Barth RF (2003) A critical assessment of boron neutron capture therapy: an overview. J Neurooncol 62:1–5. (The entire issue of the journal is devoted to a review of BNCT)

BEIR Report V (1990) Health effects of exposure to low levels of ionizing radiation. National Academy Press, Washington, DC. (Committee on the Biological Effects of Ionizing Radiation)

BEIR Report VI (1999) Health effects of exposure to radon. National Academy Press, Washington, DC. (Committee on Health Risks of Exposure to Radon)

BEIR Report VII (2005) Health risks from exposure to low levels of ionizing radiation. National Academy Press, Washington, DC. (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation)

Birch R, Marshall M (1979) Computation of bremsstrahlung X-ray spectra and comparison with spectra measured with a Ge(Li) detector. Phys Med Biol 24:505–517

Boice JD Jr (1996) Risk estimates for radiation exposures. In: Hendee WR, Edwards FM (eds). Health effects of exposure to low-level ionizing radiation. Institute of Physics, Bristol.

Brenner DJ, Elliston CD (2004) Estimated radiation risks potentially associated with full-body CT screening. Radiology 232:735–738

Broad WJ (1980) Riddle of the Nobel debate. Science 207:37–38

Brooks AL (2003) Developing a scientific basis for radiation risk estimates: goal of the DOE low dose research program. Health Phys 85(1):85–93

Brooks RA, DiChiro G (1976a) Principles of computer assisted tomography (CAT) in radiographic and radioisotope imaging. Phys Med Biol 21:689–732

Brooks RA, DiChiro G (1976b) Statistical limitations in x-ray reconstructive tomography. Med Phys 3:237–240

Cohen BL (1995) Test of the linear—no threshold theory of radiation carcinogenesis for inhaled radon decay products. Health Phys 68(2):157–174

Cohen BL (1998) Response to Lubin's proposed explanations of our discrepancy. Health Phys 75(1):18–22

Cohen BL (1999) Response to the Lubin rejoinder. Health Phys 76(4):437–439

Cohen, B. L. (2002). Cancer risk from low-level radiation. [see comment]. AJR Am J Roentgenology 179(5):1137–1143

Cohen BL (2007) The cancer risk from low-level radiation (Chapter 3). In: Tack D, Gevenois PA (eds) Radiation dose from adult and pediatric multidetector computed tomography. Springer, Berlin

Cormack AM (1980) Nobel award address: early two-dimensional reconstruction and recent topics stemming from it. Med Phys 7(4):277–282

Cowen AR, Davies AG, Sivananthan MU (2008a) The design and imaging characteristics of dynamic, solid-state, flat-panel x-ray image detectors for digital fluoroscopy and fluorography. Clin Radiol 63:1073–1085

Cowen AR, Kengyelics SM, Davies AG (2008b) Solid-state, flat-panel, digital radiography detectors and their physical imaging characteristics. Clin Radiol 63:487–498

Delaney TF, Kooy HM (2008) Proton and charged particle radiotherapy. Williams and Wilkins, Philadelphia

DeVita VT (2003) Hodgkin's disease—Clinical trials and travails. N Engl J Med 348(24):2375–2376

DiChiro G, Brooks RA (1979) The 1979 Nobel prize in physiology or medicine. Science 206:1060–1062

Doi K (2006) Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. Phys Med Biol 51:R5–R27

Doss M, Little MP, Orton C (2014) Point/Counterpoint: low-dose radiation is beneficial, not harmful. Med Phys 41:070601. doi:http://dx.doi.org/10.1118/1.4881045

Douglas JG, Koh WJ, Austin-Seymour M, Laramore GE (2003) Treatment of salivary gland neoplasms with fast neutron radiotherapy. Arch Otolaryngol Head Neck Surg 129(9):944–948

Duncan W (1994) An evaluation of the results of neutron therapy trials. Acta Oncol 33(3):299–306. (This issue of the journal is devoted to fast-neutron therapy)

Goitein M (2008) Radiation oncology: a physicist's eye view. Springer, New York

Hall EJ (2000) Radiobiology for the radiologist, 5th edn. Lippincott Williams & Wilkins, Philadelphia

Hall EJ (2002) Helical CT and cancer risk: introduction to session I. Pediatr Radiol 32:225–227

Hall EJ (2003) The bystander effect. Health Phys 85(1):31–35

Hall EJ, Giaccia AJ (2012) Radiobiology for the radiologist, 7th edn. Lippincott Williams & Wilkins, Philadelphia

Harris CC, Hamblen DP, Francis JE Jr (1969) Basic Principles of Scintillation Counting for Medical Investigators. ORNL-2808

Health Physics Society (2010) Radiation risk in perspective. Position statement of the health physics society. http://hps.org/hpspublications/positionstatements.html./risk ps010-2.pdf

Hempelmann LH (1949) Potential dangers in the uncontrolled use of shoe fitting fluoroscopes. N Engl J Med 241:335–336

Hendee WR, Ritenour ER (2002) Medical imaging physics, 4th edn. Wiley-Liss, New York

Higson DJ (2004) The bell tolls for LNT. Health Phys 87(Supplement 2):S47–S50

Hogstrom KR, Almond PR (2006) Review of electron beam therapy physics. Phys Med Biol 51:R455–489
Hounsfield GN (1980) Nobel award address: computed medical imaging. Med Phys 7(4):283–290
Hubbell JH, Seltzer SM (1996) Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients from 1 keV to 20 MeV for elements Z=1 to 92 and 48 additional substances of dosimetric interest. National Institute of Standards and Technology Report NISTIR 5632. http://www.nist.gov/pml/data/xraycoef
Hunt DC, Kirby SS, Rowlands JA (2002) X-ray imaging with amorphous selenium: x-ray to charge conversion gain and avalanche multiplication gain. Med Phys 29(11):2464–2471
ICRP (1991) The 1990 recommendations of the international commission on radiological protection. Ann ICRP 21:1–3
ICRP (2007) The 2007 recommendations of the international commission on radiation protection. Ann ICRP Publication No. 103. Elsevier, New York
ICRU Report 31 (1979) Average energy required to produce an ion pair. International Commission on Radiation Units and Measurements, Bethesda
ICRU Report 33 (1980, Reprinted 1992). *Radiation Quantities and Units*. Bethesda, MD, International Commission on Radiation Units and Measurements.
ICRU Report 37 (1984) Stopping powers for electrons and positrons. International Commission on Radiation Units and Measurements, Bethesda
ICRU Report 39 (1985) Determination of dose equivalents resulting from external radiation sources. International Commission on Radiation Units and Measurements, Bethesda
ICRU Report 41 (1986) Modulation transfer function of screen-film systems. International Commission on Radiation Units and Measurements, Bethesda
ICRU Report 54 (1996) Medical imaging—the assessment of image quality. International Commission on Radiation Units and Measurements, Bethesda
ICRU Report 74 (2005) Patient dosimetry for x rays used in medical imaging. J ICRU 5(2):1–113
Kalender WA (2011) Computed tomography: fundamentals, system technology, image quality and applications, 3rd edn. Publicis, Erlangen
Kassis AI (2004) The amazing world of Auger electrons. Rad Biol 80(11–12):789–803
Kathren RL (1996) Pathway to a paradigm: the linear nonthreshold dose-response model in historical context: the American Academy of Health Physics 1995 Radiology Centennial Hartman Oration. Health Phys 70(5):621–635
Khan FM (1986) Clinical electron beam dosimetry. In Keriakes JG, Elson HR, Born CG (eds) Radiation oncology physics. American Association of Physicists in Medicine, College Park
Khan FM (2003) The physics of radiation therapy, 3rd edn. Philadelphia, Lippincott Williams & Wilkins, p 163, 204, 210, 215
Khan FM (2010) The physics of radiation therapy, 4th edn. Lippincott Williams & Wilkins, Philadelphia
Kondo S (1993) Health effects of low-level radiation. Kinki University Press, Osaka. English translation: Medical Physics, Madison
Körner M, Weber CH, Wirth S, Pfeifer K-J, Reiser MF, Treitl M (2007) Advances in digital radiography: physical principles and system overview. Radiographics 27:675–686
Lubin JH (1998a) On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression. Health Phys 75(1):4–10
Lubin JH (1998b) Rejoinder: Cohen's response to "On the discrepancy between epidemiologic studies in individuals of lung cancer and residential radon and Cohen's ecologic regression". Health Phys 75(1):29–30
Lubin JH (1999) Response to Cohen's comments on the Lubin rejoinder. Health Phys 77(3):330–332
Lutz G (1999) Semiconductor radiation detectors. Springer, New York
Macovski A (1983) Medical imaging systems. Prentice-Hall, Englewood Cliffs
McCollough CH, Schueler BA (2000) Educational treatise: calculation of effective dose. Med Phys 27(5):828–837
McCollough CH, Chen GH, Kalender W, Leng S, Samei E, Taguchi K, Wang G, Yu L, Pettigrew RI (2012) Achieving routine submillisievert CT scanning: report from the summit on management of radiation dose in CT. Radiology 264(2):567–580
Mettler FA, Huda W, Yoshizumi TT, Mahesh M (2008) Effective doses in radiology and diagnostic nuclear medicine: a catalog. Radiology 248:254–263
Metz CE, Doi K (1979) Transfer function analysis of radiographic imaging systems. Phys Med Biol 24(6):1079–1106
Miralbell R, Lomax A, Cella L, Schneider U (2002) Potential reduction of the incidence of radiation-induced second cancers by using proton beams in the treatment of pediatric tumors. Int J Radiat Oncol Biol Phys 54(3):824–829
Mossman KL (2001) Deconstructing radiation hormesis. Health Phys 80(3):263–269
NCRP Report 94 (1987) Exposure of the population in the United States and Canada from natural background radiation. National Council of Radiation Protection and Measurements, Bethesda
NCRP Report 100 (1989) Exposure of the U.S. population from diagnostic medical radiation. National Council of Radiation Protection and Measurements, Bethesda
NCRP Report 136 (2001) Evaluation of the linear-nonthreshold dose-response model for ionizing radiation. National Council of Radiation Protection and Measurements, Bethesda
NCRP Report 160 (2009) Ionizing radiation exposure of the population of the United States. National Council on Radiation Protection and Measurements, Bethesda
Orton C (1997) Fractionation: radiobiological principles and clinical practice (Chapter 21). In: Khan FM, Gerbi B (eds) Treatment planning in radiation oncology. Williams and Wilkins, Baltimore
Pisano ED, Yaffe MJ (2005) Digital mammography. Radiology 234:353–362
Platzman RL (1961) Total ionization in gases by high-energy particles: an appraisal of our understanding. Intl J Appl Radiat Is 10:116–127
Ratliff ST (2009) Resource letter MPRT-1: medical physics in radiation therapy. Am J Phys 77:774–782
Rowlands JA (2002) The physics of computed radiography. Phys Med Biol 47: R123–R126
Schlomka JP, Roessl E, Dorscheid R, Dill S, Martens G, Istel T, Bäumer C, Herrmann C, Steadman R, Zeitler G, Livne A, Proksa R (2008) Experimental feasibility of multi-energy photon-counting K-edge imaging in pre-clinical computed tomography. Phys Med Biol 53:4031–4047
Schulz RJ, Kagan AR (2002) On the role of intensity-modulated radiation therapy in radiation oncology. Med Phys 29(7):1473–1482
Shani G (1991) Radiation dosimetry: instrumentation and methods. CRC, Boca Raton
Shani G (2001) Radiation dosimetry: instrumentation and methods, 2nd edn. CRC, Boca Raton
Smith AR (2009) Vision 20/20: proton therapy. Med Phys 36:556–568
Sobol WT (2002) High frequency x-ray generator basics. Med Phys 29(2):132–144
Steel GG (1996) From targets to genes: a brief history of radiosensitivity. Phys Med Biol 41(2):205–222
Suess C, Polacin A, Kalender WA (1995) Theory of xenon/computed tomography cerebral blood flow methodology. In: Tomonaga M, Tanaka A, Yonas H (eds). Quantitative cerebral blood flow measurements using stable xenon/CT: clinical applications. Futura, Armonk

Suit H, Urie M (1992) Proton beams in radiation therapy. J Natl Cancer Inst 84(3):155–164

Tack D, Gevenois PA (eds) (2007) Radiation dose from adult and pediatric multidetector computed tomography. Springer, Berlin

Tubiana M, Feinendegen LE, Yang C, Kaminski JM (2009) The linear no-threshold relationship is inconsistent with radiation biologic and experimental data. Radiology 251:13–22

Uffmann M, Schaefer-Prokop C (2009) Digital radiography: the balance between image quality and required radiation dose. Eur J Radiol 72:202–208

Upton AC (2003) The state of the art in the 1990's: NCRP report no. 136 on the scientific bases for linearity in the dose-response relationship for ionizing radiation. Health Phys 85(1):15–22

van Eijk CWE (2002) Inorganic scintillators in medical imaging. Phys Med Biol 47:R85–R106

Wagner HN Jr (ed) (1968) Principles of nuclear medicine. Elsevier, p 153, 162

Wagner RF (1977) Toward a unified view of radiological imaging systems. Part II: Noisy images. Med Phys 4(4):279–296

Wagner RF (1983) Low contrast sensitivity of radiologic, CT, nuclear medicine, and ultrasound medical imaging systems. IEEE Trans Med Imaging MI-2(3):105–121

Wagner RF, Weaver KE, Denny EW, Bostrom RG (1974) Toward a unified view of radiological imaging systems. Part I: Noiseless images. Med Phys 1(1):11–24

Williams CR (1949) Radiation exposures from the use of shoe-fitting fluoroscopes. N Engl J Med 241:333–335

William TM, James DC (1989) Radiation Oncology, 6th edn. Mosby, St. Louis

Xu XG, Bednarz B, Paganetti H (2008) A review of dosimetry studies on external-beam radiation treatment with respect to second cancer induction. Phys Med Biol 53:R193–R241

Yu CX, Amies CJ, Svatos M (2008) Planning and delivery of intensity-modulated radiation therapy. Med Phys 35:5233–5241

Nuclear Physics and Nuclear Medicine

17

Each atom contains a nucleus about 100,000 times smaller than the atom. The nuclear charge determines the number of electrons in the neutral atom and hence its chemical properties. The nuclear mass determines the mass of the atom. For a given nuclear charge there can be a number of nuclei with different masses or *isotopes*. If an isotope is unstable, it transforms into another nucleus through *radioactive decay*.

In this chapter we will consider some of the properties of radioactive nuclei and their use for medical imaging and for treatment, primarily of cancer (Ruth 2009; Williams 2008).

Four kinds of radioactivity measurements have proven useful in medicine. The first involves no administration of a radioactive substance to the patient. Rather, a sample from the patient (usually blood) is mixed with a radioactive substance in the laboratory, and the resulting chemical compounds are separated and counted. This is the basis of various *competitive binding assays,* such as those for measuring thyroid hormone and the availability of iron-binding sites. The most common competitive binding technique is called *radioimmunoassay.* A wide range of proteins are measured in this manner.

In the second kind of measurement, radioactive tracers are administered to the patient in a way that allows the volume of a compartment within the body to be measured. Examples of such compartments are total body water, plasma volume, and exchangeable sodium. Time-dependent measurements include red-blood-cell survival and iron and calcium kinetics. One can measure radioactivity from the whole body or from blood or urine samples drawn at different times after administration of the isotope.

For the third class of measurements, a *gamma camera* generates an image of an organ from radioactive decay of a drug that has been administered and taken up by the organ. These images are often made as a function of time.

The fourth class is an extension of these in which tomographic reconstructions of body slices are made. These include *single-photon emission computed tomography* and *positron emission tomography.*

Radioactive isotopes are also used for therapy. The patient is given a radiopharmaceutical that is selectively absorbed by a particular organ (e.g., radioactive iodine for certain thyroid diseases). The isotope emits charged particles that lose their energy within a short distance, thereby giving a high dose to the target organ. Isotopes are also used in self-contained implants for *brachytherapy.*

The first five sections introduce some of the nuclear properties that are important: size, mass, the modes of radioactive decay, and the amount of energy released.

It is important to know the dose to the patient from a nuclear medicine procedure, and a standard technique for calculating it has been developed by the Medical Internal Radiation Dose (MIRD) Committee of the Society of Nuclear Medicine and Molecular Imaging. Section 17.6 shows the steps in making these calculations. Section 17.7 describes some of the pharmacological considerations in selecting a suitable isotope.

The next few sections describe various ways of forming images. Section 17.8 describes the gamma camera, and Sect. 17.9 extends this to single-photon emission tomography. Section 17.10 describes positron emission tomography.

Radiotherapy is described in Sect. 17.11, including both the relatively common brachytherapy and the less common injection of isotopes that target particular organs.

The final section describes the nuclear decay of radon and some of the considerations in calculating the dose and the risk to the general population. It supplements the material that was introduced in Sect. 16.12.

17.1 Nuclear Systematics

An atomic nucleus is composed of Z protons and $N = A - Z$ neutrons. We call Z the *atomic number* and A the *mass number.* Neutrons and protons have very similar properties,

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_17, © Springer International Publishing Switzerland 2015

Table 17.1 Properties of nucleons, the electron, and the neutral hydrogen atom

Property	Neutron	Proton	Electron	H atom
Mass[a]	1.008664916	1.00727647	0.0005485799	1.007825032
Charge[b]	0	$+e$	$-e$	0
Rest energy m_0c^2 (MeV)	939.565	938.272	0.5110	938.783
Half-life	$\approx$ 12 min	Stable	Stable	Stable
Spin	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	...

[a] 1 u is the mass unit. The mass of ^{12}C is 12.0000000 u by definition. 1 u= 1.660539×10^{-27} kg
[b] $e = 1.602177 \times 10^{-19}$ C

as can be seen from Table 17.1. Therefore, they are classed as two different charge states of one particle, the *nucleon*.

Table 17.1 lists the *rest mass* m_0 and the *rest energy*, the rest mass times the square of the speed of light, m_0c^2. One can show using special relativity that the total energy E of an object with rest mass m_0 is related to its speed v and kinetic energy T by

$$E = \frac{m_0c^2}{(1 - v^2/c^2)^{1/2}} = m_0c^2 + T. \qquad (17.1)$$

The energy and mass of both the proton and neutral hydrogen atom are given; the distinction will be important later.

It is customary to specify a nucleus by a symbol such as the following for carbon ($Z = 6$, $N = 6$, $A = 12$):

$$^{A}_{Z}\mathrm{C} \quad \text{or} \quad ^{12}_{6}\mathrm{C}.$$

The mass number used to be written as a superscript on the right; however, this becomes confusing if the ionization state of the atom must also be specified. It is now customary to leave the right side of the symbol for atomic properties. Since the element symbol corresponds to a specific atomic number, Z is often omitted.

Different nuclei of the same element with different numbers of neutrons are called *isotopes*. Other isotopes of carbon are ^{11}C, which has five neutrons, and ^{13}C, which has seven.

The sizes of atoms are roughly constant as one goes through the periodic table, with exceptions as electron shells are filled. On the other hand, the size of nuclei grows steadily through the periodic table (Fig. 17.1). The nuclear radius R and atomic mass number are related by

$$R = R_0A^{1/3}. \qquad (17.2)$$

The precise value for R_0 depends on how the nuclear radius is measured. If it is measured from the charge distribution, then

$$R_0 = 1.07 \times 10^{-15}\,\mathrm{m}. \qquad (17.3)$$

The constancy of atomic size results from two competing effects: as Z increases the outer electrons have a larger value of the principal quantum number n. On the other hand, the greater charge means that Coulomb attraction makes the orbit radius smaller for a given n.

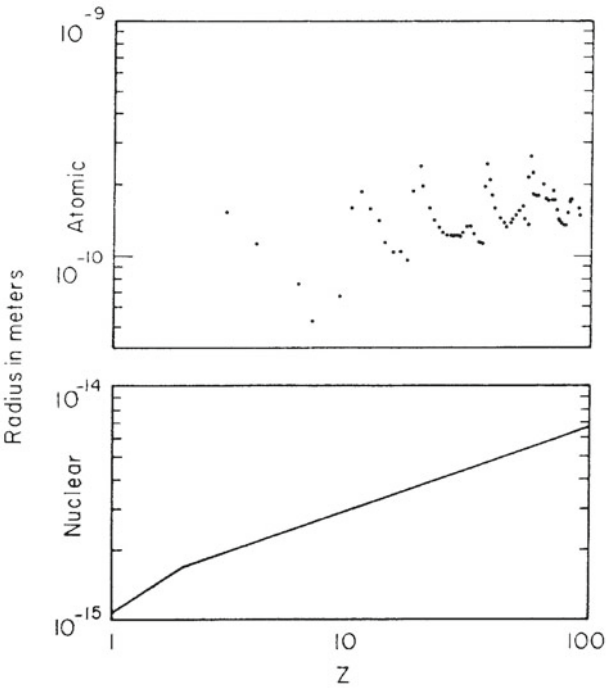


Fig. 17.1 Atomic radius and nuclear radius vs atomic number, showing the relative constancy of the atomic radius and the systematic increase of nuclear radius. Shell effects in atomic radii are quite pronounced; slight shell effects in the nuclear radius are not shown. Atomic data are from Table 7b–3 of *The American Institute of Physics Handbook*. New York, McGraw-Hill, 1957. Nuclear radii are from Eq. 17.2, using the average atomic mass to estimate A from Z

The $A^{1/3}$ dependence in the nuclear case means that the nuclear density is independent of A. To see this, note that the volume of a spherical nucleus is $4\pi R^3/3 = 4\pi R_0^3 A/3$. Since the mass and volume are both proportional to A, the density is constant. This implies that nucleons can get only so close to one another, and that as more are added, the nuclear volume increases. This constant density is the same effect we see in the aggregation of atoms in a crystal or a drop of water.

Scattering experiments measure the force between two nucleons. At large distances, there is no force between two

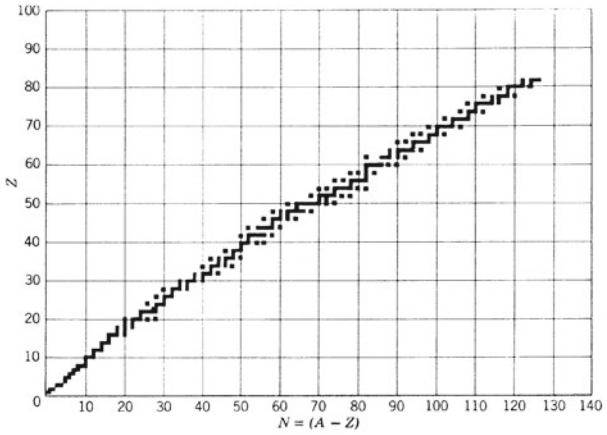


Fig. 17.2 Stable nuclei. Solid squares represent nuclei which are stable and are found in nature. (From Eisberg and Resnick 1985, p. 524. Reprinted with permission of John Wiley & Sons)

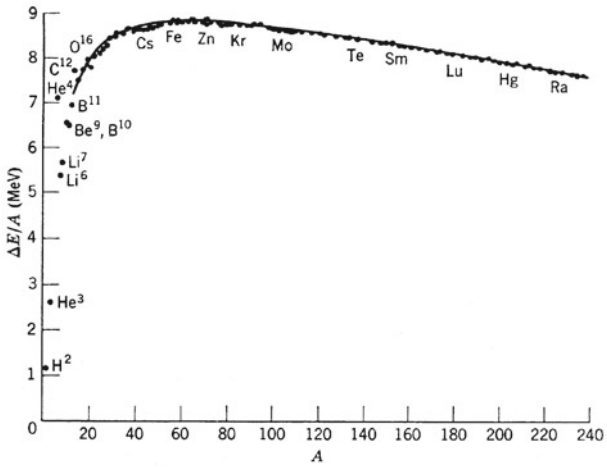


Fig. 17.3 The average binding energy per nucleon for stable nuclei. (From Eisberg and Resnick 1985, p. 524. Reprinted with permission of John Wiley & Sons)

neutrons or between a neutron and a proton. (Between two protons, of course, there is Coulomb repulsion.) As two nucleons are brought close together, a strong attractive force becomes important; at still closer distances, the nuclear force becomes repulsive.

If we look at the nuclei that are stable against radioactive decay and are therefore found in nature, we find that for light elements, $Z = N$. As Z increases, the number of neutrons becomes greater than Z; this can be seen in Fig. 17.2.

Equation 17.1 shows that when an object is at rest, its total energy (which is its internal energy) is related to its rest mass by

$$E = m_0 c^2. \tag{17.4}$$

The measurement of nuclear masses has provided one way to determine nuclear energies. It is necessary to supply energy to a stable nucleus to break it up into its constituent nucleons (or else it would not be stable). The *binding energy* (BE) of the nucleus is the total energy of the constituent nucleons minus the energy of the nucleus:

$$\text{BE} = Zm_p c^2 + (A - Z)m_n c^2 - m_{\text{nucl}} c^2. \tag{17.5}$$

It represents the amount of energy that must be added to the nucleus to separate it into its constituent neutrons and protons.

Suppose we add $Zm_e c^2$ to the first term. Then we have the energy of Z protons plus the energy of Z electrons. Except for the BE of each electron, this is the same as the mass of Z neutral hydrogen atoms, which we call $M_p c^2$. Similarly, we can add the mass of Z electrons to $m_{\text{nucl}} c^2$ and neglect the electron BE to obtain $M_{\text{atom}} c^2$. Capital M represents the mass of a neutral atom, while m stands for the mass of a bare nucleus. For the neutron, $m = M$. In Eq. 17.5, we can add $Zm_e c^2$ to the first term and add $Zm_e c^2$ to the last term, to obtain the BE in terms of the masses of the corresponding neutral particles:

$$\text{BE} = ZM_{\text{p}} c^2 + (A - Z)M_{\text{n}} c^2 - M_{\text{atom}} c^2. \tag{17.6}$$

This is fortunate, because neutral masses (or those for ions carrying one or two charges) are the quantities actually measured in mass spectroscopy.

Masses are measured in *unified mass units* u, defined so that the mass of neutral ^{12}C is exactly 12 u. Carbon is used for the standard because hydrocarbons can be made in combinations to give masses close to any desired mass. The carbon standard replaced one based on the naturally occurring mixture of oxygen isotopes in the early 1960s. (One of the troubles with the earlier standard was that the relative abundance of the various oxygen isotopes varies with time and with location on the earth.) The earlier unit was called the atomic mass unit, amu. One still finds confusion in the literature about which standard is being used, and the carbon standard is sometimes called an amu.

One unified mass unit is related to the kilogram, the joule, and the electron volt by

$$\begin{aligned} 1\,\text{u} &= 1.66054 \times 10^{-27}\ \text{kg}, \\ (1\,\text{u})(c^2) &= \begin{cases} 1.49242 \times 10^{-10}\ \text{J} \\ 931.49\ \text{MeV}. \end{cases} \end{aligned} \tag{17.7}$$

A plot of the BE per nucleon versus mass number, as in Fig. 17.3, shows that the BE per nucleon has a maximum near $A = 60$, and that the average BE (except for light elements) is about 8 MeV per nucleon. For less stable nuclei on either side of the stable line plotted in Fig. 17.2, the BE is less than that for the nuclei shown here.

The maximum near $A = 60$ is what makes both *fission* and *fusion* possible sources of energy. A heavy nucleus with A near 240 can split roughly in half, giving two fission products. Since the nucleons in each of the products are more tightly bound on the average than in the original nucleus, energy is released. This energy difference comes almost entirely from the Z^2 dependence of the Coulomb repulsion of the protons in the nuclei. In fusion, two nuclei of very low A combine to give a nucleus of higher A, for which the BE per nucleon is greater.

17.2 Nuclear Decay: Decay Rate and Half-Life

If a nucleus has more energy than it would if it were in its ground state, it can decay. If the nucleus has sufficient energy, it can emit a proton, neutron, or cluster of nucleons [α particle ($^4_2$He), deuteron ($^2_1$H), etc.]. When a nucleus has enough excitation energy to decay by nuclear emission it usually does so in such an extremely short time that the nuclei could never be introduced in the body after they were produced. An exception is the α decay of a few elements near the upper (high-Z) end of the periodic table. They are found in nature, either because their lifetimes are very long or because they are formed as the result of some other decay process that has a long lifetime.

If a nucleus has just a small amount of excess energy, it emits a γ *ray*, a photon analogous to the x-ray or visible photons emitted by an excited atom. Another process that can occur is the emission of a positive or negative electron, with the conversion of a proton to a neutron, or vice versa. This is called β *decay*. γ and β decay will be described in detail in the next two sections.

Each excited nucleus will decay or undergo a *nuclear transformation*. There can be several *transitions* associated with each transformation. For example, there might be a cascade of two or more successive gamma rays (γ_1 and γ_2 of Fig. 17.4), or competing pathways (branching) (γ_1, γ_3, β_2^-, γ_1 of Fig. 17.4).

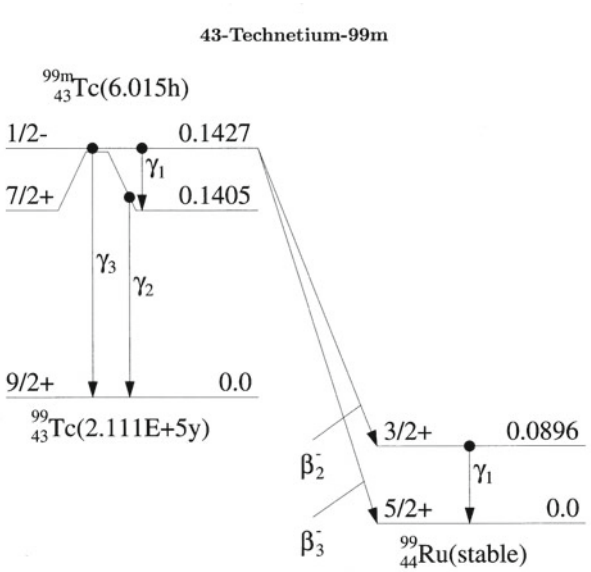
An excited nucleus has a probability λdt of transforming in time dt. When there are N nuclei present, the average number decaying in time dt is[1]

$$-dN = N\lambda\, dt.$$

This leads to the familiar exponential decay of Chap. 2:

$$N = N_0 e^{-\lambda t}.$$

[1] The decay constant is called λ in this chapter to conform to the usage in nuclear medicine.



$T_{1/2} = 6.015$ h — Decay Modes: β^- IT

Radiation	$Y(i)$ (Bq s)$^{-1}$	$E(i)$ MeV	$\Delta(i)$ Gy kg (Bq s)$^{-1}$
ce-M, γ-ray 1	8.62E−01	1.748E−03†	2.42E−16
ce-N$^+$, γ-ray 1	1.30E−01	2.173E−03†	4.52E−17
γ-ray 2	8.91E−01	1.405E−01	2.00E−14
ce-K, γ-ray 2	8.92E−02	1.195E−01	1.71E−15
ce-L$_1$, γ-ray 2	9.89E−03	1.375E−01	2.18E−16
ce-L$_2$, γ-ray 2	6.46E−04	1.377E−01	1.42E−17
ce-L$_3$, γ-ray 2	3.37E−04	1.378E−01	7.45E−18
ce-M, γ-ray 2	1.99E−03	1.401E−01†	4.47E−17
ce-N$^+$, γ-ray 2	3.80E−04	1.405E−01†	8.56E−18
ce-K, γ-ray 3	5.50E−03	1.216E−01	1.07E−16
ce-L$_1$, γ-ray 3	9.48E−04	1.396E−01	2.11E−17
ce-L$_2$, γ-ray 3	1.98E−04	1.398E−01	4.44E−18
ce-L$_3$, γ-ray 3	6.08E−04	1.400E−01	1.36E−17
ce-M, γ-ray 3	3.48E−04	1.422E−01†	7.93E−18
K - L$_2$ x-ray	2.14E−02	1.821E−02	6.23E−17
K - L$_3$ x-ray	4.06E−02	1.833E−02	1.19E−16
K - M$_3$ x-ray	6.53E−03	2.059E−02	2.16E−17
Auger KLL	1.48E−02	1.542E−02†	3.65E−17
Auger KLX	5.59E−03	1.782E−02†	1.60E−17
Auger LMM	9.03E−02	2.054E−03†	2.98E−17
Auger LMX	1.41E−02	2.333E−03†	5.26E−18
CK MMX	7.09E−01	1.142E−04†	1.30E−17
Auger MNN	1.08E+00	2.061E−04†	3.57E−17
CK NNX	2.47E+00	2.961E−05†	1.17E−17

Fig. 17.4 Energy levels and decay data for the isotope ^{99m}Tc. The various features are discussed in the text. (These results were originally published in Eckerman and Endo 2008, p. 232. © by the Society of Nuclear Medicine and Molecular Imaging, Inc.)

The *activity*, $A(t)$, is the number of decays per second:

$$A(t) = \left|\frac{dN}{dt}\right| = \lambda N.$$

The activity is measured in nuclear transformations per second or becquerel (Bq). The total number of transformations or *cumulated activity* is measured in bequerel seconds (Bq s).

The half-life $T_{1/2}$ is related to λ by Eq. 2.10:

$$T_{1/2} = \frac{0.693}{\lambda}. \tag{17.8}$$

17.3 Gamma Decay and Internal Conversion

When a nucleus is in an excited state, it can lose energy by photon emission. The energy levels of the nucleus are characterized by certain quantum numbers, and γ emission is subject to selection rules analogous to those for x-ray emission by atoms. Half-lives for γ emission range from 10^{-20} to 10^{+8} s.

Figure 17.4 shows an energy level diagram for $^{99}_{43}$Tc (technetium), an isotope widely used in nuclear medicine, along with some tabular material that we will need as we progress through this chapter. First, look at the energy level diagram. There are two important levels to consider in $^{99}_{43}$Tc. The ground state is not stable but decays by β^- decay, considered in Sect. 17.5. However, its decay rate is so small (half-life of 2.111×10^5 years) that we can ignore its decay. There is a level at an excitation of 0.1427 MeV above the ground state that has a half-life of 6.015 h for γ decay. This is an unusually long half-life; we call it a *metastable state* and denote it by ^{99m}Tc. We see that there are two modes of gamma decay from this state. The first is the emission of a 0.0022-MeV γ ray (γ_1) followed immediately by a 0.1405-MeV γ_2. The other, less common possibility, is the emission of γ_3 of energy 0.1427 MeV. The ^{99m}Tc can also undergo beta decay, considered in Sect. 17.5.

Whenever a nucleus loses energy by γ decay, there is a competing process called *internal conversion*. The energy to be lost in the transition, E_γ, is transferred directly to a bound electron, which is then ejected with a kinetic energy

$$T = E_\gamma - B, \tag{17.9}$$

where B is the binding energy of the electron.

We now turn to the tabular part of Fig. 17.4. Each line describes a unique transition associated with the nuclear transformation of the ^{99m}Tc.

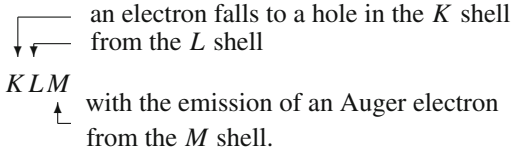
The *mean number per disintegration* $Y(i)$ in the table is the mean number of times that the indicated transition between energy levels takes place per nuclear transformation. (Think Y for yield.)

The first two lines in the table show that the only transitions associated with γ_1 are internal conversion of either an M-shell or N-shell electron (*ce* stands for conversion electron). Gamma-ray 2 is emitted 0.891 times per nuclear transformation, with internal conversion occurring 0.102 times per transformation.

17.4 Atomic Deexcitation

Once internal conversion has created a hole in the electronic structure of the atom, characteristic x rays and Auger and Coster-Kronig (CK) electrons are emitted as described in Sect. 15.9.

Characteristic x ray transitions have labels like K-L_2 x ray. The labels for Auger and CK electrons show this information:



The Auger cascade means that several of these electrons are emitted per transition. If a radionuclide is in a compound that is bound to DNA, the effect of several electrons released in the same place is to cause as much damage per unit dose as high-LET radiation. Linear energy transfer was defined in Chap. 15. A series of reports on this effect have been released by the American Association of Physicists in Medicine (AAPM) (Sastry 1992; Howell 1992; Humm et al. 1994).

Many electrons (up to 25) can be emitted for one nuclear transformation, depending on the decay scheme (Howell 1992). The electron energies vary from a few eV to a few tens of keV. Corresponding electron ranges are from less than 1 nm to 15 μm. The diameter of the DNA double helix is about 2 nm. A number of experiments (reviewed in the AAPM reports, and also in Kassis (2011)) show that when the radioactive substance is in the cytoplasm the cell damage is like that for low-LET radiation in Fig. 15.32 with relative biological effectiveness (RBE) = 1. When it is bound to the DNA, survival curves are much steeper, as with the α particles in Fig. 15.32 (RBE $\approx$ 8). When it is in the nucleus but not bound to DNA the RBE is about 4. The fraction of the Auger emitter that binds to the DNA depends on the chemical agent to which the nuclide is attached. There is also a significant bystander effect (Kassis 2011).

17.5 Beta Decay and Electron Capture

Nuclei that are not on the line of stability in Fig. 17.2 have greater internal energy and are susceptible to some kind of decay. They can lose energy by γ emission. In addition, nuclei above the line of stability have too many protons relative to the number of neutrons; nuclei below the line have relatively too many neutrons.

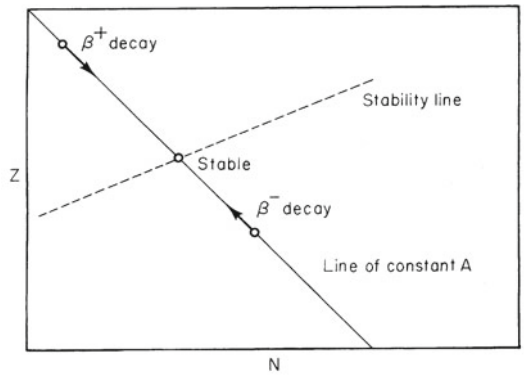


Fig. 17.5 β^- decay and β^+ decay do not change A. They do change N and Z to bring the nucleus closer to the stability line

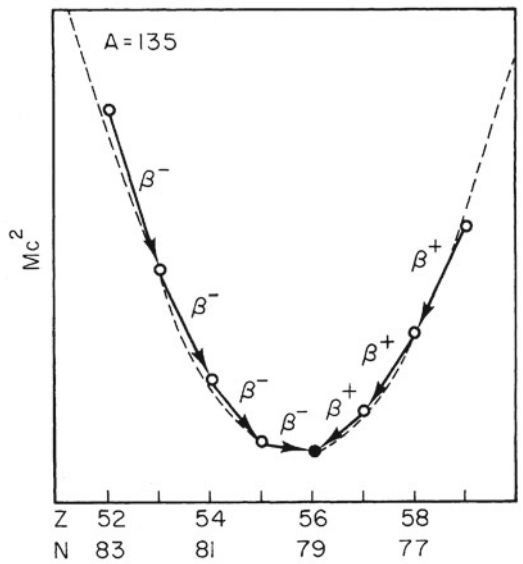


Fig. 17.6 Energy of nuclei as a function of Z for an odd value of A ($A = 135$). The only stable nucleus is $^{135}_{56}$Ba; nuclei of lower Z undergo β^- emission; those of higher Z undergo β^+ emission or electron capture

Two modes of decay allow a nucleus to approach the stable line. In *beta* (β^- or electron) *decay*, a neutron is converted into a proton. This keeps A constant, lowering N by one and raising Z by one. In *positron* (β^+) *decay*, a proton is converted into a neutron. Again A remains unchanged, Z decreases and N increases by 1. We find β^+ decay for nuclei above the line of stability and β^- decay for nuclei below the line. Figure 17.5 shows a portion of the line of stability, a line of constant A ($Z = A - N$), and the regions for β^+ and β^- decay.

We can plot the energy of the neutral atom for different nuclei along the line of constant A. Since there are one or two stable nuclei, there is some value of Z and N for which the

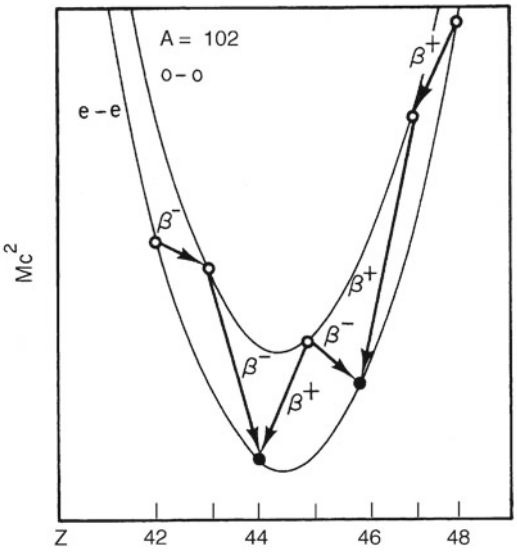


Fig. 17.7 Energy of even-A nuclei ($A = 102$) as a function of Z. Nuclei with an odd number of protons and neutrons have higher energies than those with an even number of each. This makes it possible for the same nucleus to decay by either β^- or β^+ emission

energy is a minimum. The energy increases in either direction from this minimum. The first approximation to a curve with a minimum is a parabola, as shown in Fig. 17.6 for a nucleus of odd A.[2] When Z is too small, a neutron is converted to a proton by β^- decay. If Z is too large, a proton changes to a neutron by β^+ decay or electron capture (to be described below).

When A is odd, there are an even number of protons and an odd number of neutrons (even–odd) or vice versa (odd–even). When we plot the energies of even-A nuclei, we find that the masses lie on two different parabolas (Fig. 17.7). The one for which both Z and N are odd (odd–odd) has greater energy than the parabola for which both are even. The reason is that nucleons have lower energy when they are paired with one another in such a way that their spins are antiparallel. In the even–even case, the neutrons and the protons are all paired off and have this lower energy; in the odd–odd case there are both an unpaired proton and an unpaired neutron, and the energy is higher. As we change Z by one, we jump back and forth between the odd–odd and the even–even parabolas. For odd-A nuclei, either the neutrons are paired and one proton is not, or vice versa. There is always one unpaired nucleon as Z changes, so there is only one parabola.

[2] This parabola and the general behavior of the BE with Z and A can be explained remarkably well by the semiempirical mass formula (Evans 1955, Chap. 11; Eisberg and Resnick 1985, p. 528).

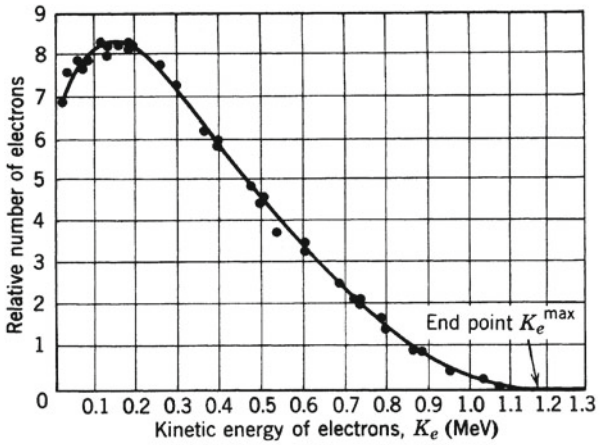


Fig. 17.8 A typical spectrum of β particles. In this case it is for the β decay of $^{210}_{83}$Bi. (From Eisberg and Resnick 1985, p. 566. Copyright ©1985 John Wiley & Sons. Reproduced by permission of John Wiley & Sons)

The existence of the two parabolas means that there are usually (but not always) two stable nuclei with an odd–odd nucleus between them that can decay by either β^- or β^+ emission.

The emission of a β^- particle is accompanied by the emission of a *neutrino* (strictly speaking, an antineutrino):

$$^A_ZX \rightarrow {}^A_{Z+1}Y + \beta^- + \overline{\nu}. \tag{17.10}$$

The neutrino has no charge and no rest mass,[3] so that like a photon, it travels with velocity c and its energy and momentum are related by $E = pc$. Neutrinos hardly interact with matter at all, so they are quite difficult to detect. Nevertheless they have been detected through certain specific nuclear reactions that take place on the rare occasions when a neutrino does interact with a nucleus. A particle that seemed originally to be an invention to conserve energy and angular momentum now has a strong experimental basis.

Suppose that β decay consisted of the ejection of only a β particle. If the original nucleus was at rest,[4] then the final nucleus would recoil in the direction opposite the β particle to conserve momentum; the ratio of its velocity to that of the β particle would be given by their mass ratio. The recoil nucleus and the β particle would each have a definite fraction of the total energy available from the decay, and the β particles would all have the same energy. However, the observed β-particle energy spectrum is not a line spectrum but a continuum ranging from zero to the expected energy, as shown in Fig. 17.8. The missing energy is carried by the neutrino. The different energies correspond to different angles of emission of the neutrino relative to the direction of the β particle. This kind of spectrum is characteristic of three bodies emerging from the reaction.

The total kinetic energy for the three emerging particles is

$$E_{\text{decay}} = m_{Z,A}c^2 - m_{Z+1,A}c^2 - m_ec^2.$$

If we add and subtract Zm_ec^2, the result is unchanged:

$$\begin{aligned} E_{\text{decay}} &= (m_{Z,A}c^2 + Zm_ec^2) - (m_{Z+1,A}c^2 \\ &\quad + Zm_ec^2 + m_ec^2) \\ &= M_{Z,A}c^2 - M_{Z+1,A}c^2. \end{aligned} \tag{17.11}$$

The energy released in the decay is given by the difference in rest energies of the initial and final neutral atoms. This energy is shared in different amounts by the three particles; it is shared mainly by the neutrino and electron, since the nucleus is so massive and its kinetic energy is $p^2/2m$. The maximum or end-point energy of the β spectrum in Fig. 17.8 corresponds to E_{decay}.

Figure 17.9 shows data for the decay of ^{24}Na, an isotope that has been used in nuclear medicine. The transition labeled β_3^- is overwhelmingly more common than β_4^-. The β_3^- emission is followed by two γ rays. The average energy of the β_3^- particle is 0.554 MeV, about 40 % of the end-point energy, 1.392 MeV.

Emission of a positron converts a proton into a neutron, and Z decreases by one. A neutrino is also emitted:

$$^A_ZX \rightarrow {}^A_{Z-1}Y + \beta^+ + \nu. \tag{17.12}$$

The decay energy is again given by

$$E_{\text{decay}} = m_{Z,A}c^2 - m_{Z-1,A}c^2 - m_ec^2.$$

However, this time, when we add Zm_ec^2 to the first term and subtract $(Z-1)m_ec^2$ from the second term to convert these to atomic masses, the electron masses do not cancel. Instead, we get

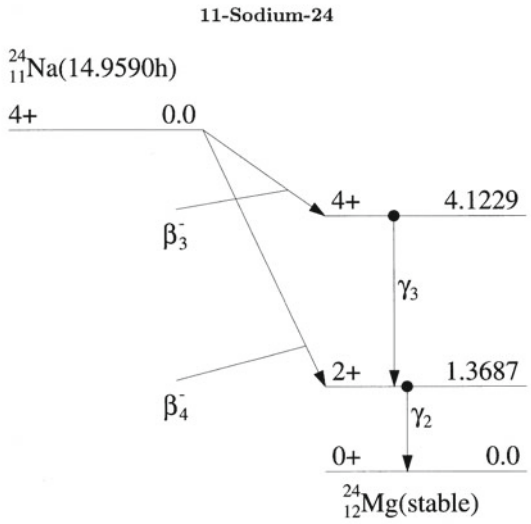
$$E_{\text{decay}} = M_{Z,A}c^2 - M_{Z-1,A}c^2 - 2m_ec^2. \tag{17.13}$$

Positron emission will not occur unless the initial neutral atomic mass exceeds the final neutral atomic mass by at least $2m_ec^2$.

Figure 17.10 shows the decay scheme for ^{18}F, which decays primarily by positron emission, with an average positron energy of 0.2498 MeV. The decay line in the energy level diagram is labelled EC$_1$, β_1^+. EC stands for *electron capture*, a process that competes with beta decay. Some of the inner electrons of the atom are close enough to the nucleus (quantum mechanically, the electron wave functions overlap the nucleus enough) so that the electron is captured by the nucleus, and a neutrino is emitted. In terms of nuclear

[3] Recent measurements indicate that the neutrino does have a rest mass, but it is too small to affect our argument.

[4] Its thermal energy of about $\frac{1}{40}$ eV is negligible compared to the energy released in decay.

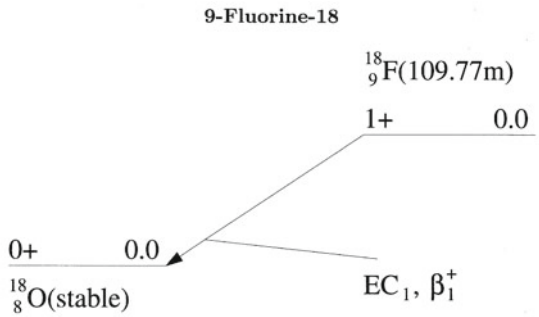


$T_{1/2}$ = 14.959 h — Decay Mode: β^-

Radiation	$Y(i)$ (Bq s)$^{-1}$	$E(i)$ MeV	$\Delta(i)$ Gy kg (Bq s)$^{-1}$
β^- 3	9.99E−01	5.540E−01†	8.88E−14
γ-ray 2	1.00E+00	1.369E+00	2.20E−13
γ-ray 3	9.99E−01	2.754E+00	4.41E−13
Listed X, γ and $\gamma\pm$ radiation			6.60E−13
Neglected X, γ and $\gamma\pm$ radiation‡			3.33E−16
Listed β and ce			8.88E−14
Neglected β and ce‡			2.11E−17
Listed radiations			7.48E−13
Neglected radiations‡			3.54E−16

†Average energy.
‡Each omitted transition contributes < 0.10% to $\Sigma\Delta(i)$.
Magnesium-24 daughter, yield of 1.0, is stable.

Fig. 17.9 Energy levels and data for the β decay of ^{24}Na. (These results were originally published in Eckerman and Endo 2008, p. 56. © by the Society of Nuclear Medicine and Molecular Imaging, Inc.)



$T_{1/2}$ = 109.77 m — Decay Modes: EC β^+

Radiation	$Y(i)$ (Bq s)$^{-1}$	$E(i)$ MeV	$\Delta(i)$ Gy kg (Bq s)$^{-1}$
β^+ 1	9.67E−01	2.498E−01†	3.88E−14
$\gamma\pm$	1.93E+00	5.110E−01	1.58E−13
Listed X, γ and $\gamma\pm$ radiation			1.58E−13
Listed β and ce			3.88E−14
Listed radiations			1.97E−13

†Average energy.
Oxygen-18 daughter, yield of 1.0, is stable.

Fig. 17.10 Energy levels and data for the β^+ decay of ^{18}F. (These results were originally published in Eckerman and Endo 2008, p. 52. © by the Society of Nuclear Medicine and Molecular Imaging, Inc.)

masses, an electron rest energy is added to the parent nucleus (we ignore its kinetic energy):

$$E_{\text{e.c.}} = m_e c^2 + m_{Z,A} c^2 - m_{Z-1,A} c^2.$$

If we add and subtract $(Z-1)m_e c^2$, we have

$$E_{\text{e.c.}} = M_{Z,A} c^2 - M_{Z-1,A} c^2. \tag{17.14}$$

A K electron is usually captured. The energy from the nuclear transition is given to a neutrino. No electron or positron emerges from the nucleus, but there are K x rays and Auger electrons, as there are any time a vacancy in the K shell occurs, and these contribute to the radiation dose. Electron capture and positron emission can both occur in proton-rich isotopes. In the case of ^{18}F (and many other low-atomic-number isotopes) the decay is mainly by positron emission, with relatively little electron capture. In many heavier nuclei, electron capture dominates over positron emission. For instance, ^{125}I decays by electron capture, and the resulting cascade of Auger electrons makes a significant contribution to the dose.

The second entry, labeled $\gamma_\pm$, stands for *annihilation radiation*. Once a positron has been emitted, it slows down like any other charged particle. At some point it combines with an electron (since the positron and electron constitute a particle–antiparticle pair), and all of the rest energy of both particles goes into two photons.[5] The energy conservation equation is

$$2m_e c^2 = 2h\nu. \tag{17.15}$$

For each original positron emitted, two photons are produced, each of energy $m_e c^2 = 0.511$ MeV. Note that Y_i for the annihilation gamma rays is twice the value for positron emission.

17.6 Calculating the Absorbed Dose from Radioactive Nuclei within the Body: the MIRD Method

When a radiopharmaceutical is given to a patient for either diagnosis or therapy, the nuclei end up in different organs in varying amounts; for example, ^{99m}Tc-labeled albumin

[5] Three photons are occasionally emitted.

microspheres injected intravenously lodge in the lungs. The problem is to calculate the whole-body absorbed dose, the dose to the lungs, and the dose to other organs.

The dose calculation in this chapter follows the technique and notation recommended by the MIRD Committee of the Society of Nuclear Medicine and Molecular Imaging (Loevinger et al. 1988; ICRU 2002; Stabin et al. 2005; Stabin 2008; Bolch et al. 2009). It is carried out in the following way:

1. Calculate the total number of nuclear transformations or disintegrations in organ h. It is called the *cumulated activity* $\tilde{A}_h$ or N_h.
2. Calculate the mean energy emitted per unit cumulated activity for each type of photon or particle emitted.
   a) If the radioactive nucleus can emit several types of particles or photons per transformation, call Y_i the mean number of particles or photons of type i (transitions) emitted per transformation. These include γ rays, electrons, x rays and Auger electrons. The data are also available in electronic form (Eckerman et al. 1994; Stabin and da Luz 2002; RADAR (the Radiation Group Assessment Resource), www.doseinfo-radar.com; and the National Nuclear Data Center, www.nndc.bnl.gov/mird/).
   b) For each transition i determine E_i, the *mean energy per transition*.
   c) Calculate or look up $\Delta_i = Y_i E_i$, *the mean energy emitted per unit cumulated activity*, for each type of particle or photon emitted. (In earlier MIRD literature, this was called the *equilibrium absorbed dose constant*.)
3. Calculate $\phi_i(r_k \leftarrow r_h)$, the fraction of the radiation of type i emitted in source region r_h that is absorbed in target region r_k, and divide by the mass of the target region to get the *specific absorbed fraction*

$$\Phi_i(r_k \leftarrow r_h) = \frac{\phi_i(r_k \leftarrow r_h)}{m_k}.$$

   (Φ has the units of inverse mass.)
4. The *mean absorbed dose* in organ k due to activity in organ h, $\overline{D}$ (in J kg^{-1} or Gy) is

$$\overline{D}(r_k \leftarrow r_h) = \tilde{A}_h \sum_i \Delta_i \Phi_i(r_k \leftarrow r_h). \tag{17.16}$$

5. If several organs are radioactive, a sum must be taken over each organ:

$$\overline{D}(r_k) = \sum_h \tilde{A}_h \sum_i \Delta_i \Phi_i(r_k \leftarrow r_h). \tag{17.17}$$

Some tables (Snyder et al. 1976, 1978) give values of Φ_i for photons of various energies. It is necessary to multiply by Δ_i and sum for the isotope of interest. The sum is called the *mean absorbed dose per unit cumulated activity*:

$$S(r_k \leftarrow r_h) = \sum_i \Delta_i \Phi_i(r_k \leftarrow r_h), \tag{17.18}$$

$$\overline{D}(r_k \leftarrow r_h) = \tilde{A}_h S(r_k \leftarrow r_h), \tag{17.19}$$

$$\overline{D}(r_k) = \sum_h \tilde{A}_h S(r_k \leftarrow r_h). \tag{17.20}$$

These sums must be repeated over and over again for common radionuclides. A table of S for many common radionuclides is available (Snyder et al. 1976). The tables cannot be summed over h because the values of $\tilde{A}_h$ depend on how the isotope is administered. A computer program OLINDA/EXM is most commonly used for these calculations (Stabin et al. 2005). These authors call S the *dose factor*.

To discuss units, imagine there is only one type of radiation. In SI units the dose is simply

$$D\ (\text{Gy}) = \left[\tilde{A}\ (\text{dimensionless})\right][\Delta_i\ (\text{J})]\left[\Phi_i\ (\text{kg}^{-1})\right]. \tag{17.21a}$$

In day-to-day calculations, it is often easier to use mixed units and write

$$D\ (\text{Gy}) = k\left(\frac{\text{Gy kg}}{\text{MBq s MeV}}\right)\left[\tilde{A}\ (\text{MBq s})\right] \times [\Delta_i\ (\text{MeV})]\left[\Phi_i\ (\text{kg}^{-1})\right]. \tag{17.21b}$$

The numerical value of k in these units is 1.6×10^{-7}. In an older system of units, where the dose is in rad and the total number of transitions is in microcurie-hour (see below), the equation is

$$D\ (\text{rad}) = \left[\tilde{A}\ (\mu\text{Ci h})\right]\left[\Delta_i\ (\text{g rad}\ \mu\text{Ci}^{-1}\ \text{h}^{-1})\right]\left[\Phi_i\ (\text{g}^{-1})\right]. \tag{17.21c}$$

The next three subsections discuss cumulated activity, the mean energy emitted, and the absorbed fraction of the energy. Then all of these concepts are combined with examples of absorbed dose calculations.

17.6.1 Activity and Cumulated Activity

The activity $A(t)$ is the number of radioactive transitions (or transformations or disintegrations) per second. The SI unit of activity is the *becquerel* (Bq):

$$1\ \text{Bq} = 1\ \text{transition}\ \text{s}^{-1}. \tag{17.22}$$

The earlier unit of activity, which is still used occasionally, is the *curie* (Ci):

$$\begin{aligned} 1\,\text{Ci} &= 3.7 \times 10^{10}\,\text{Bq}, \\ 1\,\mu\text{Ci} &= 3.7 \times 10^{4}\,\text{Bq}. \end{aligned} \tag{17.23}$$

The cumulated activity $\tilde{A}$ is the total number of transitions that take place. The SI unit of cumulated activity is the transition or the Bq s. Both are dimensionless. The old unit of cumulated activity is the μCi h:

$$1\,\mu\text{Ci}\,\text{h} = 1.332 \times 10^{8}\ \text{Bq s}. \tag{17.24}$$

Consider a sample of N_0 radioactive nuclei at time $t = 0$. The total number of nuclei remaining at time t is $N(t) = N_0 e^{-\lambda t}$, and the total activity is $A(t) = \lambda N(t) = A_0 e^{-\lambda t}$. The cumulated activity between times t_1 and t_2 is

$$\begin{aligned} \tilde{A}(t_1, t_2) &= \int_{t_1}^{t_2} A(t)\,dt = \frac{A_0}{\lambda}\left(e^{-\lambda t_1} - e^{-\lambda t_2}\right) \\ &= N(t_1) - N(t_2). \end{aligned} \tag{17.25}$$

If all times are considered, $t_1 = 0$ and $t_2 = \infty$,

$$\tilde{A} = \tilde{A}(0, \infty) = \frac{A_0}{\lambda} = \frac{A_0 T_{1/2}}{0.693} = 1.443 A_0 T_{1/2}. \tag{17.26}$$

This is, as we would expect, N_0.

17.6.1.1 The General Distribution Problem: Residence Time

Suppose that a radioactive substance is introduced in the body by breathing, ingestion, or injection. It may move into and out of many organs before decaying, and it may even leave the body. The details of how it moves depend on the pharmaceutical to which it is attached.

The cumulated activity in organ h is the total number of disintegrations in that organ:

$$N_h = \tilde{A}_h = \int_0^{\infty} A_h(t)\,dt. \tag{17.27}$$

The dose to organ k is then

$$D_k = \textstyle\sum_h N_h S(r_k \leftarrow r_h). \tag{17.28}$$

The units of N_h are disintegrations (dimensionless) or Bq s. If initial activity A_0 (Bq) is administered to the patient, the ratio N_h/A_0 is called the *residence time*[6]

$$\tau_h = \frac{N_h}{A_0} = \frac{\tilde{A}_h}{A_0} = \frac{\tilde{A}_h(0, \infty)}{A_0}. \tag{17.29}$$

[6] Stabin (2008) says that residence time is confusing. He recommends that the ratio $\tilde{A}_h/A_0$ should be called the *normalized cumulative activity* which has units of Bq s per Bq administered.

The residence time is the length of time that activity at a constant rate A_0 would have to reside in the organ to give that cumulated activity. The residence time for a given substance and organ must be determined by measurement, guided by the use of appropriate models. Many residence times have been determined and published. The presence of an abnormality in some organ can drastically alter the residence time.

We now calculate the cumulated activity and residence time for some simple situations.

17.6.1.2 Immediate Uptake with No Biological Excretion

This is the simplest example. A certain fraction of the radiopharmaceutical is taken up very rapidly in some organ, and it stays there. This is a good model for $^{99\text{m}}$Tc–sulfur colloid, which is used for liver imaging. About 85 % is trapped in the liver; the remainder goes to the spleen and elsewhere (Loevinger et al. 1988, p. 23). The activity in the organ is $A_h(t) = A_h e^{-\lambda t}$. [Note the difference between the activity in organ h as a function of time, $A_h(t)$, the initial activity in organ h, A_h, and the cumulated activity in organ h, $N_h = \tilde{A}_h$.] Let the fraction of the activity in the organ be F_h. The cumulated activity is

$$\tilde{A}_h = A_h \int_0^{\infty} e^{-\lambda t}\,dt = \frac{A_h}{\lambda} = \frac{F_h A_0}{\lambda}.$$

The residence time is

$$\tau_h = \frac{\tilde{A}_h}{A_0} = \frac{F_h}{\lambda} = 1.443 F_h T_{1/2}. \tag{17.30}$$

17.6.1.3 Immediate Uptake with Exponential Biological Excretion

Suppose that in addition to physical decay with decay constant λ, the pharmaceutical moves to another organ while it is still radioactive. Such a process can be complicated, perhaps involving storage in the gut or bladder. In other cases, the disappearance from a particular organ may be close to exponential with a biological disappearance constant λ_j. (Assume for now that all the radioactive nuclei can disappear biologically. If some are bound in different chemical forms, this might not be true.) If N is the number of radioactive nuclei in the organ (not the total number originally administered), then the rate of change of N is

$$\frac{dN}{dt} = -(\lambda + \lambda_j)N,$$

the solution to which is $N(t) = N_0 e^{-(\lambda+\lambda_j)t}$. The activity is λN, not $|dN/dt|$. Since it is proportional to N, we can again write

$$A_h(t) = A_h e^{-(\lambda+\lambda_j)t} = \lambda N_0 e^{-(\lambda+\lambda_j)t}. \tag{17.31}$$

Again, $N_0 = A_h/\lambda$. The decay constant $\lambda + \lambda_j$ is larger than the physical decay constant. The effective half-life is

$$(T_j)_{\text{eff}} = \frac{0.693}{\lambda + \lambda_j}. \tag{17.32}$$

In terms of the physical and biological half-lives T and T_j, this is

$$\frac{1}{(T_j)_{\text{eff}}} = \frac{1}{T} + \frac{1}{T_j} \tag{17.33}$$

or

$$(T_j)_{\text{eff}} = \frac{TT_j}{T + T_j}. \tag{17.34}$$

The cumulated activity is

$$\begin{aligned} \tilde{A}_h(t_1, t_2) &= A_h \int_{t_1}^{t_2} e^{-(\lambda+\lambda_j)t}\, dt \\ &= \frac{A_h}{\lambda + \lambda_j} \left(e^{-(\lambda+\lambda_j)t_1} - e^{-(\lambda+\lambda_j)t_2}\right). \end{aligned} \tag{17.35}$$

The cumulated activity for all time is

$$\tilde{A}_h = \frac{A_h}{\lambda + \lambda_j} = 1.443\,(T_j)_{\text{eff}}\, A_h. \tag{17.36}$$

17.6.1.4 Immediate Uptake Moving through Two Compartments

Consider the simplest two-compartment model. A total of N_0 nuclei are administered that move immediately to the first compartment. They then move exponentially from the first compartment to the second but do not move back. The number in the first compartment is given by

$$\frac{dN_1}{dt} = -(\lambda_1 + \lambda)N_1. \tag{17.37}$$

The radioactive decay constant is λ and the biological disappearance rate is λ_1. In compartment 2, the substance enters from compartment 1 and is biologically removed with constant λ_2:

$$\frac{dN_2}{dt} = +\lambda_1 N_1 - (\lambda + \lambda_2)N_2. \tag{17.38}$$

Suppose we start with no nuclei in either compartment and inject N_0 nuclei in compartment 1 at $t = 0$. Then one can show (see Problem 13) that

$$N_1(t) = N_0 e^{-(\lambda+\lambda_1)t} \tag{17.39}$$

so

$$\frac{dN_2}{dt} = \lambda_1 N_0 e^{-(\lambda+\lambda_1)t} - (\lambda + \lambda_2)N_2, \tag{17.40}$$

the solution to which is

$$N_2(t) = N_0 \frac{\lambda_1}{\lambda_1 - \lambda_2} \left(e^{-(\lambda+\lambda_2)t} - e^{-(\lambda+\lambda_1)t}\right). \tag{17.41}$$

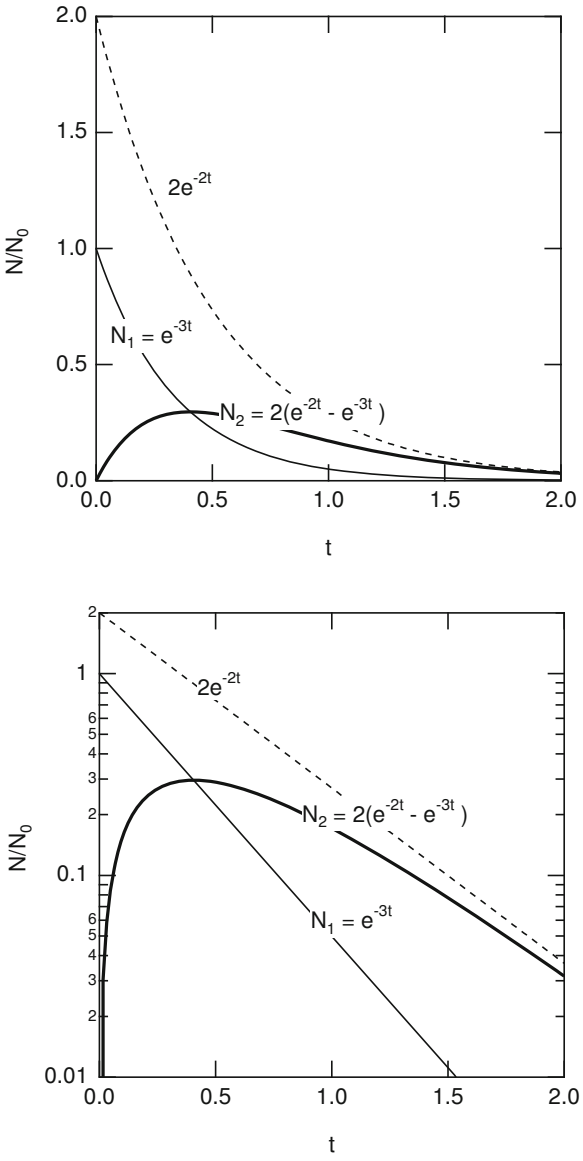


Fig. 17.11 An example of two-compartment transfer when $\lambda = 1$, $\lambda_1 = 2$, and $\lambda_2 = 1$

These solutions are worth examining. They are plotted in Fig. 17.11 for $\lambda = 1$, $\lambda_1 = 2$, and $\lambda_2 = 1$. The number of nuclei in compartment 1 is $N_0 e^{-3t}$. At first, many of the particles leaving compartment 1 enter compartment 2, and N_2 rises. When there is no more of the substance entering the second compartment from the first, N_2 decays at a rate $\lambda + \lambda_2 = 2$. This corresponds to the vanishing of the second term in Eq. 17.41. The larger the value of λ_1, the faster the second term vanishes. For very large values of λ_1, the second term vanishes quickly, the factor $\lambda_1/(\lambda_1 - \lambda_2)$ approaches unity, and the decay is nearly $N_2(t) = N_0 e^{-(\lambda+\lambda_2)t}$. The case

$\lambda_1 = \lambda_2$ is discussed in Problem 15. The activities are

$$A_1(t) = \lambda N_1(t), \qquad A_2(t) = \lambda N_2(t)$$

and the cumulated activities are obtained by integration:

$$\begin{aligned} \tilde{A}_1 &= \frac{A_0}{\lambda + \lambda_1}, \\ \tilde{A}_2 &= \frac{A_0 \lambda_1}{(\lambda + \lambda_1)(\lambda + \lambda_2)}. \end{aligned} \tag{17.42}$$

The residence times are

$$\begin{aligned} \tau_1 &= \frac{1}{\lambda + \lambda_1}, \\ \tau_2 &= \frac{\lambda_1}{(\lambda + \lambda_1)(\lambda + \lambda_2)}. \end{aligned} \tag{17.43}$$

17.6.1.5 More Complicated Situations

A number of more complicated situations are solved by Loevinger et al. (1988). These include situations where substances move between compartments in both directions, the experimental data for the activity have been fit with a series of exponentials, and convolution techniques are used. All of these cases are for isotopes and pharmaceuticals used in clinical practice.

17.6.1.6 Activity per Unit Mass

It is sometimes convenient to use the *mean initial activity per unit mass*

$$C_h = \frac{A_h}{m_h}\,\mathrm{Bq\,kg^{-1}} \tag{17.44}$$

and the *cumulated mean activity per unit mass*

$$\tilde{C}_h = \frac{\tilde{A}_h}{m_h} = \frac{\tau_h A_0}{m_h}\,\mathrm{kg^{-1}}. \tag{17.45}$$

Earlier units for these were $\mu\mathrm{Ci\,g^{-1}}$ and $\mu\mathrm{Ci\,h\,g^{-1}}$.

17.6.2 Mean Energy Emitted Per Unit Cumulated Activity

The mean energy emitted per unit cumulated activity Δ_i is determined by knowing Y_i and E_i for each particle or photon that is emitted. For a given nuclear transformation, the Y_i and E_i must include all photons (whether γ rays or x rays) and all electrons (betas, internal conversion electrons, and Auger electrons). In SI units,

$$\Delta_i \text{ (in J)} = Y_i E_i \text{ (in J)}. \tag{17.46a}$$

If E_i is expressed in MeV, we must use the conversion factor $1\ \mathrm{MeV} = 1.6 \times 10^{-13}$ J. In the old system of units, there is the conversion factor:

$$\begin{aligned} \Delta_i\ (\mathrm{g\,rad\,\mu Ci^{-1}\,h^{-1}}) &= Y_i E_i\ (\mathrm{MeV}) \\ &\times (3.7 \times 10^4\,\mathrm{s^{-1}\,\mu Ci^{-1}})(1.6 \times 10^{-13}\,\mathrm{J\,MeV^{-1}}) \\ &\times (10^7\,\mathrm{erg\,J^{-1}})(3.6 \times 10^3\,\mathrm{s\,h^{-1}})(10^{-2}\,\mathrm{rad\,g\,erg^{-1}}) \\ \Delta_i &= 2.13 Y_i E_i. \end{aligned} \tag{17.46b}$$

17.6.3 Calculation of the Absorbed Fraction

The remaining part of the dose determination problem is the most difficult: the calculation of $\phi(r_k \leftarrow r_h)$, the fraction of the radiation of a certain type emitted in region r_h that is absorbed in region r_k. A lot has been published on this problem; this section provides only an introduction.

17.6.3.1 Nonpenetrating Radiation

The simplest case is for charged particles or photons of very low energy that lose all their energy after traveling a short distance. If the source volume is much larger than this distance, we can say that the target volume is the same as the source volume:

$$\phi(r_k \leftarrow r_h) = \begin{cases} 0, & r_k \neq r_h \\ 1, & r_k = r_h. \end{cases} \tag{17.47}$$

17.6.3.2 Infinite Source in an Infinite Medium

Suppose that a radioactive source is distributed uniformly throughout a region that is so large that edge effects can be neglected. The activity per unit mass is C, so the total activity is $\tilde{A} = M\tilde{C}$, where M is the mass of the material. The energy released is $\tilde{A}\Delta$. This is absorbed in mass M, so the fractional absorbed energy is 1, as in case 1. The dose is

$$D = \frac{M\tilde{C}\Delta}{M} = \tilde{C}\Delta. \tag{17.48}$$

(This is why Δ used to be called the *equilibrium absorbed dose constant*.)

17.6.3.3 Point Source of Monoenergetic Photons in Empty Space

Another simple case is a point source of monoenergetic photons in empty space. The total amount of energy released by the source is $\tilde{A}\Delta_i$. If the energy of the radiation is E_i, the number of photons is $\tilde{A}\Delta_i/E_i$. At distance r the number per unit area is $\tilde{A}\Delta_i/4\pi r^2 E_i$. If a small amount of substance of

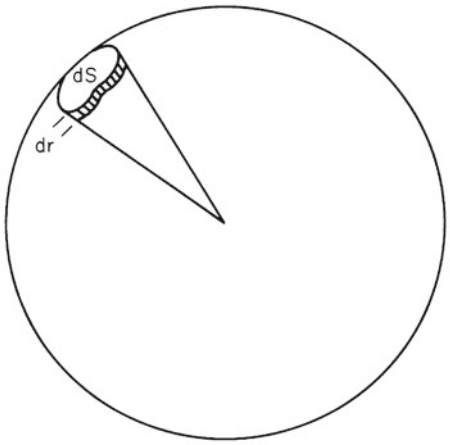


Fig. 17.12 A small volume of absorbing material is introduced at distance r from a point source of γ rays

area dS, density ρ, thickness dr, and energy absorption coefficient μ_{en} is introduced as in Fig. 17.12, the amount of energy absorbed in it is E_i times the number of photons absorbed: $\delta E = \tilde{A}\Delta_i dS\,\mu_{\text{en}} dr/4\pi r^2$. Therefore, the absorbed fraction is

$$\phi = \frac{\delta E}{\tilde{A}\Delta_i} = \frac{\mu_{\text{en}}\, dr\, dS}{4\pi r^2}. \tag{17.49}$$

This is exactly what we expect from the definition of ϕ. If the source radiates its energy isotropically, the fraction passing through dS is $dS/(4\pi r^2)$. The fraction of that energy absorbed in dr is $\mu_{\text{en}}\, dr$. The specific absorbed fraction is

$$\Phi = \frac{\phi}{M} = \frac{\phi}{\rho dS dr} = \frac{\mu_{\text{en}}}{4\pi r^2 \rho}. \tag{17.50}$$

17.6.3.4 Point Source of Monoenergetic Photons in an Infinite Isotropic Absorber

If the source is not in empty space but in an infinite, homogeneous, isotropic absorbing medium, the number of photons at distance r from the source is modified by the factor $e^{-\mu_{\text{atten}} r} B(r)$, where the *buildup factor* $B(r)$ accounts for secondary photons. Therefore,

$$\phi = \frac{\mu_{\text{en}}\, dr\, dS\, e^{-\mu_{\text{atten}} r}}{4\pi r^2} B(r) \tag{17.51}$$

and

$$\Phi = \frac{\mu_{\text{en}}}{\rho} \frac{e^{-\mu_{\text{atten}} r}}{4\pi r^2} B(r). \tag{17.52}$$

See also Sect. 15.17. The buildup factor has been tabulated for photons of various energies in water (Berger 1968).

17.6.3.5 More Complicated Cases—the MIRD Tables

For more realistic geometries, the calculation of ϕ is quite complicated. Tables for humans of average build have been prepared by the MIRD Committee (Snyder et al. 1975, 1976, 1978) . A *Monte Carlo* computer calculation was used. The description below shows how it works in principle; the actual calculations, though equivalent, are different in detail to save computer time. The radioactive nuclei are assumed to be distributed uniformly throughout the source organ. A point within the source organ is picked. The model emits a photon of energy E in some direction, picked at random from all possible directions. This photon is followed along its path; for every element ds of its path, the probability of its interacting, $\mu_{\text{atten}} ds$, is calculated. The computer program then "flips a coin" with this probability of having heads. If a head occurs, the photon is considered to interact at that point. If the interaction is Compton scattering, the angle is picked at random with a relative probability given by the differential cross section. The energy of a recoil electron for that scattering angle is calculated and deposited at the interaction site. Similar procedures are followed for the photoelectric effect and pair production. The scattered photon is then followed in the same way. If a tail occurred on the first flip, the photon is allowed to travel another distance ds and the probability of interaction is again calculated. This procedure is repeated until all the energy has been absorbed.

To determine what kind of material the photons are traveling through, a model of the body called a *phantom* is used. An example of a phantom is shown in Fig. 17.13.

This entire procedure is repeated many times for each organ, until one has a map of the radiation deposited in all organs by γ rays leaving that point in the source organ. The procedure is described in much greater detail by Snyder et al. (1976). Table 17.2 shows a portion of a table for ϕ. A computer code (OLINDA/EXM) is usually used to make the calculations (Stabin et al. 2005). Recently 3-d imaging has made it possible to make patient-specific dose calculations (Dewaraja et al. 2012).

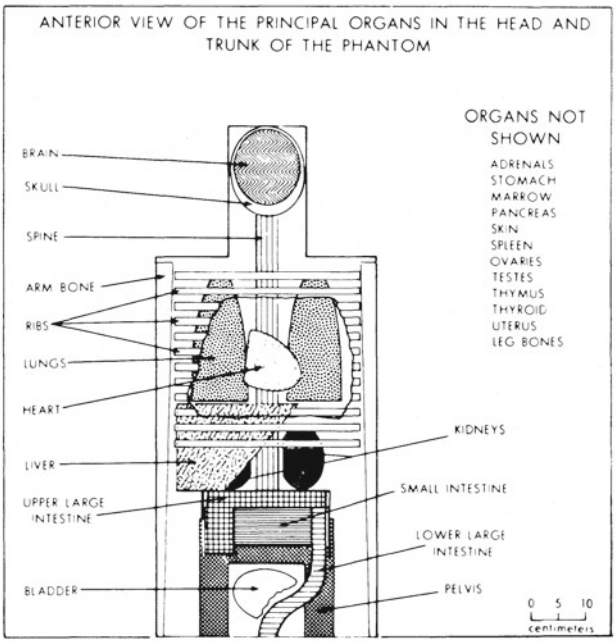
Arqueros and Montesinos (2003) provide a pedagogical discussion of Monte Carlo simulation of γ-ray transport. A pedagogical program for whole-body Monte Carlo calculations has been developed by Hunt et al. (2004). It is available through the RADAR (Radiation Dose Assessment Resource) web site: www.doseinfo-radar.com.

Often most of the isotope is taken up in one or two organs, and the rest of it distributes fairly uniformly through the rest of the body. Using the subscript h for the organs with the greatest activity, TB to mean total body, and RB to mean the rest of the body,

$$\tilde{A}_{\text{RB}} = \tilde{A}_{\text{TB}} - \textstyle\sum_h \tilde{A}_h. \tag{17.53}$$

The dose is then

$$D_k = \textstyle\sum_h \tilde{A}_h S(r_k \leftarrow r_h) + \tilde{A}_{\text{RB}} S(r_k \leftarrow \text{RB}). \tag{17.54}$$



(a)

HEART

The heart is half an ellipsoid capped by a hemisphere which is cut by a plane. A rotation and translation are then effected. The heart (Fig. 4) is represented by

$$x_1 = 0.6943\,(x + 1) - 0.3237\,(y + 3) - 0.6428\,(z - 51),$$

$$y_1 = 0.4226\,(x + 1) + 0.9063\,(y + 3),$$

$$z_1 = 0.5826\,(x + 1) - 0.2717\,(y + 3) + 0.7660\,(z - 51),$$

$$\left(\frac{x_1}{8}\right)^2 + \left(\frac{y_1}{5}\right)^2 + \left(\frac{z_1}{5}\right)^2 \leqq 1,$$

$$x_1^2 + y_1^2 + z_1^2 \leqq (5)^2 \quad \text{if} \quad x_1 < 0,$$

$$\frac{x_1}{3} + \frac{z_1}{5} \geqq -1 \quad \text{if} \quad x_1 < 0$$

and has a volume of 603.1 cm^3.

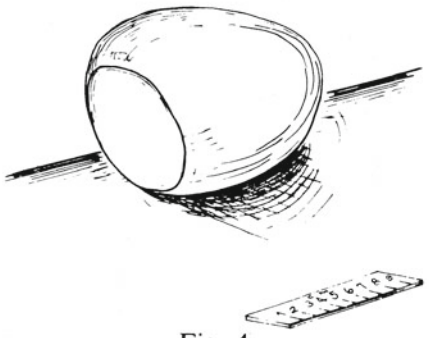


Fig. 4

(b)

Fig. 17.13 An early phantom used by the MIRD Committee for calculations of the absorbed fraction. **a** A view of the whole body. **b** Details of the heart boundaries. (Reprinted by permission of the Society of Nuclear Medicine and Molecular Imaging from W. S. Snyder, M. R. Ford, G. G. Warner, and H. L. Fisher. MIRD Pamphlet No. 5. Estimates of Absorbed Fractions for Monoenergetic Photon Sources Uniformly Distributed in Various Organs of a Heterogeneous Phantom. *J Nucl Med* 1969; **10** (Suppl. 3): 5–52, Figs. 4 and 5)

The quantity $S(r_k \leftarrow \mathrm{RB})$ cannot easily be tabulated, since it depends on what organs are included in the sum over h. Substituting the tabulated quantity $S(r_k \leftarrow \mathrm{TB})$ introduces errors because the "hot" organs that have significant activity are included a second time. One solution to this problem is to modify the cumulated activities (Coffey and Watson 1979). First, define a uniform total body cumulated activity that has the same cumulated activity per unit mass as the rest of the body:

$$\tilde{A}_u = \frac{m_{\mathrm{TB}}}{m_{\mathrm{RB}}} \tilde{A}_{\mathrm{RB}}. \tag{17.55}$$

This activity in the total body would give a dose

$$D_k = \tilde{A}_u S(r_k \leftarrow \mathrm{TB}).$$

Table 17.2 Absorbed fractions for a uniform source of ^{99m}Tc in the lungs, calculated using the tables in Snyder et al. (1976)

Target organ	ϕ
Adrenals	1.39×10^{-4}
Bladder	8.83×10^{-5}
GI(stomach)	2.58×10^{-3}
GI(SI)	1.46×10^{-3}
GI(ULI)	4.49×10^{-4}
GI(LLI)	7.02×10^{-5}
Heart	1.35×10^{-2}
Kidneys	9.14×10^{-4}
Liver	1.66×10^{-2}
Lungs	4.95×10^{-2}
Marrow	2.16×10^{-2}
Pancreas	5.35×10^{-4}
Skeleton (rib)	2.03×10^{-2}
Skeleton (pelvis)	4.21×10^{-4}
Skeleton (spine)	7.59×10^{-3}
Skeleton(skull)	1.11×10^{-3}
Skeleton (total)	5.21×10^{-2}
Skin	5.43×10^{-3}
Spleen	1.46×10^{-3}
Thyroid	4.14×10^{-5}
Uterus	1.55×10^{-5}
Trunk	3.71×10^{-1}
Legs	4.05×10^{-4}
Head	7.70×10^{-3}
Total body	3.79×10^{-1}

Then define for each organ of interest the quantity $\tilde{A}_h^*$, which is the difference between the actual activity in organ h and that assuming the substance is uniformly distributed in the total body:

$$\tilde{A}_h^* = \tilde{A}_h - \frac{m_h}{m_{\text{RB}}} \tilde{A}_{\text{RB}}. \tag{17.56}$$

Then the dose to organ k is

$$D_k = \sum_h \tilde{A}_h^* S(r_k \leftarrow r_h) + \tilde{A}_u S(r_k \leftarrow \text{TB}). \tag{17.57}$$

Problem 31 shows that Eqs. 17.55–17.57 are consistent with Eqs. 17.53 and 17.54 if

$$\frac{m_{\text{RB}}}{m_{\text{TB}}} S(r_k \leftarrow \text{RB}) + \sum_h \frac{m_h}{m_{\text{TB}}} S(r_k \leftarrow r_h) = S(r_k \leftarrow \text{TB}), \tag{17.58}$$

which is consistent with a uniform source $\tilde{A}_u$ distributed throughout the body. The dose can be determined either by calculating the modified activities and using the total body S in Eq. 17.57, or by calculating S for the rest of the body from Eq. 17.58 and using the unmodified activities. Problem 32 shows how these reformulations work in a simple case.

17.6.4 Sample Dose Calculation

We pull this discussion together by making a simplified calculation of the dose to various organs from ^{99m}Tc-labeled microspheres used in a lung scan. We assume that 37 MBq of ^{99m}Tc is injected, that it all lodges in the capillaries of the lung, and that it remains there long enough so that the half-life is the physical half-life.[7] The residence time is then $\tau_h = 1.443 T_{1/2} = (1.443)(6) = 8.658$ h, so the cumulated activity is $\tilde{A}_{\text{lung}} = (3.7 \times 10^7)(8.658 \times 60 \times 60) = 1.153 \times 10^{12}$ Bq s (Table 17.3).

The dose to the lungs is considerably greater than in a chest x ray; however, a chest x ray is almost useless for diagnosing a pulmonary embolus. The whole body dose is not unreasonable.

Table 17.4 shows some typical doses from various nuclear medicine procedures. The effective dose is defined on page 491.

17.7 Radiopharmaceuticals and Tracers

A radioactive nucleus by itself is not very useful. It must usually be attached to some substance that will give it the desired biological properties, for example, to be preferentially absorbed in the region of interest. It must also be prepared in a sterile form, free of toxins that produce a fever (*pyrogens*) so that it can be injected in the patient. This section surveys some of the properties of *radiopharmaceuticals*. Much more detail can be found in Cherry et al. (2012) and Kowalsky and Falen (2011).

The radioactive nuclei used in nuclear medicine are not found in nature. They are produced by bombarding a stable isotope with neutrons (from a nuclear reactor) or protons (from a cyclotron). The bombardment may yield fission fragments or an isotope that is useful as produced. In other cases the isotope produced has a half life that is long enough to ship it to a hospital. Its decay product has a shorter half life and is the isotope used in the pharmaceutical. See Cherry et al. (2012), Chap. 5.

17.7.1 Physical Properties

The half-life must be short enough so that a reasonable fraction of the radioactive decays take place during the diagnostic procedure; any decays taking place later gives the patient a dose that has no benefit. (This requirement can be relaxed if

[7] The last is not a good assumption. The ^{99m}Tc leaches from the microspheres into the general circulation. A more accurate calculation requires measurements and the use of a convolution integral, as described in Loevinger et al. (1988, pp. 79–81). The principal residence times are 4.3 h in the lung, 1.8 h in the extravascular space, 0.83 h in the urine, 0.7 h in the kidney, and 0.6 h in the blood.

Table 17.3 Values of Δ_i, E_i, and ϕ_i for ^{99m}Tc in the lung

i, Line in Fig. 17.4	Δ_i (J)	E_i (keV) (e denotes an electron)	ϕ_i				
			Lung	Heart	Liver	Head	Whole body
1	2.42×10^{-16}	e	1	0	0	0	1
2	4.52×10^{-17}	e	1	0	0	0	1
3	2.00×10^{-14}	140.5	0.0495	0.0135	0.0166	0.0077	0.3785
4	1.71×10^{-15}	e	1	0	0	0	1
5	2.18×10^{-16}	e	1	0	0	0	1
6	1.42×10^{-17}	e	1	0	0	0	1
7	7.45×10^{-18}	e	1	0	0	0	1
8	4.47×10^{-17}	e	1	0	0	0	1
9	8.56×10^{-18}	e	1	0	0	0	1
10	1.07×10^{-16}	e	1	0	0	0	1
11	2.11×10^{-17}	e	1	0	0	0	1
12	4.44×10^{-18}	e	1	0	0	0	1
13	1.36×10^{-17}	e	1	0	0	0	1
14	7.93×10^{-18}	e	1	0	0	0	1
15	6.23×10^{-17}	18.21	1	0	0	0	1
16	1.19×10^{-16}	18.33	1	0	0	0	1
17	2.16×10^{-17}	20.59	1	0	0	0	1
18	3.65×10^{-17}	e	1	0	0	0	1
19	1.60×10^{-17}	e	1	0	0	0	1
20	2.98×10^{-17}	e	1	0	0	0	1
21	5.26×10^{-18}	e	1	0	0	0	1
22	1.30×10^{-17}	0.11	1	0	0	0	1
23	3.57×10^{-17}	e	1	0	0	0	1
24	1.17×10^{-17}	0.029	1	0	0	0	1
$\sum \Delta_i \phi_i$			3.79×10^{-15}	2.70×10^{-16}	3.32×10^{-16}	1.54×10^{-16}	1.04×10^{-14}
m (kg)			0.999	0.603	1.833	5.278	70.036
$S = \sum \Delta_i \phi_i / m$			3.79×10^{-15}	4.48×10^{-16}	1.81×10^{-16}	2.92×10^{-17}	1.48×10^{-16}
Dose (Gy) $A_0 = 37$ MBq			4.37×10^{-3}	5.16×10^{-4}	2.09×10^{-4}	3.36×10^{-5}	1.71×10^{-4}

Table 17.4 Some typical doses for nuclear medicine procedures. (Adapted from Table 9-3 in Zanzonico et al. 1995)

Study and agent	A_0 (MBq)	Organ and highest dose (mSv)		Total body dose (mSv)	Effective dose (mSv)
Bone ^{99m}Tc–pyrophosphate	555	Bladder wall	51	2.0	4.4
Heart ^{201}Tl–chloride	55	Kidneys	20	3.6	13
Liver ^{99m}Tc–sulfur colloid	185	Bladder wall	17	0.9	2.6

the biological excretion is rapid.) On the other hand, the lifetime must be long enough so that the radiopharmaceutical can be prepared and delivered to the patient.

For diagnostic work, the decay scheme should minimize the amount of nonpenetrating radiation. Such radiation provides a dose to the patient but never reaches the detector. This means that there should be as few charged particles (β particles) as possible. The ideal source then is a γ source, which means that the nucleus is in an excited state (an isomer). Such states are usually very short-lived. Not only should the nucleus be a γ emitter, but the internal conversion coefficient should be small, since internal conversion produces nonpenetrating electrons. Positron emitters are more desirable than are β^- emitters because the positrons produce 0.5-MeV radiation that can reach an external detector. For therapy, on the other hand, nonpenetrating radiation is ideal.

It is also necessary that the decay product have no undesirable radiations. If the decay is a β^- or β^+ decay, the product has different chemical properties from the parent and may be taken up selectively by a different organ. If it is also radioactive, this can confuse a diagnosis and give an undesirable dose to the other organ.

Ease of chemical separation of the radioactive substance from whatever carrier it is produced with is also important. It is necessary to remove the radioactive isotope from stable isotopes of the same element, because the chemicals are

usually toxic. This toxicity is avoided by giving the chemical in minute amounts, which can only be done if the specific activity is high.

17.7.2 Biological Properties

For diagnostic work, a pharmaceutical is needed that is taken up more by the diseased tissue to give a *hot spot* or taken up less to give a *cold spot*. The former is easier to see with small amounts of radioactivity, but both techniques are used. For therapy one wishes to have selective absorption of the pharmaceutical so that the radiations will destroy the target organ but not the rest of the body. There are several mechanisms by which a pharmaceutical may be localized.

1. *Active transport.* The drug is concentrated by a specific organ against a concentration gradient. Examples are the selective concentration of iodine in the thyroid, salivary, and gastric glands. (It is rapidly excreted from the last two but is retained in the thyroid). This technique is also effective for certain drugs in the kidney.
2. *Phagocytosis*. Particles in the size range 1–1000 nm may be phagocytized—taken up by specialized cells of the reticuloendothelial system. This can take place in liver, bone marrow, and spleen. Particles of size 1 nm go to the Kupfer cells of the liver and to the marrow, while larger particles (100–1000 nm) are gathered by phagocytes in the liver and spleen.
3. *Sequestration*. Still larger particles, such as red blood cells that have been denatured by heat, are gathered in the spleen or liver by the process called sequestration. The particles are trapped as the blood percolates through the pulp of the spleen and are later phagocytized.
4. *Capillary blockade*. The capillaries have a diameter of 7-10 μm. Particles from 20 to 40 μm diameter injected into a vein will find progressively larger vessels as they work their way through the right heart and will be stopped in the capillaries of the lung.
5. *Diffusion*. It is also possible for a pharmaceutical to move through a membrane to a region of lower concentration. There is a blood–brain barrier between the blood and the central nervous system that is relatively impermeable even to small ions. In a brain scan the chemical is not concentrated in normal brain tissue but leaks into tissue where the blood–brain barrier is compromised by a lesion.
6. *Compartmental localization*. A suitable pharmaceutical injected in the blood may remain there a long time, mixing well and allowing the blood volume to be determined.

The most widely used isotope is ^{99m}Tc. As its name suggests, it does not occur naturally on earth, since it has no stable isotopes. We consider it in some detail to show how an isotope is actually used. Its decay scheme has been discussed above.

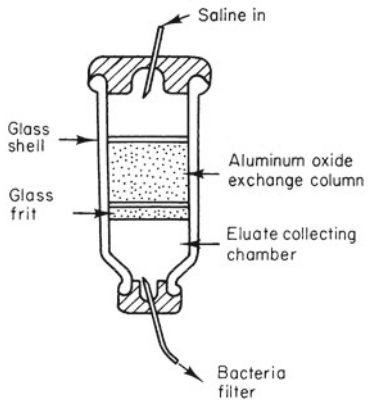


Fig. 17.14 A ^{99}Mo–^{99m}Tc generator system. Molybdenum is trapped in the aluminum oxide layer. Eluant introduced at the top flows through and is collected at the bottom

There is a nearly monoenergetic 140-keV γ ray. Only about 10 % of the energy is in the form of nonpenetrating radiation. The isotope is produced in the hospital from the decay of its parent, ^{99}Mo, which is a fission product of ^{235}U and can be separated from about 75 other fission products. The ^{99}Mo decays to ^{99m}Tc.

Technetium is made available to hospitals through a *generator* that was developed at Brookhaven National Laboratories in 1957 and is easily shipped. Isotope ^{99}Mo, which has a half-life of 67 h, is adsorbed on an alumina substrate in the form of molybdate (MoO_4^{2-}). From 8 to 100 GBq of ^{99}Mo can be provided. The heart of such a generator (without the lead shielding) is shown in Fig. 17.14. As the ^{99}Mo decays, it becomes pertechnetate (TcO_4^-). Sterile isotonic eluting solution is introduced under pressure above the alumina and passes through after filtration into an evacuated eluate container. After removal of the technetium, the continued decay of ^{99}Mo causes the ^{99m}Tc concentration to build up again. A generator lasts about a week.

Several steps must be taken to prepare the pertechnetate as a radiopharmaceutical. First, it must be checked for breakthrough of the ^{99}Mo. The Nuclear Regulatory Commission allows 1.5×10^{-4} Bq of ^{99}Mo per Bq of ^{99m}Tc. The purity is checked by placing the eluate in a lead sleeve that attenuates the ^{99m}Tc γ ray much more than the $\approx$ 750-keV γ rays from ^{99}Mo and measuring the activity. It is also checked with a colorimetric test for the presence of aluminum ion.

The eluate can be used directly for imaging brain, thyroid, salivary gland, urinary bladder, and blood pool, or it can be combined with phosphate, albumin or aggregated albumin, colloidal sulfur, or $FeCl_3$. Commercial kits are available for making these preparations.

For example, kits for labeling aggregated human albumin are commercially available. A vial containing 10 ml of saline solution is enough for ten doses. The aggregated albumin

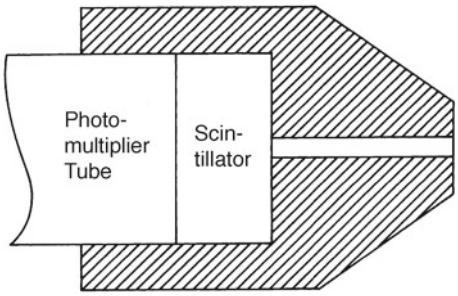


Fig. 17.15 A scintillator with a lead collimator to give directional sensitivity

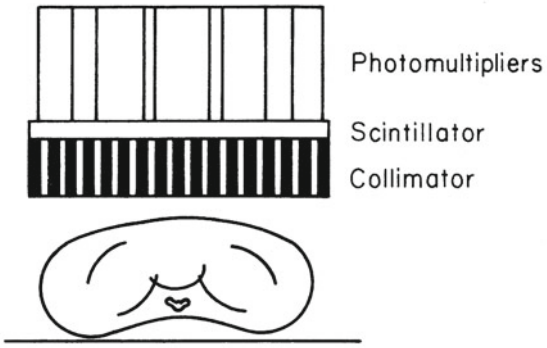


Fig. 17.16 Side view of a scintillation camera. A collimator allows photons from the patient to strike the scintillator directly above the source. An array of photomultiplier tubes records the position and energy of the detected photon

particles are 10–70 μm in diameter. Each milliliter of solution contains $(4 - 8) \times 10^5$ particles. Tin is attached to the microspheres and serves to bind technetium. Up to 10^9 Bq of technetium pertechnetate is added to the vial by the user. A typical adult dose is 10–40 MBq (3.51×10^5 albumin particles). The problems consider attaching Tc to the microspheres and what fraction of the capillaries are blocked by this kind of study.

Other common isotopes are ^{201}Tl, ^{67}Ga, and ^{123}I. Thallium, produced in a cyclotron (see Sect. 8.1), is chemically similar to potassium and is used in heart studies, though it is being replaced by ^{99m}Tc–sestamibi and ^{99m}Tc–tetrofosmin. Gallium is used to image infections and tumors. Iodine is also produced in a cyclotron and is used for thyroid studies. For many more details see Cherry et al. (2012) or Kowalsky and Falen (2011).

17.8 Detectors; The Gamma Camera

Nuclear medicine images do not have the inherent spatial resolution of diagnostic x-ray images; however, they provide functional information: the increase and decrease of activity as the radiopharmaceutical passes through the organ being imaged (Zanzonico 2012).

Early measurements were done with single detectors such as the scintillation detector[8] shown in Fig. 17.15. Directional sensitivity is provided by a collimator, which can be cylindrical or tapered. Single detectors are still used for in vitro measurements and for thyroid uptake studies.

Two-dimensional images can be taken with the *scintillation camera* or *gamma camera* shown in Figs. 17.16 and 17.17. The scintillator is 6–12 mm thick and about 60 cm

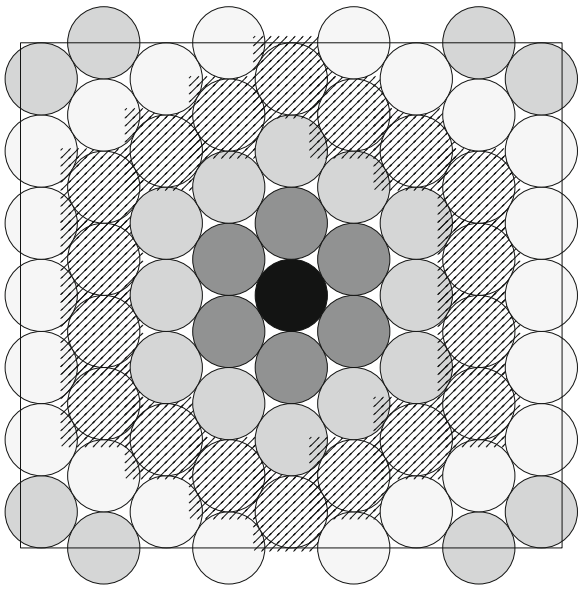


Fig. 17.17 A square scintillator viewed by an array of 67 photomultiplier tubes. The hexagonal arrangement of the tubes above the scintillator gives the closest spacing between tubes

across. Modern scintillators are rectangular. The scintillator is viewed by an array of 50–100 photomultiplier tubes arranged in a hexagonal array. The tube nearest where the photon interacts receives the greatest signal. Signals from each tube are combined to give the total energy signal and x and y position signals.

The collimator is a critical component of the gamma camera. The channels are usually hexagonal, with walls just thick enough to stop most of the photons which do not pass down the collimator opening. The collimator usually has parallel channels. Single pinholes, diverging, and converging channels are sometimes used and can lead to geometric distortions of the image (Cherry et al. 2012). The spatial resolution depends on the distance from the source to the collimator, as

[8] Scintillation detectors were discussed in Sect. 16.3.

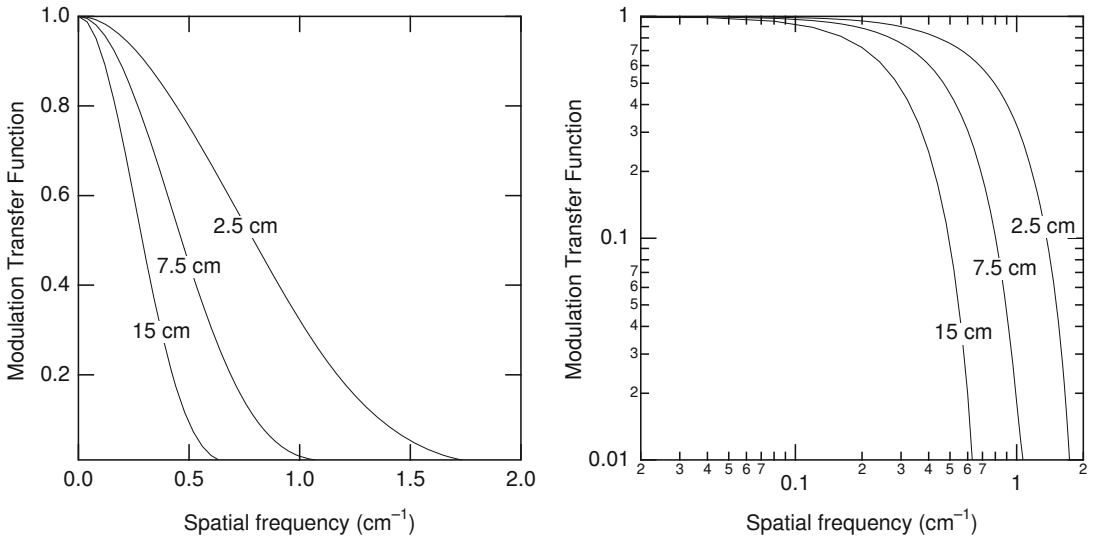


Fig. 17.18 Modulation transfer function curves for a typical parallel-hole collimator for different source-to-collimator distances. Both linear and log-log plots are shown. The source-to-collimator distances are 2.5, 7.5, and 15 cm. (Data are from Erhardt et al. 1978, p. 39)

shown for one collimator in Fig. 17.18. There are trade-offs between sensitivity and resolution (Links and Engdahl 1995). Some of the aspects of collimator design are discussed in Problems 49–51.

Figure 17.19 shows a bone scan of a child taken with a gamma camera. The ^{99m}Tc–diphosphonate is taken up in areas of rapid bone growth. Bone growth at the epiphyses at the end of each bone can be seen. There are also hot spots at the injection site, in one kidney, and in the bladder.

Nuclear medicine can show physiologic function. For example, if the isotope is uniformly distributed in the blood, viewing the heart and synchronizing the data accumulation with the electrocardiogram (gating) allows one to measure blood volume in the heart when it is full and contracted, and to calculate the *ejection fraction*, the fraction of blood in the full left ventricle that is pumped out. Fig. 17.20, shows pictures and contours of the heart at end-systole and end-diastole. The imaging agent was ^{99m}Tc-labeled human red blood cells.

Figure 17.21 shows a series of images taken at six different angles around a patient who has had a lung transplant. The left lung is new and shows considerably more activity than the diseased right lung.



Fig. 17.19 A scintillation camera *bone scan* of a 7-year-old male who received a ^{99m}Tc–diphosphonate injection. An anterior view is on the left, and a posterior view is on the right. The scan shows an area of decreased uptake surrounded by a dark ring in the right anterior skull, consistent with an *eosinophilic granuloma*. Identifiable hot regions are the injection site in the right elbow, an attempted injection site in the right hand, the bladder, and the left kidney, which is probably not remarkable on this exam, along with the ends of the long bones. (Photograph courtesy of B. Hasselquist, Ph.D., Department of Diagnostic Radiology, University of Minnesota)

17.9 Single-Photon Emission Computed Tomography

Still another detection scheme, *single-photon emission computed tomography* (SPECT), is analogous to computed tomography (CT). The detector is sensitive to all radioactivity

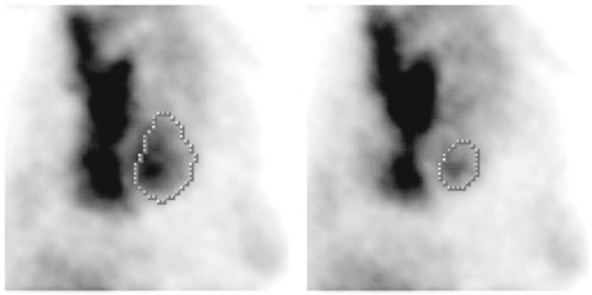


Fig. 17.20 Two gated-scintillation camera views of the heart, imaged with ^{99m}Tc-labeled red blood cells. The dots outline the left ventricle. On the left is end diastole (left ventricle filled with blood). On the right is end systole (left ventricle at smallest volume). The ejection fraction is 66 %. (Photograph courtesy of B. Hasselquist, Ph.D., Department of Diagnostic Radiology, University of Minnesota)

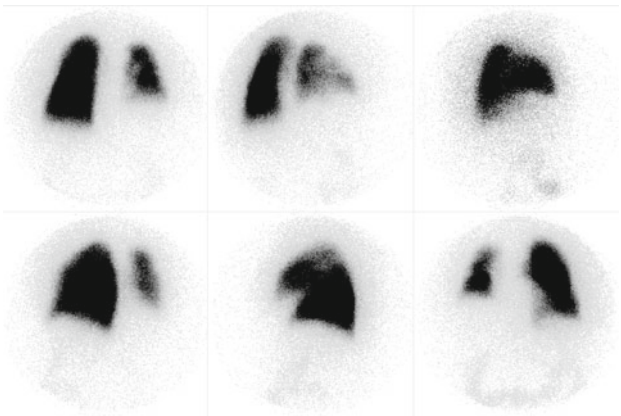


Fig. 17.21 Lung scans of a patient who has received a lung transplant. The upper left is a posterior view; each successive view is rotated about the patient, ending with an anterior view on the lower right. The left lung is the transplant. It has much more activity than the diseased right lung. (Photograph courtesy of B. Hasselquist, Ph.D., Department of Diagnostic Radiology, University of Minnesota)



Fig. 17.22 Single photon emission computed tomographic (SPECT) slices of the heart. The patient was injected with ^{99m}Tc–tetrofosmin, an agent that is taken up by myocardium. The images have been reconstructed in planes parallel to the axis of the heart. The dark myocardium surrounds the blood in the left ventricle. (Photograph courtesy of B. Hasselquist, Ph.D., Department of Diagnostic Radiology, University of Minnesota)

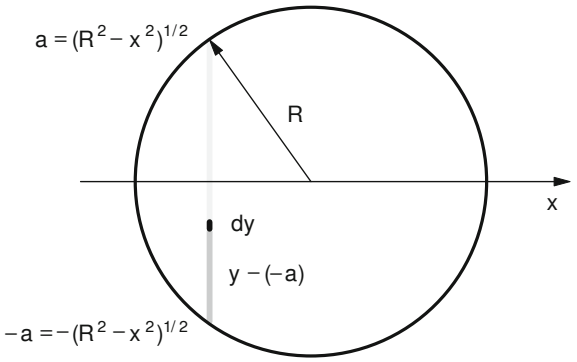


Fig. 17.23 Projection perpendicular to the x axis for a radioactive source of uniform concentration, including the effect of photon attenuation

along a line passing through the patient. The counting rate is thus proportional to a projection through the patient, and a cross-sectional slice can be reconstructed from a series of projections, just as was done with x-ray CT using the techniques in Chap. 12. A series of images like those in Fig. 17.21, but at more angles, are used to reconstruct a three-dimensional image that can then be viewed from any direction, with slices at any desired depth. A SPECT scan is shown in Fig. 17.22. There are five reconstructed slices in planes parallel to the long axis of the heart. The left ventricle is prominent, and the right ventricle can be seen faintly in the last few slices.

One of the problems with SPECT is photon attenuation along the projection line. This is shown in Fig. 17.23 for a cylindrical source with uniform activity throughout. Let A_V be the activity per unit volume, and ignore variations in $1/r^2$. The projection $F(x)$ is

$$F(x) = \int_{-a}^{a} A_V(x, y)\, \Delta x\, \Delta z\, e^{-\mu(y+a)}\, dy, \qquad (17.59)$$

where $dy\, \Delta x\, \Delta z$ is the volume detected. When $A_V(x, y)$ is constant (a uniform activity distribution), this can be integrated to give

$$F = \frac{A_V \Delta x \Delta z}{\mu} \left(1 - e^{-2\mu(R^2 - x^2)^{1/2}}\right). \qquad (17.60)$$

This is plotted in Fig. 17.24 for $\mu = 0$, $\mu = 10\ \mathrm{m}^{-1}$ (511-keV annihilation radiation) and $\mu = 15\ \mathrm{m}^{-1}$ (140-keV ^{99m}Tc). When $\mu = 0$, $F(x) = A_V(2a\Delta x \Delta z)$, where $2a$ is the thickness of the source along the projection. Corrections for attenuation are made in a number of ways.[9] Other nuclides used besides ^{99m}Tc are ^{81m}Kr, ^{133}Xe, ^{131}I, ^{67}Ga, ^{123}I, and ^{201}Tl.

[9] See Cherry et al. (2012), pp. 288–303.

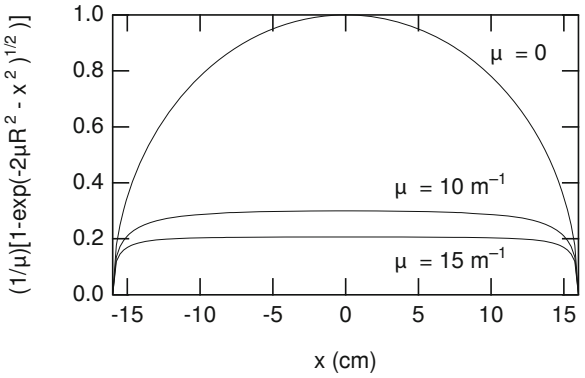


Fig. 17.24 Plot of the projection including attenuation, Eq. 17.60, for $\mu = 0$, $\mu = 10\ \mathrm{m}^{-1}$ (corresponding to 511-keV annihilation radiation) and $\mu = 15\ \mathrm{m}^{-1}$ (corresponding to the photons from ^{99m}Tc). The radius of the circle is 16 cm

More details can be found in Cherry et al. (2012) and Zanzonico (2012).

17.10 Positron Emission Tomography

If a positron emitter is used as the radionuclide, the positron comes to rest and annihilates an electron, emitting two annihilation photons back to back. In *positron emission tomography* (PET) these are detected in coincidence. This simplifies the attenuation correction, because the total attenuation for both photons is the same for all points of emission along each γ ray through the body (see Problem 55); also the photons have a higher energy (511 keV) and lower attenuation coefficient than those used in SPECT. Positron emitters are short-lived, and it is necessary to have a cyclotron for producing them in or near the hospital. This is proving to be less of a problem than initially imagined. Commercial cyclotron facilities deliver isotopes to a number of nearby hospitals. Patterson and Mosley (2005) found that 97 % of the people in the United States live within 75 miles of a clinical PET facility. Muehllehner and Karp (2006) review the history and uses of PET. See also Zanzonico (2004).

We have mentioned that nuclear medicine procedures have the potential to measure function, as the molecules to which the isotopes are bound move from organ to organ in the body. This is particularly true for some of the lighter positron emitters, which have the advantage of being natural constituents of molecules in the body or similar to them (Table 17.5). PET can provide a *functional image* with information about metabolic activity. A very common positron agent is ^{18}F fluorodeoxyglucose—glucose in which a hydroxyl group has been replaced with ^{18}F. The PET signal is largest in those cells that have taken up the ^{18}F because they are actively metabolizing glucose. PET has become particularly important in studies of brain function, where active neurons are detected by an increase in their metabolism, and in locating metastatic cancer. The number of installed PET scanners is growing very rapidly. Most of them have built-in CT scanners to provide accurate fused PET/CT images (Christian and Waterstram-Rich 2012).

Table 17.5 Positron emitters used in nuclear medicine

Nuclide	Half-life
$^{11}_{6}$C	20.3 min
$^{13}_{7}$N	10.0 min
$^{15}_{8}$O	2.1 min
$^{18}_{9}$F	109.7 min

There is great interest in using ^{11}C, in spite of its short (20-min) half life, because it can be incorporated in molecules of biological interest.

A PET scan overlaid on a magnetic resonance (MRI) image is shown in Fig. 17.25 The positron emitter is ^{15}O-labeled water (2.1-min half life). The views are described in the caption. The subject is sequentially touching each finger of the left hand with the thumb. Activity can be seen in the right cerebral sensorimotor cortex (slice, upper right) and in the left cerebellum (slice, lower left). The technique is described by Rehm et al. (1994) and Strother et al. (1995).

17.11 Brachytherapy and Internal Radiotherapy

Brachytherapy (*brachy* means short) involves implanting directly in a tumor sources for which the radiation falls off rapidly with distance because of attenuation, short range, or $1/r^2$. Originally the radioactive sources (seeds) were implanted surgically, resulting in high doses to the operating room personnel. In the *afterloading* technique, developed in the 1960s, hollow catheters are implanted surgically and the sources inserted after the surgery. Remote afterloading, developed in the 1980s, places the sources by remote control, so that only the patient receives a radiation dose.

We saw in Chap. 16 that fractionation of the dose results in better sparing of normal tissue for a given probability of killing the tumor. Afterloading allows the sources to be placed and removed, but it is often difficult for the patient to tolerate the catheters for long periods of time. This has led to the development of *high-dose-rate brachytherapy* (HDR), in which the dose is given in one or a few fractions over the course of a day or two (Nag 1994). Though this is much easier for the patient, tissue sparing is not as great as with

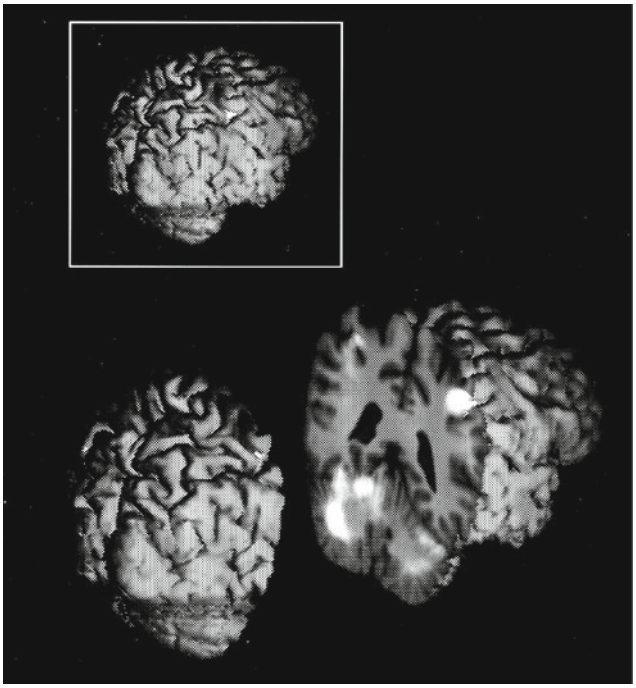


Fig. 17.25 A positron emission tomography (*PET*) scan is overlayed on an MR image. At the upper left is a three dimensional MRI of the brain viewed from above and to the right. At the bottom the image has been sliced through the motor strip and cerebellum, and the two pieces are separated. The PET image has been overlaid on the slice. The positron emitter is ^{15}O-labeled water. The subject is sequentially touching each finger of the left hand with the thumb. Activity can be seen in the right cerebral sensorimotor cortex (slice, upper right) and in the left cerebellum (slice, lower left). (Image courtesy of Prof. Kelly Rehm, University of Minnesota and the PET Imaging Service, Veterans Administration Medical Center, Minneapolis)

a longer treatment. Current practice seeks to compensate for this by meticulous treatment planning based on an extended version of the linear-quadratic model, and by making sure that the tumor receives much higher doses than the surrounding normal tissue.

Radium was the first brachytherapy source, but it has been replaced by a number of nuclei that decay by β^- emission or electron capture. One common source is ^{137}Cs. It undergoes β^- decay to a metastable state of ^{137}Ba, which then emits a 662-keV gamma ray. The β particles are absorbed in the stainless steel tube enclosing the cesium, so the dose is due to the gamma rays (Khan 2010, Chap. 15). Conventional low-dose-rate brachytherapy is delivered at 0.4–1.0 Gy hr^{-1}. High dose rates are about 1 Gy min^{-1}.

Patients with coronary artery disease are often treated with *balloon angioplasty*, in which a coronary artery is dilated by inserting a balloon on the end of a catheter into the femoral artery in the leg and from there through the aorta and into the coronary artery. One problem is *restenosis* or reclosure of the artery. Restenosis can be reduced by placing a *stent*—a helical coil of wire—in the artery at the time of the angioplasty. Restenosis sometimes occurs within a stent, though the rate of recurrence is reduced by using a stent which elutes (gives off) a restenosis-inhibiting drug. If restenosis does occur, it can be treated by placing a string of radioactive seeds in the stent. Treatments may use either a gamma emitter, ^{192}Ir, for 20 min, or a beta emitter (^{90}Sr/^{90}Y) for 3 min (Kaluza and Raizner 2004; Fox 2002).

Internal radiotherapy treats the patient with a radionuclide in a chemical that is selectively taken up by the tumor. The classic example is the administration by mouth of capsules containing ^{131}I for treatment of hyperthyroidism and thyroid cancer. Other nuclides are being used for breast cancer, neuroendocrine tumors, and melanoma (Fritzberg and Wessels 1995). A radionuclide for this purpose should emit primarily nonpenetrating radiation, have a physical half-life long compared to the biological half-life, have a large activity per unit mass, and exhibit a high degree of specificity for the tumor. If the nuclide can be delivered within the cell, then Auger electrons can be exploited. One way to achieve high concentrations in the tumor is *radioimmunotherapy*: monoclonal antibodies are tagged with the radionuclide such as ^{125}I (see the special issue of *Medical Physics* edited by Buchsbaum and Wessels 1993). It turns out that double-strand DNA breaks from Auger cascades occur more often than had been expected, and that the bystander effect is important. The use of Auger electrons from nuclides attached to the appropriate antibodies for cancer therapy is under active development; see Kassis (2011). The MIRD formulation can be adapted to the dose calculations (Watson et al. 1993). Radionuclide therapy is described for a general audience by Coursey and Nath (2000).

17.12 Radon

The naturally occurring radioactive nuclei are either produced continuously by cosmic γ ray bombardment, or they are the products in a decay chain from a nucleus whose half-life is comparable to the age of the earth. Otherwise they would have already decayed. There are three naturally-occurring radioactive decay chains near the high-Z end of the periodic table. One of these is the decay products from $^{238}_{92}$U, shown in Fig. 17.26. The half-life of ^{238}U is 4.5×10^9 yr, which is about the same as the age of the earth. A series of α and β decays lead to radium, ^{226}Ra, which undergoes α decay with a half-life of 1620 yr to radon, ^{222}Rn.

Uranium, and therefore radium and radon, are present in most rocks and soil. Radon, a noble gas, percolates through grainy rocks and soil and enters the air and water in different concentrations. Although radon is a noble gas, its decay

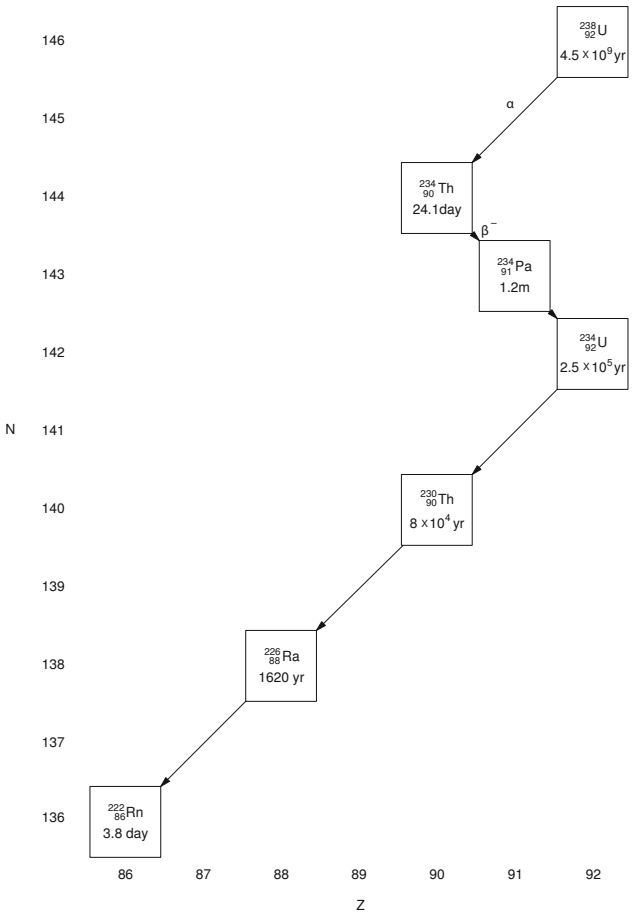


Fig. 17.26 Decay of ^{238}U to radon

products have different chemical properties and attach to dust or aerosol droplets which can collect in the lungs. High levels of radon products in the lungs have been shown by both epidemiological studies of uranium miners and by animal studies to cause lung cancer (Committee on the Biological Effects of Ionizing Radiations, BEIR IV 1988; BEIR VI 1999). The deposition process is quite complicated. A certain fraction of the decay products attach to aerosol droplets. That fraction is an important parameter in estimating the dose, because the unattached particles are deposited in the airways; those that have attached to aerosols are also deposited in the airways, the site depending on the droplet size. The rate at which natural mucus clearing from the lungs removes them is also variable.

The $^{222}_{86}$Rn decay scheme is shown in Fig. 17.27. (Alternate branches that occur very rarely are not shown.) The shaded nuclides are the greatest contributors to the dose. Radon is a noble gas; once it decays the other shaded nuclides decay shortly after. Radon dosimetry is described on pp. 137–158 of BEIR IV (1988) and in BEIR VI (1999). Typical uranium activities in soil are 20 Bq kg^{-1} (range 7–40), leading to radon concentrations in the air over average soil of about 4 Bq m^{-3}.

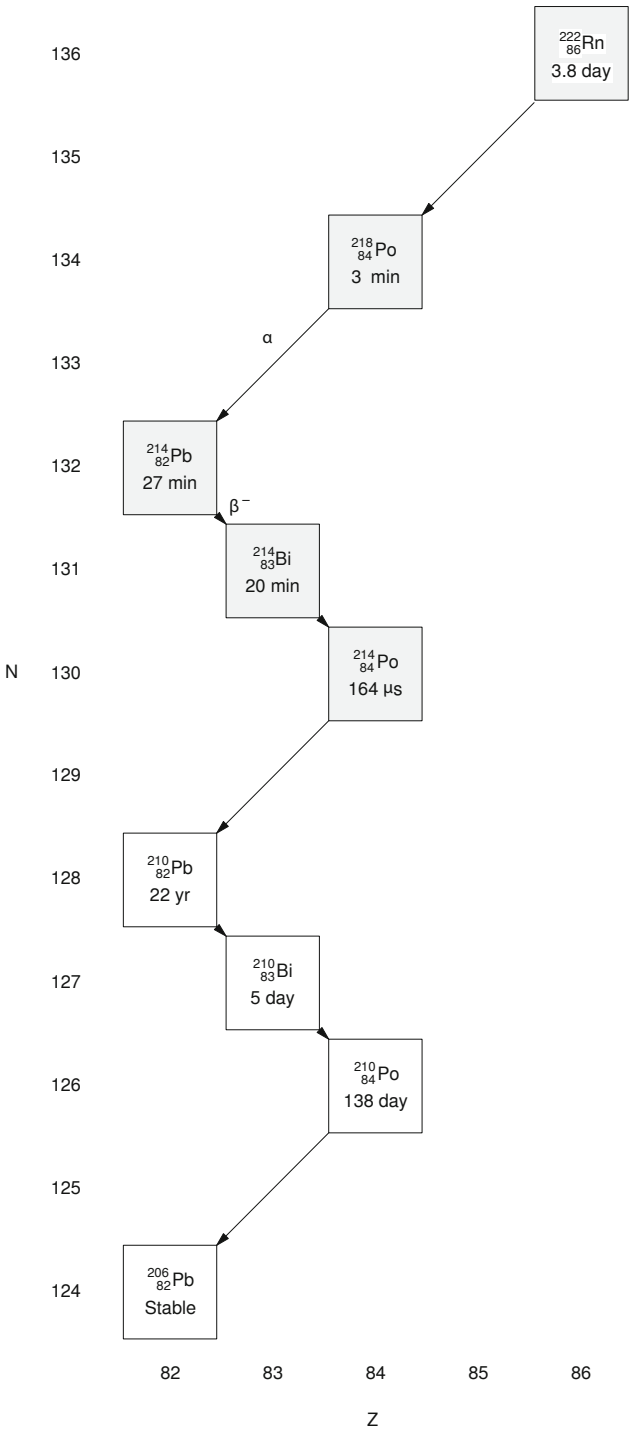


Fig. 17.27 Decay of radon. The decay of the shaded nuclides is most significant in determining dose

The *working level* (WL) has been defined to be any combination of the shaded isotopes in Fig. 17.27 in 1 l of air at ambient temperature and pressure that results in the ultimate emission of 1.3×10^5 MeV of α-particle energy. This is about the energy liberated by the decay products in equilibrium with 100 pCi (3.7 Bq) of radon. Thus 1 WL corresponds

to 3.7 Bq l^{-1} or 3700 Bq m^{-3}. More recently, the activity of radon and its decay products has been described by the *Potential Alpha Energy Concentration* (PAEC) (BEIR VI 1999, p. 179). Its units are J m^{-3}.

The *working-level month* (WLM) measures the total radon exposure and is 1 WL for 170 h (1 month of 40-h work weeks). Another unit is the PAEC multiplied by the number of hours exposure, measured in J h m^{-3}. There are 3.5×10^{-3} J h m^{-3} per WLM.

Dose estimates for the miners and for the general population require models of aerosol size, unattached fraction, target cells, exercise level, and occupancy factors that are described in BEIR IV (1988). Averaging over all of these variables shows a dose in the lungs of about 6 mGy per WLM, with a factor-of-2 uncertainty because of these variables.

The report uses a time-since-exposure model to estimate the risk of lung cancer on the basis of four studies of groups of miners. The model predicts a relative risk ratio that is unity for no exposure and increases linearly to 3.5 for a continuous exposure of 5 WLM per year over a lifetime.[10] The report uses the linear-no-threshold model to estimate risks to the general population at small exposures. The issue of applying the linear-no-threshold model was discussed in Sect. 16.12. See particularly the data from the Cohen study in Fig. 16.53. Typical radon concentrations in houses are usually less than $4r_0$ or 4 pCi l^{-1} (128 Bq m^{-3}) or 0.04 WL. (We saw in Sect. 16.12 that $r_0 = 37\,\mathrm{Bq\,m^{-3}} = 1\,\mathrm{pCi\ l^{-1}}$). Exposure to r_0 for 24 h per day for one year gives 0.5 WLM. The miners had exposures of 5–100 WLM per year, over periods of 3–20 years.

Symbols Used in Chapter 17

Symbol	Use	Units	First used page
a	Distance	m	522
b	Source to collimator distance	m	530
c	Speed of light	m s^{-1}	504
d	Width of collimator channel	m	530
e	Electron charge	C	504
g	Detector efficiency		530
h	Planck's constant	J s	510
h, k	Denote specific organs		511
l	Collimator thickness	m	530
m_0	Rest mass	kg	504
m_x	Rest mass of particle type x	kg	505
p	Momentum	kg m s^{-1}	509
r_h, r_k	Source and target regions		511
r	Distance	m	515
r_0	Radon concentration unit	Bq m^{-3}	526
s	Path length	m	515
t	Time	s	506
t	Collimator septum thickness	m	530
v	Speed	m s^{-1}	504
w	Distance across collimator wall in the direction of photon travel	m	530
x, y, z	Position	m	522
A	Mass number		503
A, A_0	Activity	Bq	507
$\tilde{A}_h$	Cumulated activity in organ h	Bq s	511
B	Buildup factor		515
B, B_K, B_L	Binding energy	eV	507
$C_h, \tilde{C}_h$	Activity and cumulated activity per unit mass in organ h	Bq kg^{-1}; kg^{-1}	514
D	Dose	J kg^{-1} (Gy)	511
E, E_γ	Energy	J, eV	504
F_h	Fraction of activity in organ h		512
F	Projection	Bq	522
K	Geometric factor		530
M, M_X	Mass	kg	509
N	Neutron number		503
N, N_0	Number of nuclei		506
R, R_0	Nuclear radius	m	504
R	Radius of disk	m	522
R_t, R_o	True and observed counting rates	s^{-1}	530
S	Mean absorbed dose per unit cumulated activity	J kg^{-1}	511
S	Area	m^2	515
T	Kinetic energy	J, eV	504
T	Time	s	527
$T_{1/2}$	Half-life	s	507
T_j	Half-life for jth biological disappearance process	s	513
Y_i	Mean number (fraction) of transitions of type i per transformation		511
Z	Atomic number (number of protons)		503
α_h	Fraction of total activity in organ h		528
β^-, β^+	Electron and positron (in β decay)		507
λ	Physical decay constant	s^{-1}	506
λ_j	Decay constant for jth biological process	s^{-1}	512
μ	Attenuation coefficient	m^{-1}	522
μ_{en}	Energy absorption coefficient	m^{-1}	515
$\nu, \overline{\nu}$	Neutrino, antineutrino		509
ν	Photon frequency	Hz	510
ρ	Density	kg m^{-3}	515
ϕ_i	Absorbed fraction		511
τ	Detector dead time	s	530
τ_h	Residence time in organ h	s	512
Δ_i	Mean energy emitted in radiation type i	J	511
Φ_j	Specific absorbed fraction	kg^{-1}	511

[10] BEIR IV (1988), Fig. 2.2. This is averaged by BEIR over smokers and nonsmokers and by us over sex.

Problems

Section 17.1

Problem 1. In the 1940s, a pressing question in biology was whether DNA or protein was responsible for the transmission of genetic information. A simple system to study this is a *bacteriophage*, a virus that injects a substance into *Escherichia coli*, thereby transforming the bacteria's genetic material. Design an experiment using radioactive tracers that could determine whether DNA or protein was the injected substance. Hint: DNA contains many phosphorus atoms but no sulfur, whereas protein has many sulfur atoms but no phosphorus. Alfred Hershey and Martha Chase performed such an experiment in 1952.

Problem 2. An alpha particle is fired directly at a stationary aluminum nucleus. Assume the only interaction is the electrostatic repulsion between the alpha particle and the nucleus, and the nucleus is so heavy that it is stationary. Calculate the distance of their closest approach as a function of the initial kinetic energy of the alpha particle. This calculation is consistent with Ernest Rutherford's famous alpha particle scattering formula for energies lower than 3 MeV, but deviates from his formula for energies higher than 3 MeV. If the alpha particle enters the nucleus, the nuclear force dominates and the formula no longer applies. Estimate the radius of the aluminum nucleus.

Problem 3. The best current (2010) value for the mass of the proton is 1.007276467 u. The mass of the electron is $5.485799095 \times 10^{-4}$ u. The BE of the electron in the hydrogen atom is 13.6 eV. Calculate the mass of the neutral hydrogen atom.

Problem 4. Solve Eq. 17.1 for the kinetic energy, T. Show that when $v \ll c$, it reduces to the familiar $T = m_0 v^2/2$.

Problem 5. The rest energy of the ${}^{184}_{74}\mathrm{W}$ nucleus is 171303 MeV. The average binding energies of the electrons in each shell are

Shell	Number of electrons	BE per electron (eV)
K	2	69,525
L	8	11,015
M	18	2125
N	32	213
O	12	49
P	2	≈ 6

Calculate the atomic rest energy of tungsten.

Section 17.5

Problem 6. Refer to Figs. 17.2 and 17.5. Uranium splits roughly in half when it undergoes nuclear fission. Will the fission fragments decay by β^+ or β^- emission?

Problem 7. The following nuclei of mass 15 are known: ${}^{15}_{6}\mathrm{C}$, ${}^{15}_{7}\mathrm{N}$, and ${}^{15}_{8}\mathrm{O}$. Of these, ${}^{15}\mathrm{N}$ is stable. How do the others decay?

Problem 8. Look up the decay schemes of the following isotopes (for example, in the *Handbook of Chemistry and Physics,* CRC Press or at www.nndc.bnl.gov/). Comment on their possible medical usefulness: ${}^{3}\mathrm{H}$, ${}^{15}\mathrm{O}$, ${}^{13}\mathrm{N}$, ${}^{18}\mathrm{F}$, ${}^{22}\mathrm{Na}$, ${}^{68}\mathrm{Ga}$, ${}^{64}\mathrm{Cu}$, ${}^{11}\mathrm{C}$, ${}^{123}\mathrm{I}$, and ${}^{56}\mathrm{Ni}$.

Problem 9. Look up the half lives of the isotopes in Fig. 17.6 (for instance in the Handbook of Chemistry and Physics. CRC Press or at www.nndc.bnl.gov/). Relate qualitatively the half life to the position of the isotope on the parabola.

Section 17.6

Problem 10. Calculate the conversion factor k of Eq. 17.21b.

Section 17.6.1

Problem 11. Show that 1 μCi h $= 1.332 \times 10^8$ disintegrations or Bq s.

Problem 12. Obtain a numerical value for the residence time for ${}^{99\mathrm{m}}$Tc-sulfur colloid in the liver if 85 % of the drug injected is trapped in the liver and remains there until it decays.

Problem 13. Derive Eqs. 17.39–17.41.

Problem 14. Calculate numerical solutions of Eqs. 17.39 and 17.41 and plot them on semilog paper. Use $\lambda = 2$, $\lambda_1 = 0.5$, $\lambda_2 = 3$.

Problem 15. Eq. 17.41 is not valid if $\lambda_1 = \lambda_2$. In that case, try a solution of the form $N_2 = Bte^{-\alpha t}$ where α is to be determined, and obtain a solution.

Problem 16. Derive Eqs. 17.42 and 17.43.

Problem 17. The biological half-life of iodine in the thyroid is about 25 days. ${}^{125}\mathrm{I}$ has a half-life of 60 days. ${}^{132}\mathrm{I}$ has a half-life of 2.3 h. Find the effective half-life in each case.

Problem 18. For Sect. 17.6.1.4, with $\lambda = 0.05\ \mathrm{h}^{-1}$, $\lambda_1 = 1\ \mathrm{h}^{-1}$, and $\lambda_2 = 0.1\ \mathrm{h}^{-1}$, find $\tilde{A}_1$ and $\tilde{A}_2$ in terms of the initial activity A_0 and in terms of the initial number of nuclei N_0.

Problem 19. N_0 radioactive nuclei with physical decay constant λ are injected in a patient at $t = 0$. The nuclei move into the kidney at a rate λ_1, so that the number in the rest of the body falls exponentially: $N(t) = N_0 e^{-(\lambda+\lambda_1)t}$. Suppose that the nuclei remain in the kidney for a time T before moving out in the urine. (This is a crude model for the radioactive nuclei being filtered into the glomerulus and then passing through the tubules before going to the bladder.)

(a) Calculate the cumulated activity and the residence time in the kidney by finding the total number of nuclei entering the kidney and multiplying by the probability that a nucleus decays during the time T that it is in the kidney.
(b) Calculate the cumulated activity and residence time in the bladder, assuming that the patient does not void.

Problem 20. Suppose that at $t = 0$, ^{99m}Tc with an activity of 370 kBq enters a patient's bladder and stays there for 2 h, at which time the patient voids, eliminating all of it. What is the cumulated activity? What is the cumulated activity if the time is 4 h?

Problem 21. Suppose that the ^{99m}Tc of the previous problem does not enter the bladder abruptly at $t = 0$, but that it accumulates linearly with time. At the end of 2 h the activity is 370 kBq and the patient voids, eliminating all of it. What is the cumulated activity?

Problem 22. A radioactive substance has half-life $T_{1/2}$. It is excreted from the body with biological half-life T_1. N_0 radioactive nuclei are introduced in the body at $t = 0$. Find the total number that decay inside the body.

Problem 23. The *fractional distribution function* α_h is the fraction of the total activity that is in organ h: $\alpha_h(t) = A_h(t)/A(t) = A_h(t)/A_0 e^{-\lambda t}$.
(a) Show that $\tau_h = \int_0^\infty \alpha_h(t) e^{-\lambda t} dt$.
(b) Calculate $\alpha_1(t)$ and $\alpha_2(t)$ for Eqs. 17.39 and 17.41 and show that integration of these expressions leads to Eqs. 17.43.

Problem 24. Suppose that the fractional distribution function (defined in the previous problem) is $\alpha(t) = 1,\ t < T$; $\alpha(t) = b,\ t > T$; $(b < 1)$. Find the residence time. This is a simple model for the situation where a bolus (a fixed amount in a short time) of some substance passes through an organ once and is then distributed uniformly in the blood.

Problem 25. The *distribution function* $q_h(t)$ is defined to be the activity in organ h corrected for radioactive decay to a reference time. If the correction is from time t to time 0, find an expression for $q_h(t)$ in terms of $A_h(t)$.

Problem 26. The "official" definition of the fractional distribution function $\alpha_h(t)$ is the ratio of the distribution function $q_h(t)$ produced by a bolus administration to the patient, divided by the activity A_0 in the bolus. Show that this is equivalent to the definition in Problem 23.

Problem 27. Show that if the uptake in a compartment is not instantaneous but exponential, with subsequent exponential decay, the cumulated activity is $\tilde{A} = 1.443 A_0 (T_e T_{ue}/T_u)$, where T_e is the effective half-life for excretion, and $T_{ue} = T_u T_{1/2}/(T_u + T_{1/2})$. Hint: see Eq. 17.42.

Section 17.6.2

Problem 28. Rearrange the data of Fig. 17.4. Find the total Δ for emission of photons below 30 keV and charged particles. Rank the radiations in the order they contribute to the dose.

Problem 29. Nitrogen-13 has a half-life of 10 min. All of the disintegrations emit a positron with end point energy 1.0 MeV (average energy 0.488 MeV). There is no electron capture. Make a table of radiations that must be considered for calculating the absorbed dose and determine E_i and Δ_i for each one.

Problem 30. A patient swallows 3.5×10^9 Bq of ^{131}I. The half-life of the iodine is 8 days. Ten min later the patient vomits all of it. If none had yet left the stomach and all was vomited, determine the cumulated activity and residence time in the stomach.

Section 17.6.3

Problem 31. Derive Eq. 17.57 by substituting Eqs. 17.55 and 17.56 in Eq. 17.54. You will also have to justify and use Eq. 17.58.

Problem 32. The body consists of two regions. Region 1 has mass m_1 and cumulated activity $\tilde{A}_1$. It is completely surrounded by region 2 of mass m_2 and cumulated activity $\tilde{A}_2 = \tilde{A}_0 - \tilde{A}_1$. We can say that the mass of the total body is $m_{TB} = m_1 + m_2 = m_1 + m_{\mathrm{RB}}$. A single radiation is emitted with disintegration energy Δ. The radiation is nonpenetrating so that

$$\phi(1 \leftarrow 1) = \phi(2 \leftarrow 2) = 1,$$

$$\phi(1 \leftarrow 2) = \phi(2 \leftarrow 1) = 0.$$

(a) What are $\phi(\mathrm{TB} \leftarrow 1)$ and $\phi(\mathrm{TB} \leftarrow 2)$?
(b) What are the corresponding values of Φ and S?
(c) Show that directly from the definition, Eq. 17.54

$$D_1 = \tilde{A}_1 \Delta / m_1,$$

$$D_2 = D_{\mathrm{RB}} = \tilde{A}_2 \Delta / m_2,$$

$$D_{\mathrm{TB}} = \tilde{A}_0 \Delta / (m_1 + m_2)$$

(d) Calculate $\tilde{A}_u$ and $\tilde{A}_1^*$.
(e) What is $S(1 \leftarrow \mathrm{TB})$? Remember that ϕ is calculated for activity uniformly distributed within the source region.
(f) Calculate the dose to region 1 using Eq. 17.57 and show that it agrees with (c).
(g) Evaluate $S(1 \leftarrow \mathrm{RB})$ using Eq. 17.58 and show that it agrees with $S(1 \leftarrow 2)$.

Problem 33. The body consists of two regions. Region 1 has mass m_1 and cumulated activity $\tilde{A}_1$. It is completely surrounded by region 2 of mass m_2 and cumulated activity $\tilde{A}_2$. A single radiation is emitted with disintegration energy Δ. The characteristics of the radiation are such that

$$\phi(1 \leftarrow 1) + \phi(2 \leftarrow 1) = 1,$$

$$\phi(1 \leftarrow 2) + \phi(2 \leftarrow 2) + \phi(0 \leftarrow 2) = 1,$$

where $\phi(0 \leftarrow 2)$ represents energy from region 2 that has escaped from the body. Obtain expressions for the dose to each region and the whole body dose.

Problem 34. Consider the decay of a parent at rate λ_1 to an offspring that decays with rate λ_2.

(a) Write a differential equation for the amount of offspring present.
(b) Solve the equation.
(c) Discuss the solution when $\lambda_2 > \lambda_1$.
(d) Discuss the solution when $\lambda_2 < \lambda_1$.
(e) Plot the solution for a technetium generator that is eluted every 24 h.

Problem 35. N_0 nuclei of ^{99m}Tc are injected into the body. What is the maximum activity for the decay of the metastable state? When does the maximum activity for decay of the ground state occur if no Tc atoms are excreted? What is the ratio of the maximum metastable state activity to the maximum ground-state activity?

Problem 36. If 1 μCi of ^{99m}Tc is injected in the blood and stays there, relate the activity in a sample drawn time t later to the volume of the sample and the total blood volume. If the gamma rays are detected with 100 % efficiency, what will be the counting rate for a 10-ml sample of blood if the blood volume is 5 l? (Using non-SI units was intentional.)

Problem 37. Assume that aggregated human albumin is in the form of microspheres. A typical dose of albumin microspheres is 0.5 mg of microspheres containing 80 MBq of ^{99m}Tc and 15 μg of tin. There are 1.85×10^6 microspheres per mg.

(a) How many ^{99m}Tc atoms are there per microsphere?
(b) How many tin atoms per microsphere?
(c) How many technetium atoms per tin atom?
(d) What fraction of the surface of a microsphere is covered by tin? Assume the sphere has a density of 10^3 kg m^{-3}.

Problem 38. It is estimated that the total capillary surface area in the lung is 90 m^2. Assume each capillary has 50 segments, each 10 μm long, and a radius of 5 μm.

(a) How many capillaries are there in the lung?
(b) There are about 3×10^8 alveoli in both lungs. How many capillaries per alveolus are there?
(c) An alveolus is 150–300 μm in diameter. Are the above answers consistent?
(d) A typical dose of albumin microspheres is 0.5 mg with an average diameter of 25 μm. There are 1.85×10^6 spheres per mg. What fraction of the capillaries are blocked if there is good mixing?

Section 17.6.4

Problem 39. Look up the decay schemes and half-lives for ^{123}I and ^{131}I. Explain why ^{123}I is used to image the thyroid and ^{131}I is used to treat thyroid cancer.

Problem 40. Identify all the isotopes in Fig. 17.7 using the ^A_ZSymbol notation. What are the stable isotopes? What isotope can decay by both β^- and β^+ emission?

Problem 41. The half-life of ^{99m}Tc is 6.0 h. The half-life of ^{131}I is 8.07 day. Assume that the same initial activity of each is given to a patient and that all of the substance remains within the body.

(a) Find the ratio of the cumulated activity for the two isotopes.
(b) ^{99m}Tc emits 0.141-MeV photons. For each decay of ^{131}I the most important radiations are 0.89 β^- of average energy 0.192 MeV and 0.81 photons of 0.365 MeV. If all of the decay energy were absorbed in the body, what would be the ratio of doses for the same initial activity?

Problem 42. A patient is given an isotope that spreads uniformly through the lungs. It emits a single radiation: a γ ray of energy 50 keV. There are no internal-conversion electrons. The cumulated activity is 40 GBq s. Find the absorbed dose in the liver ($m = 1.83$ kg).

Problem 43. The decay of ^{99m}Tc can be approximated by lumping all of the decays into two categories:

Radiation	E_i (MeV)	Δ_i (J)
γ	0.14	2×10^{-14}
Electrons and soft x rays		2.76×10^{-15}

Sulfur colloid labeled with 100 MBq of ^{99m}Tc is given to a patient and is taken up immediately by the liver. Assume it stays there. Find the dose to the liver, spleen, and whole body. Use the following information:

Absorbed fraction for a source in the liver

Target organ	Mass (kg)	$E(\gamma) =$ 0.14 MeV
Liver	1.833	0.161
Spleen	0.176	0.000629
Whole body	70.0	0.431

Problem 44. An ionization type smoke detector contains 4.4 μCi of ^{241}Am. This isotope emits α particles (which we will ignore) and a 60-keV γ ray, for which $n = 0.36$. The half-life is 458 yr.

(a) How many moles of ^{241}Am are in the source?
(b) Ignoring attenuation, backscatter, and buildup in any surrounding material (such as the cover of the smoke detector), what is the absorbed dose in a small sample of muscle located 2 m away, if the muscle is under the detector for 8 h per day for 1 year?

Problem 45. One mCi of a radioactive substance lodges permanently in a patients lungs. The substance emits a single 80-keV γ ray. It has a half-life of 12 h. Find the cumulated activity and the dose to the liver (mass 1833 g).

Problem 46. The dose calculation for microspheres in the lung was an oversimplification because technetium leaches

off the spheres. The footnote in Sect. 17.6.4 lists some more realistic residence times. If none of the technetium is excreted from the body, the sum of all the residence times will still be 8.7 h. Assume that the residence time in the lungs is 4.3 h and the residence time in the rest of the body is 4.4 h.

(a) Show that $\tilde{A}_u = 4.46 \times 3600 \times A_0$ and $\tilde{A}^*_{\text{lung}} = 4.24 \times 3600 \times A_0$.

(b) For a source distributed uniformly throughout the total body, the absorbed fractions for 140-keV photons are $\phi(\text{lung} \leftarrow \text{TB}) = 0.0053$, $\phi(\text{TB} \leftarrow \text{TB}) = 0.3572$. Split the radiation into penetrating and nonpenetrating components:

$$S(\text{lung} \leftarrow \text{TB}) = (\phi_{\text{nonpen}} \Delta_{\text{nonpen}} + \phi_{\text{penetrating}} \Delta_{\text{penetrating}})/m_{\text{lung}}.$$

Remember that for activity uniformly distributed in the total body, $\phi(\text{lung} \leftarrow \text{TB}) = m_{\text{lung}}/m_{\text{TB}}$ and use some of the information in Table 17.3 to show that

$$S(\text{lung} \leftarrow \text{TB}) = 1.463 \times 10^{-16}\,\text{J kg}^{-1},$$

$$S(\text{TB} \leftarrow \text{TB}) = 1.414 \times 10^{-16}\,\text{J kg}^{-1}.$$

(c) Calculate the dose to the lungs and the total body dose for an initial activity of 37 MBq. Compare the values to those in Table 17.3.

Section 17.8

Problem 47. Nuclear counting follows Poisson statistics. Show that for a fixed average counting rate R (counts per second) the standard deviation of a sum of N measurements each of length T is the same as a single measurement of duration NT. (Hint: You will first have to consider the situation where one measures $y = x_1 + x_2 + \cdots$ and find the variance of y in terms of the variances of the x_i when there is no correlation between the x_i.)

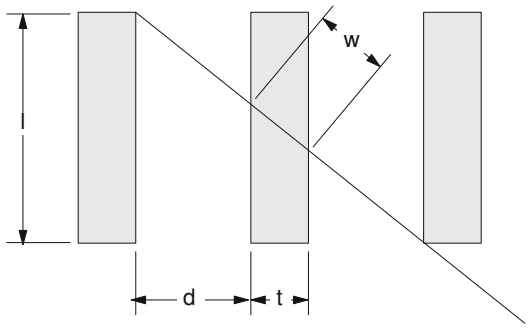
Problem 48. The interaction of a photon in a nuclear detector (an "event") initiates a process in the detector that lasts for a certain length of time. A second event occurring within a time τ of the first event is not recorded as a separate event. Suppose that the true counting rate is R_t. A counting rate R_o is observed.

(a) A nonparalyzable counting system is "dead" for a time τ after each recorded event. Additional events that occur during this dead time are not recorded but do not prolong the dead time. Show that $R_t = R_o(1 - R_o\tau)$ and $R_o = R_t/(1 + R_t\tau)$.

(b) A paralyzable counting system is unable to record a second event unless a time τ has passed since the last event. In other words, an event occurring during the dead time is not only not recorded, it prolongs the dead time. Show that in this case $R_o = R_t e^{-R_t\tau}$. (Hint: Use the Poisson distribution of Appendix J to find the fraction of events separated by a time greater than τ. The probability that the next event occurs between t and $t + dt$ is the probability of no event during time t multiplied by the probability of an event during dt.)

(c) Plot R_o vs R_t for the two cases when τ is fixed. The easiest way to do this is to plot $R_o\tau$ vs $R_t\tau$.

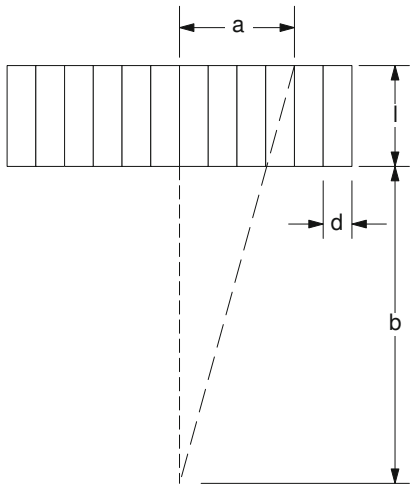
Problem 49. Two channels of a collimator for a gamma camera are shown in cross section, along with the path of a photon that encounters the minimum thickness of collimator septum (wall).



(a) Show that if $(d + t)/l \ll 1$, then $w/t = l/(2d + t)$.

(b) If transmission through the septum is to be less than 5 %, what is the relationship between t, d, l, and μ? Evaluate this for $^{99\text{m}}$Tc and for a positron emitter.

Problem 50. Photons from a point source a distance b below a collimator pass through channels out to a distance a from the perpendicular to the collimator passing through the source.



(a) Find an expression for a in terms of b, d, and l.

(b) Assume that a is related to the spatial frequency k for which the modulation transfer function (MTF) = 0.5 in Fig. 17.18 by $a = K/k$, where K is a constant. Calculate the thickness l of the collimator.

Problem 51. The collimator efficiency of a gamma camera is defined to be the fraction of the γ rays emitted isotropically by a point source that pass through the collimator into the scintillator.

(a) Consider a circular channel of diameter d in the collimator directly over the source. Show that the fraction of the photons striking the scintillator after passing through that channel is $d^2/16(l+b)^2$. (Assume that any which strike the septum are lost).

(b) Use the result of the previous problem to estimate the number of channels through which at least some photons from the point source pass. Assume that the fraction of collimator area that is occupied by channels rather than lead is $\left[d/(d+t)\right]^2$.

(c) Calculate the geometric efficiency g assuming that all channels that pass any photons have the same efficiency as the one on the perpendicular from the source. Show that it is of the form

$$g = K^2 \left(\frac{d}{l}\right)^2 \left(\frac{d}{d+t}\right)^2$$

and evaluate K. More detailed calculations show that K is about 0.24 for a hexagonal array of round holes and 0.26 for hexagonal holes.[11]

(d) How does the detector efficiency relate to the collimator resolution?

Section 17.9

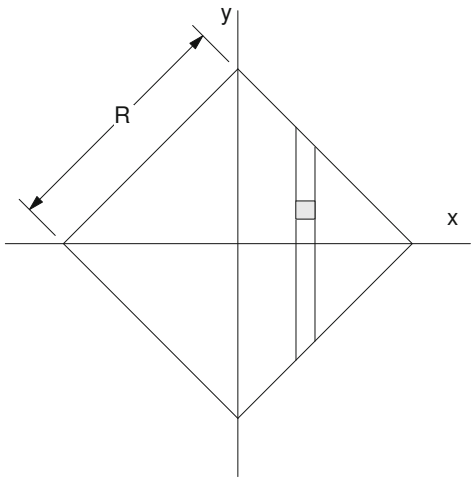
Problem 52.

(a) Derive Eq. 17.60 from Eq. 17.59.

(b) Calculate the limit of Eq. 17.60 when there is no attenuation.

Problem 53. The attenuation distortion for SPECT can be reduced by making measurements on opposite sides of the patient and taking the geometric mean. The geometric mean of variables x_1 and x_2 is $(x_1 x_2)^{1/2}$. Calculate the geometric mean of two SPECT measurements on opposite sides of the patient. Ignore possible $1/r^2$ effects.

Problem 54. Consider a radioactive source having a uniform activity per unit volume A_V and the square geometry shown below.

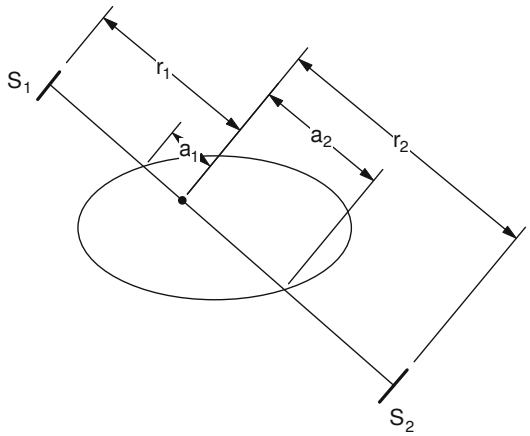


(a) Calculate the projection $F(x)$ including the effects of attenuation with coefficient μ.

(b) Plot $F(x)$ for $\mu = 0$ and for $\mu R = 3$.

Section 17.10

Problem 55. Suppose that A positrons are emitted from a point per second. They come to rest and annihilate within a short distance of their source. When a positron annihilates, two photons are emitted in opposite directions. Two photon detectors are set up on opposite sides of the source. The source is distance r_1 from the first detector, of area S_1, and r_2 from the second detector of area S_2. The area S_2 is large enough so that the second photon will definitely enter detector 2 if the first photon enters detector 1. Assume that both detectors count with 100 % efficiency.



(a) Show that the number of counts in the first detector would be $2AS_1/4\pi r_1^2$ if there were no attenuation between source and detector, and that it is $(2AS_1/4\pi r_1^2)e^{-\mu a_1}$ if attenuation in a thickness a_1 of the body is considered.

[11] Cherry et al. (2012, p. 222), Grenier et al. (1974).

(b) Detector 2 detects the second photon for every photon that strikes detector 1. Assuming a uniform attenuation coefficient and body thickness a_2, find an expression for the number of events in which both photons are detected.

Problem 56. Positron emission tomography relies on simultaneous detection of the back-to-back annihilation gamma rays (a *coincidence*). In addition to true coincidences, there can be "scatter coincidences" in which annihilation photons coming from a point that is not on the line between the two detectors enter both detectors. There can also be "random coincidences" which arise from photons from completely independent decays that occur nearly simultaneously. Consider a ring of detectors around a patient. Make three drawings showing true coincidences, scatter coincidences and random coincidences.

Section 17.12

Problem 57. The half-life of ^{235}U is 7×10^8 yr. The age of the earth is 4.5 billion years. What fraction of the ^{235}U that existed on the earth when it was first formed is present now?

Problem 58. There are three naturally-occurring decay series beginning with three long-lived isotopes: ^{238}U (Figs. 17.26 and 17.27), ^{235}U, and ^{232}Th. The ^{232}Th series begins with the α decay of ^{232}Th (half life $= 1.4 \times 10^{10}$ yr) to nucleus A which undergoes β^- decay to nucleus B which undergoes β^- decay to nucleus C which undergoes α decay to nucleus D which undergoes α decay to nucleus E, etc. Make a chart like Fig. 17.26 showing the first five steps in the series, and identify the five nuclei A–E.

Problem 59. One way to determine the age of biological remains is *carbon-14 dating*. The common isotope of carbon is stable ^{12}C. The rare isotope ^{14}C decays with a half-life of 5370 yr. ^{14}C is constantly created in the atmosphere by cosmic rays. The equilibrium between production and decay results in about 1 of every 10^{12} atoms of carbon in the atmosphere being ^{14}C, mostly as part of a CO_2 molecule. As long as the organism is alive, the ratio of ^{12}C to ^{14}C in the body is the same as in the atmosphere. Once the organism dies, it no longer incorporates ^{14}C from the atmosphere, and the number of ^{14}C nuclei begins to decrease. Suppose the remains of an organism have one ^{14}C for every 10^{13} ^{12}C nuclei. How long has it been since the organism died?

Problem 60. Consider a fictitious two-step decay series analogous to the more complex series shown in Fig. 17.26. The series starts with isotope A which decays at rate λ_1 to isotope B. Isotope B decays to isotope C with rate λ_2. Isotope C is stable.

(a) Derive the differential equation governing the number of nuclei N_A, N_B, and N_C. Where else in this chapter have you seen the same equations?

(b) Solve the differential equations using the initial conditions $N_A(0) = N$, $N_B(0) = N_C(0) = 0$. Make sure your solutions make sense for $t \to 0$ and $t \to \infty$.

(c) Find N_B/N_A in the limit $\lambda_2 \gg \lambda_1$. Ignore short times. Also find activities A_A and A_B. Explain physically how such a small number of nuclei N_B can contribute so much to the total activity.

References

Arqueros F, Montesinos GD (2003) A simple algorithm for the transport of gamma rays in a medium. Am J Phys 71(1):38–45

BEIR Report IV (1988) Committee on the biological effects of ionizing radiations. Health risks of radon and other internally deposited alpha-emitters. National Academy Press, Washington, DC

BEIR Report VI (1999) Committee on health risks of exposure to radon. Health effects of exposure to radon. National Academy Press, Washington, DC

Berger MJ (1968) Energy deposition in water by photons from point isotropic sources. NM/MIRD Pamphlet 2. Society of Nuclear Medicine, New York

Bolch WE, Eckerman KF, Sgouros G, Thomas SR (2009) MIRD pamphlet No. 21: a generalized schema for radiopharmaceutical dosimetry—standardization of nomenclature. J Nucl Med 50:477–484. doi:10.2967/jnumed.108.056036

Buchsbaum D, Wessels BW (1993) Introduction: radiolabeled antibody tumor dosimetry. Med Phys 20(2, Pt. 2):499–501

Cherry SR, Sorenson JA, Phelps ME (2012) Physics in nuclear medicine, 4th edn. Saunders, Philadelphia

Christian P, Waterstram-Rich KM (2012) Nuclear medicine and PET/CT: technology and techniques, 7th edn. Elsevier, St. Louis

Coffey JL, Watson EE (1979) Calculating dose from remaining body activity: a comparison of two methods. Med Phys 6(4):307–308

Coursey BM, Nath R (2000) Radionuclide therapy. Phys Today 53(4):25–30

Dewaraja YK, Frey EC, Sgouros G, Brill AB, Roberson P, Zanzonico PB, Ljungberg M (2012) MIRD pamphlet No. 23: quantitative SPECT for patient-specific 3-dimensional dosimetry in internal radionuclide therapy. J Nucl Med 53:1310–1325. doi:10.2967/jnumed.111.100123

Eckerman KF, Endo A (2008) MIRD: radionuclide data and decay schemes, 2nd edn. Society of Nuclear Medicine and Molecular Imaging, Reston

Eckerman KF, Westfall RJ, Ryman JC, Cristy M (1994) Availability of nuclear decay data in electronic form, including beta spectra not previously published. Health Phys 67(4):338–345

Eisberg R, Resnick R (1985) Quantum physics of atoms, molecules, solids, nuclei and particles, 2nd edn. Wiley, New York

Erhardt JC, Oberly LW, Cuevas JM (1978) Imaging ability of collimators in nuclear medicine. U.S. Dept. HEW, Rockville. Publ. No. (FDA)79-8077

Evans RD (1955) The atomic nucleus. McGraw-Hill, New York

Fox RA (2002) Intravascular brachytherapy of the coronary arteries. Phys Med Biol 47:R1–R30

Fritzberg AR, Wessels BW (1995) Therapeutic radionuclides. In: Wagner HN Jr, Szabo Z, Buchanan JW (eds) Principles of nuclear medicine, 2nd edn. Saunders, Philadelphia, pp 229–234

Grenier RP, Bender MA, Jones RH (1974) A computerized multicrystal scintillation gamma camera, vol. 2, Chap. 3. In: Hine HG, Sorenson JA (eds) Instrumentation in nuclear medicine. Academic, New York

Howell RW (1992) Radiation spectra for Auger-emitting radionuclides: report No. 2 of the AAPM nuclear medicine task group No. 6. Med Phys 19(6):1371–1383

Humm JL, Howell RW, Rao DV (1994) Dosimetry of Auger-electron-emitting radionuclides: report No. 3 of the AAPM nuclear medicine task group No. 6. Med Phys 12(12):1901–1915

Hunt JG, da Silva FC, Mauricio CL, dos Santos DS (2004) The validation of organ dose calculations using voxel phantoms and Monte Carlo methods applied to point and water immersion sources. Rad Prot Dosimetry 108(1):85–89

ICRU Report 67 (2002) Absorbed dose specification in nuclear medicine. J ICRU 2(1):1–113

Kaluza GL, Raizner AE (2004) Brachytherapy for restenosis after stenting for coronary artery disease: its role in the drug-eluting stent era. Curr Opin Cardiol 19:601–607

Kassis AI (2011) Molecular and cellular radiological effects of Auger emitting radionuclides. Radiat Prot Dosim 143:241–247

Khan FM (2010) The physics of radiation therapy, 4th edn. Wolters Kluwer, Philadelphia

Kowalsky RJ, Falen SW (2011) Radiopharmaceuticals in nuclear pharmacy and nuclear medicine, 3rd edn. American Pharmacists Association, Washington, DC

Links JM, Engdahl JC (1995) Planar imaging, Chap. 17. In: Wagner HN Jr, Szabo Z, Buchanan JW (eds) Principles of nuclear medicine, 2nd edn. Saunders, Philadelphia

Loevinger R, Budinger TF, Watson EE (1988) MIRD primer for absorbed dose calculations. Society of Nuclear Medicine and Molecular Imaging, New York

Muehllehner G, Karp JS (2006) Positron emission tomography. Phys Med Biol 51:R117–R137

Nag S (ed) (1994) High dose rate brachytherapy: a textbook. Futura, Armonk

Patterson JC, Mosley ML (2005) How available is positron emission tomography in the United States? Mol Imaging Biol 7(3):197–200

Rehm K, Strother SC, Anderson JR, Schaper KA, Rottenberg DA (1994) Display of merged multimodality brain images using interleaved pixels with independent color scales. J Nucl Med 35:1815–1821

Ruth TJ (2009) The uses of radiotracers in the life sciences. Rep Prog Phys 72:016701. dx.doi.org/10.1088/0034-4885/72/1/016701

Sastry KSR (1992) Biological effects of the Auger emitter iodine-125: a review. Report No. 1 of the AAPM Nuclear Medicine Task Group No. 6. Med Phys 19(6):1361–1370

Snyder WS, Ford MR, Warner GG, Watson SB (1975) S, Absorbed dose per unit cumulated activity for selected radionuclides and organs. NM/MIRD Pamphlet 11. Society of Nuclear Medicine and Molecular Imaging, New York

Snyder WS, Ford MR, Warner GG (1976) Specific absorbed fractions for radiation sources uniformly distributed in various organs of a heterogeneous phantom. NM/MIRD Pamphlet 5, revised. Society of Nuclear Medicine, New York

Snyder WS, Ford MR, Warner GG, Fisher HL (1978) Estimates of absorbed fractions for photon sources uniformly distributed in various organs of a heterogeneous phantom. NM/MIRD Pamphlet 5, revised. Society of Nuclear Medicine and Molecular Imaging, New York

Stabin MG (2008) Fundamentals of nuclear medicine dosimetry. Springer, New York

Stabin MG, da Luz LCQP (2002) Decay data for internal and external dose assessment. Health Phys 83(4):471–475

Stabin MG, Sparks RB, Crowe E (2005) OLINDA/EXM: the second-generation personal computer software for internal dose assessment in nuclear medicine. J Nucl Med 46(6):1023–1027

Strother SC, Anderson JR Jr, Schaper KA, Sidtis JJ, Liow JS, Woods RP, Rottenberg DA (1995) Principal component analysis and the scaled subprofile model compared to intersubject averaging and statistical parametric mapping: I. Functional connectivity of the human motor system studied with ^{15}O-water PET. J Cerebr Blood F Met 15(5):738–753

Watson EE, Stabin MG, Siegel JA (1993) MIRD formulation. Med Phys 20(2, Pt. 2):511–514

Williams LE (2008) Anniversary paper: nuclear medicine: fifty years and still counting. Med Phys 35:3020–3029

Zanzonico P (2004) Positron emission tomography: a review of basic principles, scanner design and performance, and current systems. Semin Nucl Med 34(2):87–111

Zanzonico P (2012) Principles of nuclear medicine imaging: planar, SPECT, PET, multi-modality and autoradiography systems. Rad Res 177:349–364. doi:10.1667/RR2577.1

Zanzonico PB, Brill AB, Becker DV (1995) Radiation dosimetry, Chap. 9. In: Wagner HN Jr, Szabo Z, Buchanan JW (eds) Principles of nuclear medicine, 2nd edn. Saunders, Philadelphia

Magnetic Resonance Imaging

18

Magnetic resonance imaging (MRI) (formerly called nuclear magnetic resonance imaging) provides very-high-resolution images without ionizing radiation. There is also the potential for more elaborate imaging, including flow, diffusion, and the signature of particular atomic environments.

Magnetic resonance phenomena are more complicated than x-ray attenuation or photon emission by a radioactive nucleus. MRI depends upon the behavior of atomic nuclei in a magnetic field; in particular, the orientation and motion of the nuclear magnetic moment in the field. The patient is placed in a strong static magnetic field (typically 1–4 T). This is usually provided by a hollow cylindrical (solenoidal) magnet, though some machines use other configurations so that the physician can carry out procedures on the patient while viewing the MRI image. Other coils apply time-varying spatial gradients to the magnetic field, along with radio-frequency signals that cause the magnetization changes described below. Still other coils detect the very weak radio-frequency signals resulting from these changes.

First, we must understand the property that we are measuring. Section 18.1 describes the behavior of a magnetic moment in a static magnetic field, and Sect. 18.2 shows how the nuclear spin is related to the magnetic moment. Section 18.3 introduces the concept of the magnetization vector, which is the magnetic moment per unit volume, while Sect. 18.4 develops the equations of motion for the magnetic moment. In order to describe the motion of the magnetization, it is convenient—in fact, almost essential—to use the rotating coordinate system described in Sect. 18.5.

To make a measurement, the nuclear magnetic moments originally aligned with the static magnetic field are made to rotate or *precess* in a plane perpendicular to the static field, after which the magnetization gradually returns to its original value. This relaxation phenomenon is described in Sect. 18.6. Sections 18.7 and 18.8 describe ways in which the magnetization can be manipulated for measurement.

Imaging techniques are finally introduced in Sect. 18.9. Sections 18.10 and 18.11 describe how chemical shifts and blood flow can affect the image or can themselves be imaged. The last two sections describe functional MRI (fMRI) and diffusion effects.

18.1 Magnetic Moments in an External Magnetic Field

Magnetic resonance imaging detects the magnetic dipoles in the nuclei of atoms in the human body. We saw in Chap. 8 that isolated magnetic monopoles have never been observed (see Eq. 8.8), and that magnetic fields are produced by moving charges or electric currents. In some cases, such as bar magnets, the external field is the same as if there were magnetic charges occurring in pairs or *dipoles*.[1] The strength of a dipole is measured by its *magnetic dipole moment* $\boldsymbol{\mu}$. (In Chap. 8 the magnetic dipole moment was called $\mathbf{m}$ to avoid confusion with μ_0. In this chapter we use $\boldsymbol{\mu}$ to avoid confusion with the quantum number m and to be consistent with the literature in the field.) The magnetic dipole moment is analogous to the electric dipole moment of Chap. 7; however, it is produced by a movement of charge, such as charge moving in a circular path. The units of $\boldsymbol{\mu}$ are J T^{-1} or A m^2. We saw that when a magnetic dipole is placed in a magnetic field as in Fig. 18.1, it is necessary to apply an external torque $\boldsymbol{\tau}_{\text{ext}}$ to keep it in equilibrium. This torque, which is required to cancel the torque exerted by the magnetic field, vanishes if the dipole is aligned with the field. The torque exerted on the dipole by the magnetic field $\mathbf{B}$ is

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B}. \tag{18.1}$$

(This is Eq. 8.4.)

[1] Dipoles can be arranged so that their fields nearly cancel, giving rise to still-higher-order moments such as the quadrupole moment or the octupole moment (see Chap. 7). A configuration for which the quadrupole moment is important is two magnets in a line arranged as N-S-S-N.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1_18, © Springer International Publishing Switzerland 2015

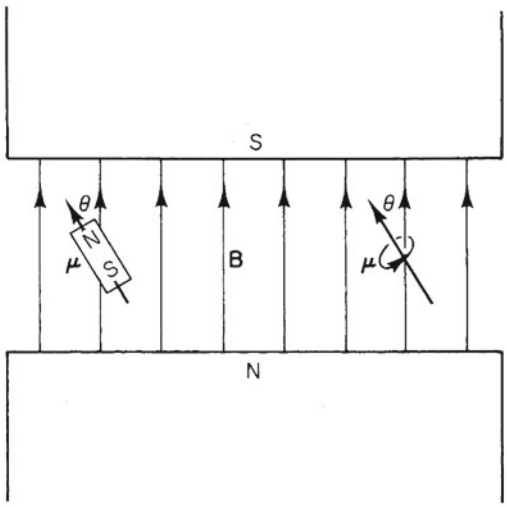


Fig. 18.1 A magnetic dipole in a magnetic field. The dipole can be either a bar magnet or a current loop

The potential energy of the dipole is the work that must be done by $\boldsymbol{\tau}_{\text{ext}}$ to change the dipole's orientation in the magnetic field without changing any kinetic energy it might have. To increase angle θ by an amount $d\theta$ requires that work be done on the dipole-magnetic field system. This work is the increase in potential energy of the system:

$$dU = \mu B \sin\theta \, d\theta. \tag{18.2}$$

This can be integrated to give the change in potential energy when the angle changes from θ_1 to θ_2:

$$U(\theta_2) - U(\theta_1) = -\mu B(\cos\theta_2 - \cos\theta_1).$$

If the energy is considered to be zero when the dipole is at right angles to **B**, then the potential energy is

$$U(\theta) = -\mu B \cos\theta = -\boldsymbol{\mu} \cdot \mathbf{B}. \tag{18.3}$$

In many cases the moving charges that give rise to the magnetic moment of an object possess angular momentum **L**. Often the magnetic moment is parallel to and proportional to the angular momentum: $\boldsymbol{\mu} = \gamma \mathbf{L}$. The proportionality factor γ is called the *gyromagnetic ratio* (sometimes called the magnetogyric ratio). When such an object is placed in a uniform magnetic field, the resulting motion can be quite complicated. The torque on the object is $\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} = \gamma \mathbf{L} \times \mathbf{B}$. It is not difficult to show (Problem 1) that the torque is the rate of change of the angular momentum, $\boldsymbol{\tau} = d\mathbf{L}/dt$. Therefore the equation of motion is

$$\gamma(\mathbf{L} \times \mathbf{B}) = \frac{d\mathbf{L}}{dt} \tag{18.4a}$$

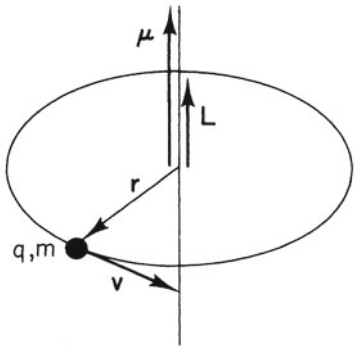


Fig. 18.2 A particle of charge q and mass m travels in a circular orbit. It has a magnetic moment $\boldsymbol{\mu}$ and angular momentum **L**. If the charge is positive, $\boldsymbol{\mu}$ and **L** are parallel; if it is negative they are in opposite directions

or

$$\gamma(\boldsymbol{\mu} \times \mathbf{B}) = \frac{d\boldsymbol{\mu}}{dt}. \tag{18.4b}$$

Solutions to these equations are discussed in Sect. 18.4.

18.2 The Source of the Magnetic Moment

Atomic electrons and the protons and neutrons in the atomic nucleus can possess both angular momentum and a magnetic moment. The magnetic moment of a particle is related to its angular momentum. We can derive this relationship for a charged particle moving in a circular orbit. We saw in Chap. 8 that the magnitude of the magnetic moment of a current loop is the product of the current i and the area of the loop S:

$$|\boldsymbol{\mu}| = \mu = iS. \tag{18.5}$$

The direction of the vector is perpendicular to the plane of the loop. Its direction is defined by a right-hand rule: curl the fingers of your right hand in the direction of current flow and your thumb will point in the direction of $\boldsymbol{\mu}$ (see the right-hand part of Fig. 18.1). This is the same right-hand rule that relates the circular motion of a particle to the direction of its angular momentum.

Suppose that a particle of charge q and mass m moves in a circular orbit of radius r as in Fig. 18.2. The speed is v and the magnitude of the angular momentum is $L = mvr$. The effective current is the charge q multiplied by the number of times it goes past a given point on the circumference of the orbit in one second: $i = qv/2\pi r$. The magnetic moment has magnitude $\mu = iS = i\pi r^2 = qvr/2$. Since the angular momentum is $L = mvr$ and $\boldsymbol{\mu}$ and **L** are both perpendicular

Table 18.1 Values of the spin and gyromagnetic ratio for a free electron and various nuclei of interest

Particle	Spin	$\gamma = \omega_{\text{Larmor}}/B$ ($\text{s}^{-1}\ \text{T}^{-1}$)	ν/B ($\text{MHz}\ \text{T}^{-1}$)
Electron	$\frac{1}{2}$	1.7608×10^{11}	2.8025×10^{4}
Proton	$\frac{1}{2}$	2.6753×10^{8}	42.5781
Neutron	$\frac{1}{2}$	1.8326×10^{8}	29.1667
^{23}Na	$\frac{3}{2}$	0.7076×10^{8}	11.2618
^{31}P	$\frac{1}{2}$	1.0829×10^{8}	17.2349

to the plane of the orbit, we can write

$$\boldsymbol{\mu} = \left(\frac{q}{2m}\right)\mathbf{L} = \gamma\,\mathbf{L}. \tag{18.6}$$

The quantity $\gamma = q/2m$ is the gyromagnetic ratio for this system. The units γ of are $\text{T}^{-1}\ \text{s}^{-1}$ (see Problem 2). The magnetic moment and the orbital angular momentum are parallel for a positive charge and antiparallel for a negative charge.

An electron or a proton also has an intrinsic magnetic moment quite separate from its orbital motion. It is associated with and proportional to the intrinsic or *spin* angular momentum **S** of the particle. We write

$$\boldsymbol{\mu} = \gamma\,\mathbf{S}. \tag{18.7}$$

The value of γ for a spin is *not* equal to $q/2m$.

Two kinds of spin measurements have biological importance. One is associated with electron magnetic moments and the other with the magnetic moments of nuclei. Most neutral atoms in their ground state have no magnetic moment due to the electrons. Exceptions are the transition elements that exhibit paramagnetism. Free radicals, which are often of biological interest, have an unpaired electron and therefore have a magnetic moment. In most cases this magnetic moment is due almost entirely to the spin of the unpaired electron.

Magnetic resonance imaging is based on the magnetic moments of atomic nuclei in the patient. The total angular momentum and magnetic moment of an atomic nucleus are due to the spins of the protons and neutrons, as well as any orbital angular momentum they have inside the nucleus. Table 18.1 lists the spin and gyromagnetic ratio of the electron and some nuclei of biological interest.

If the nuclear angular momentum is **I** with quantum number I, the possible values of the z component of **I** are $m\hbar$, where $m = -I, (-I+1), \ldots, I$. For $I = \frac{1}{2}$, the values are $-1/2$ and $1/2$, while for $I = \frac{3}{2}$ they are $-3/2, -1/2, 1/2$, and $3/2$. The direction of the external magnetic field defines the z axis, and the energy of a spin is given by $-\boldsymbol{\mu}\cdot\mathbf{B} = -\gamma\,\mathbf{I}\cdot\mathbf{B} = -\gamma m\hbar B$. The difference between adjacent energy levels is $\gamma B\hbar$, and the angular frequency of a photon corresponding to that difference is $\omega_{\text{photon}} = \gamma B$.

18.3 The Magnetization

The MR image depends on the *magnetization* of the tissue. The magnetization of a sample, **M**, is the average magnetic moment per unit volume. In the absence of an external magnetic field to align the nuclear spins, the magnetization is zero. As an external magnetic field **B** is applied, the spins tend to align in spite of their thermal motion, and the magnetization increases, proportional at first to the external field. If the external field is strong enough, all of the nuclear magnetic moments are aligned, and the magnetization reaches its saturation value.

We can calculate how the magnetization depends on **B**. Consider a collection of spins of a single nuclear species in an external magnetic field. This might be the hydrogen nuclei (protons) in a sample. The spins do not interact with each other but are in thermal equilibrium with the surroundings, which are at temperature T. We do not consider the mechanism by which they reach thermal equilibrium. Since the magnetization is the average magnetic moment per unit volume, it is the number of spins per unit volume, N, times the average magnetic moment of each spin: $\mathbf{M} = N\,\langle\boldsymbol{\mu}\rangle$.

To obtain the average value of the z component of the magnetic moment, we must consider each possible value of quantum number m. We multiply the value of μ_z corresponding to each value of m by the probability that m has that value. Since the spins are in thermal equilibrium with the surroundings, the probability is proportional to the Boltzmann factor of Chap. 3, $e^{-(U/k_BT)} = e^{\gamma m\hbar B/k_BT}$. The denominator in Eq. 18.8 normalizes the probability:

$$\langle\mu_z\rangle = \frac{\gamma\hbar\sum_{m=-I}^{I} m\,e^{\gamma m\hbar B/k_BT}}{\sum_{m=-I}^{I} e^{\gamma m\hbar B/k_BT}}. \tag{18.8}$$

At room temperature $\gamma I\hbar B/k_BT \ll 1$ (see Problem 4), and it is possible to make the approximation $e^x \approx 1+x$. The sum in the numerator then has two terms:

$$\sum_{m=-I}^{I} m + \frac{\gamma\hbar B}{k_BT}\sum_{m=-I}^{I} m^2.$$

The first sum vanishes. The second is $I(I+1)(2I+1)/3$. The denominator is

$$\sum_{m=-I}^{I} 1 + \frac{\gamma\hbar B}{k_BT}\sum_{m=-I}^{I} m.$$

The first term is $2I+1$; the second vanishes. Therefore we obtain

$$\langle\mu_z\rangle = \frac{\gamma^2\hbar^2 I(I+1)}{3k_BT}B. \tag{18.9}$$

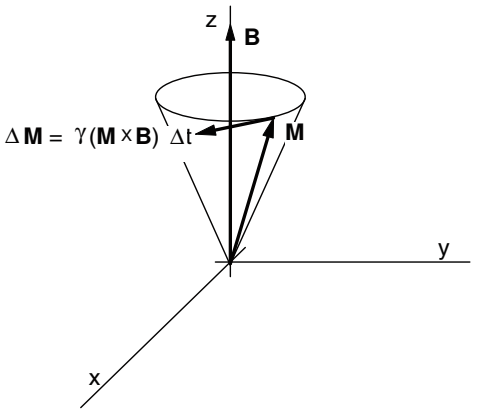


Fig. 18.3 The system with initial magnetization **M** has been given just enough additional angular momentum to precess about the direction of the static magnetic field **B**. The rate of change of **M** is perpendicular to both **M** and **B**. For short time intervals, $\Delta\mathbf{M} = \gamma(\mathbf{M} \times \mathbf{B})\,\Delta t$

The z component of **M** is

$$M_z = N\,\langle\mu_z\rangle = \frac{N\gamma^2\hbar^2 I(I+1)}{3k_B T}B, \tag{18.10}$$

which is proportional to the applied field.

18.4 Behavior of the Magnetization Vector

A remarkable result of quantum mechanics is that the average or expectation value of a spin obeys the classical Eq. 18.4b:

$$\frac{d\,\langle\boldsymbol{\mu}\rangle}{dt} = \gamma\,(\langle\boldsymbol{\mu}\rangle \times \mathbf{B}) \tag{18.11}$$

whether or not **B** is time dependent (Slichter 1990). Multiplying by the number of spins per unit volume we obtain

$$\frac{d\mathbf{M}}{dt} = \gamma\,(\mathbf{M} \times \mathbf{B})\,. \tag{18.12}$$

This equation can lead to many different behaviors of **M**, some of which are quite complicated.

The simplest motion occurs if **M** is parallel to **B**, in which case **M** does not change because there is no torque. Another relatively simple motion, called *precession*, is shown in Fig. 18.3. With the proper initial conditions **M** (and $\langle\boldsymbol{\mu}\rangle$) precess about the direction of **B**. That is, they both rotate about the direction of **B** with a constant angular velocity and at a fixed angle θ with the direction of **B**. Since $\mathbf{M} \times \mathbf{B}$ is always at right angles to **M**, $d\mathbf{M}/dt$ is at right angles to **M**, and the angular momentum does not change magnitude. The analytic solution can be investigated by writing Eq. 18.12 in Cartesian coordinates when **B** is along the z axis:

$$\begin{aligned}\frac{dM_x}{dt} &= \gamma M_y B_z,\\ \frac{dM_y}{dt} &= -\gamma M_x B_z,\\ \frac{dM_z}{dt} &= 0.\end{aligned} \tag{18.13}$$

One possible solution to these equations is

$$\begin{aligned}M_z &= M_\parallel = \text{const},\\ M_x &= M_\perp \cos(-\omega t),\\ M_y &= M_\perp \sin(-\omega t).\end{aligned} \tag{18.14}$$

You can verify that these are a solution for arbitrary values of $M_\perp$ and $M_\parallel$ as long as $\omega = \omega_0 = \gamma B_z$. This is called the *Larmor precession frequency*. The minus sign means that for positive γ the rotation is clockwise in the xy plane. The classical Larmor frequency is equal to the frequency of photons corresponding to the energy difference given by successive values of $\boldsymbol{\mu} \cdot \mathbf{B}$. For this solution the initial values of **M** at $t = 0$ are $M_x(0) = M_\perp$, $M_y(0) = 0$, and $M_z(0) = M_\parallel$.

We need to modify the equation of motion, Eq. 18.12, to include changes in **M** that occur because of effects other than the magnetic field. Suppose that **M** has somehow been changed so that it no longer points along the z axis with the equilibrium value given by Eq. 18.10. Thermal agitation will change the populations of the levels so that M_z returns to the equilibrium value, which we call M_0. We *postulate* that the rate of exchange of energy with the reservoir is proportional to how far the value of M_z is from equilibrium:

$$\frac{dM_z}{dt} = \frac{1}{T_1}\,(M_0 - M_z)\,.$$

The quantity T_1, which is the inverse of the proportionality constant, is called the *longitudinal relaxation time* or *spin–lattice relaxation time*.

We also *postulate* an exponential disappearance of the x and y components of **M** with a *transverse relaxation time* T_2 (sometimes called the *spin–spin relaxation time*). (This assumption is often not a good one. For example, the decay of M_x and M_y in ice is more nearly Gaussian than exponential.) The equations are

$$\frac{dM_x}{dt} = -\frac{M_x}{T_2}, \qquad \frac{dM_y}{dt} = -\frac{M_y}{T_2}.$$

The transverse relaxation time is always shorter than T_1. Here is why. A change of M_z requires an exchange of energy with the reservoir. This is not necessary for changes confined to the xy plane, since the potential energy $(\boldsymbol{\mu} \cdot \mathbf{B})$ does not change in that case. M_x and M_y can change as M_z

changes, but they can also change by other mechanisms, such as when individual spins precess at slightly different frequencies, a process known as *dephasing*. The angular velocity of precession can be slightly different for different nuclear spins because of local variations in the static magnetic field; the angular velocity can also fluctuate as the field fluctuates with time. These variations and fluctuations are caused by neighboring atomic or nuclear magnetic moments or by inhomogeneities in the external magnetic field. Figure 18.4 shows how dephasing occurs if several magnetic moments precess at different rates.

Combining these approximate equations for relaxation in the absence of an applied magnetic field with Eq. 18.12 for the effect of a magnetic field gives the *Bloch equations*:[2]

$$\begin{aligned} \frac{dM_z}{dt} &= \frac{1}{T_1}(M_0 - M_z) + \gamma\,(\mathbf{M}\times\mathbf{B})_z\,, \\ \frac{dM_x}{dt} &= -\frac{M_x}{T_2} + \gamma\,(\mathbf{M}\times\mathbf{B})_x\,, \\ \frac{dM_y}{dt} &= -\frac{M_y}{T_2} + \gamma\,(\mathbf{M}\times\mathbf{B})_y\,. \end{aligned} \tag{18.15}$$

While these equations are not rigorous and there is no reason for the relaxation to be strictly exponential, they have proven to be quite useful in explaining many facets of nuclear spin magnetic resonance.

One can demonstrate by direct substitution the following solution to Eqs. 18.15 for a static magnetic field **B** along the z axis:

$$\begin{aligned} M_x &= M_0 e^{-t/T_2}\cos(-\omega_0 t), \\ M_y &= M_0 e^{-t/T_2}\sin(-\omega_0 t), \\ M_z &= M_0(1 - e^{-t/T_1}), \end{aligned} \tag{18.16}$$

where $\omega_0 = \gamma B$. This solution corresponds to what happens if **M** is somehow made to precess in the xy plane. (We will see how to accomplish this in Sect. 18.5.) The magnetization in the xy plane is initially M_0, and the amplitude decays exponentially with time constant T_2. The initial value of M_z is zero, and it decays back to M_0 with time constant T_1. A perspective plot of the trajectory of the tip of vector **M** is shown in Fig. 18.5.

[2] Felix Bloch and Edward Purcell shared the 1952 Nobel Prize in physics for their discovery of nuclear magnetic resonance.

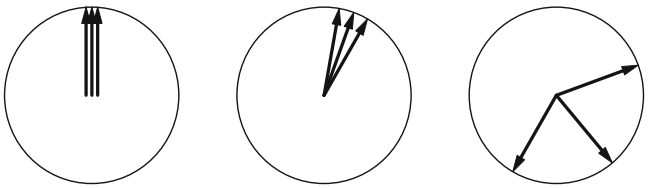


Fig. 18.4 If several spins precess in the xy plane at slightly different rates, the total spin amplitude decreases due to dephasing

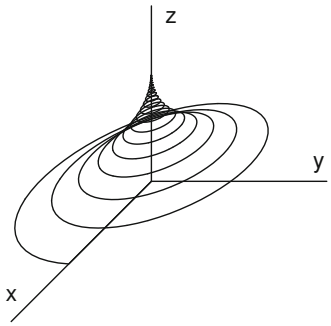


Fig. 18.5 The locus of the tip of the magnetization **M** when it relaxes according to Eqs. 18.16

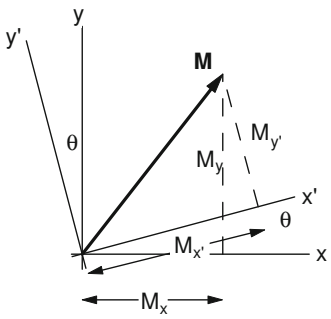


Fig. 18.6 The vector **M** can be represented by components along x and y or along x' and y'

18.5 A Rotating Coordinate System

18.5.1 Transforming to the Rotating Coordinate System

It is *much* easier to describe the motion of **M** in a coordinate system that is rotating about the z axis at the Larmor frequency. Figure 18.6 shows a vector **M** and two coordinate systems, xy and $x'y'$. The z component of **M** is unchanged. By considering the other components in Fig. 18.6, we see that

$$\begin{aligned} M_x &= M_{x'}\cos\theta - M_{y'}\sin\theta, \\ M_y &= M_{x'}\sin\theta + M_{y'}\cos\theta, \\ M_z &= M_{z'}. \end{aligned} \tag{18.17a}$$

This can also be written in matrix form. A rotation through angle θ around the z axis gives $\mathbf{M} = \mathbf{R}\mathbf{M}'$ or

$$\begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} M_{x'} \\ M_{y'} \\ M_{z'} \end{pmatrix}. \tag{18.17b}$$

Rotations about the other axes are discussed in Problem 12. Note that rotating the coordinate system that describes a fixed vector is equivalent to rotating the vector in the opposite direction, so the results quoted in Problem 12 apply to both situations.

For a three-dimensional coordinate system rotating clockwise around the z axis, $\theta = -\omega t$, the z-component of $\mathbf{M}$ is unchanged, and the transformation equations are

$$\begin{aligned} M_x &= M_{x'}\cos(-\omega t) - M_{y'}\sin(-\omega t), \\ M_y &= M_{x'}\sin(-\omega t) + M_{y'}\cos(-\omega t), \\ M_z &= M_{z'}. \end{aligned} \tag{18.18}$$

The time derivative of $\mathbf{M}$ is obtained by differentiating each component and remembering that $\mathbf{M}'$ can also depend on t:

$$\begin{aligned} \frac{dM_x}{dt} &= \frac{dM_{x'}}{dt}\cos(-\omega t) - \frac{dM_{y'}}{dt}\sin(-\omega t) \\ &\quad + \omega M_{x'}\sin(-\omega t) + \omega M_{y'}\cos(-\omega t), \\ \frac{dM_y}{dt} &= \frac{dM_{x'}}{dt}\sin(-\omega t) + \frac{dM_{y'}}{dt}\cos(-\omega t) \\ &\quad - \omega M_{x'}\cos(-\omega t) + \omega M_{y'}\sin(-\omega t), \\ \frac{dM_z}{dt} &= \frac{dM_{z'}}{dt}. \end{aligned} \tag{18.19}$$

We can use these expressions to write the equations of motion in the rotating frame. First consider a system without relaxation effects and with a static field B_z along the z axis. We will show that the components of $\mathbf{M}$ in a system rotating at the Larmor frequency are constant. The equations of motion are given in Eqs. 18.13. In terms of variables in the rotating frame, the equation for dM_x/dt becomes

$$\begin{aligned} &\frac{dM_{x'}}{dt}\cos(-\omega t) - \frac{dM_{y'}}{dt}\sin(-\omega t) + \omega M_{x'}\sin(-\omega t) \\ &\qquad + \omega M_{y'}\cos(-\omega t) \\ &\quad = \gamma\left[M_{x'}\sin(-\omega t) + M_{y'}\cos(-\omega t)\right]B_z. \end{aligned}$$

If the frame rotates at the Larmor frequency $\omega_0 = \gamma B_z$, the third and fourth terms on the left are equal to the right-hand side. The equation becomes

$$\frac{dM_{x'}}{dt}\cos(-\omega_0 t) - \frac{dM_{y'}}{dt}\sin(-\omega_0 t) = 0.$$

Under the same circumstances, the equation for dM_y/dt gives

$$\frac{dM_{x'}}{dt}\sin(-\omega_0 t) + \frac{dM_{y'}}{dt}\cos(-\omega_0 t) = 0.$$

Solving these simultaneously shows that $dM_{x'}/dt = 0$ and $dM_{y'}/dt = 0$. Therefore, in the rotating system $M_{x'}$ and $M_{y'}$ are constant. Equation 18.13 showed that $M_{z'}$ is constant, so the components of $\mathbf{M}$ are constant in the frame rotating at the Larmor frequency. Using Eqs. 18.18 to transform back to the laboratory system gives the solution Eq. 18.14.[3]

18.5.2 An Additional Oscillating Field

The next problem we consider in the rotating coordinate system is the addition of an oscillating magnetic field $B_1\cos(\omega t)$ along the x axis, fixed in the laboratory system. We will show that if the applied field is at the Larmor frequency, the equations of motion in the rotating system, Eqs. 18.25, are quite simple but very important. They are given below.

They are derived as follows. From the x component of Eq. 18.12, and remembering that $B_y = 0$,

$$\frac{dM_x}{dt} = \gamma M_y B_z,$$

we obtain (remembering that the $x'y'$ system is rotating at the Larmor frequency ω_0)

$$\begin{aligned} &\frac{dM_{x'}}{dt}\cos(-\omega_0 t) - \frac{dM_{y'}}{dt}\sin(-\omega_0 t) \\ &\quad + \omega_0 M_{x'}\sin(-\omega_0 t) + \omega_0 M_{y'}\cos(-\omega_0 t) \\ &\quad = \gamma B_z\left[M_{x'}\sin(-\omega_0 t) + M_{y'}\cos(-\omega_0 t)\right]. \end{aligned}$$

Since $\omega_0 = \gamma B_z$, the last two terms on the left cancel the terms on the right, leaving

$$\frac{dM_{x'}}{dt}\cos(-\omega_0 t) - \frac{dM_{y'}}{dt}\sin(-\omega_0 t) = 0. \tag{18.20}$$

[3] For those familiar with vector analysis, the general relationship between the time derivative of any vector $\mathbf{M}$ in the laboratory system and a system rotating with angular velocity $\boldsymbol{\Omega}$ is

$$\left(\frac{d\mathbf{M}}{dt}\right)_{\text{lab}} = \left(\frac{\partial\mathbf{M}}{\partial t}\right)_{\text{rot}} + \boldsymbol{\Omega}\times\mathbf{M}.$$

This can be applied to the magnetization combined with Eq. 18.12 to give

$$\left(\frac{\partial\mathbf{M}}{\partial t}\right)_{\text{rot}} = \gamma(\mathbf{M}\times\mathbf{B}) - \boldsymbol{\Omega}\times\mathbf{M} = \gamma\mathbf{M}\times\left(\mathbf{B} + \frac{\boldsymbol{\Omega}}{\gamma}\right),$$

which vanishes if $\gamma\mathbf{B} = -\boldsymbol{\Omega}$.

Similarly, the y-component of Eq. 18.12,

$$\frac{dM_y}{dt} = \gamma(M_z B_x - M_x B_z),$$

transforms and reduces to (remembering that $M_z = M_{z'}$)

$$\frac{dM_{x'}}{dt}\sin(-\omega_0 t) + \frac{dM_{y'}}{dt}\cos(-\omega_0 t) = \gamma B_1 M_{z'}\cos(-\omega t). \tag{18.21}$$

The z-component of Eq. 18.12 is, with $B_y = 0$

$$\frac{dM_z}{dt} = -\gamma M_y B_x,$$

which transforms to

$$\frac{dM_{z'}}{dt} = -\gamma B_1 M_{x'}\cos(\omega t)\sin(-\omega_0 t) - \gamma B_1 M_{y'}\cos(\omega t)\cos(-\omega_0 t). \tag{18.22}$$

It is possible to eliminate $M_{x'}$ from Eqs. 18.20 and 18.21 by multiplying Eq. 18.20 by $-\sin(-\omega_0 t)$, multiplying Eq. 18.21 by $\cos(-\omega_0 t)$, and adding. The result is

$$\frac{dM_{y'}}{dt} = \gamma B_1 M_{z'}\cos(\omega t)\cos(-\omega_0 t). \tag{18.23}$$

A similar technique can be used to eliminate $M_{y'}$ from these two equations:

$$\frac{dM_{x'}}{dt} = \gamma B_1 M_{z'}\cos(\omega t)\sin(-\omega_0 t). \tag{18.24}$$

18.5.3 Nutation

Equations 18.22–18.24 are the equations of motion for the components of **M** in the rotating system. If $\omega \neq \omega_0$, the motion is complicated, but averaged over many Larmor periods the right-hand side of each equation is zero. If the applied field oscillates at the Larmor frequency, $\omega = \omega_0$, then the $\cos^2(-\omega_0 t)$ factors average to $\frac{1}{2}$ while factors like $\sin(-\omega_0 t)\cos(-\omega_0 t)$ average to zero.

The averaged equations are a very important result:

$$\frac{dM_{x'}}{dt} = 0, \tag{18.25a}$$

$$\frac{dM_{y'}}{dt} = \frac{\gamma B_1}{2} M_{z'}, \tag{18.25b}$$

$$\frac{dM_{z'}}{dt} = -\frac{\gamma B_1}{2} M_{y'}. \tag{18.25c}$$

The first equation says that if $M_{x'}$ is initially zero, it remains zero. Let us define a new angular frequency

$$\omega_1 = \frac{\gamma B_1}{2}. \tag{18.26}$$

It is the frequency of rotation caused by B_1 oscillating at the Larmor frequency. It is much lower than the Larmor frequency because $B_1 \ll B_z$. In terms of ω_1, Eqs. 18.25b and 18.25c become

$$\frac{dM_{z'}}{dt} = -\omega_1 M_{y'}, \qquad \frac{dM_{y'}}{dt} = \omega_1 M_{z'}.$$

These are a pair of coupled linear differential equations with constant coefficients. Differentiating one and substituting it in the other gives

$$\frac{d^2 M_{z'}}{dt^2} = -\omega_1 \frac{dM_{y'}}{dt} = -\omega_1^2 M_{z'}, \tag{18.27}$$

which has a solution (a and b are constants of integration)

$$M_{z'} = a\sin(\omega_1 t) + b\cos(\omega_1 t). \tag{18.28}$$

From Eq. 18.25c we get

$$M_{y'} = -\frac{1}{\omega_1}\frac{dM_{z'}}{dt} = -a\cos(\omega_1 t) + b\sin(\omega_1 t). \tag{18.29}$$

The values of a and b are determined from the initial conditions. For example, if **M** is initially along the z axis, $a = 0$ and $b = M_0$. Then

$$\begin{aligned} M_{x'} &= 0, \\ M_{y'} &= M_0\sin(\omega_1 t), \\ M_{z'} &= M_0\cos(\omega_1 t). \end{aligned} \tag{18.30}$$

This kind of motion—precession about the z axis combined with a change of the projection of **M** on z—is called *nutation*.

18.5.4 π and π/2 Pulses

From Eqs. 18.30 it is easy to see that turning B_1 on for a quarter of a period of ω_1 (a 90° pulse or $\pi/2$ *pulse*, $t = T/4 = \pi/2\omega_1$) nutates **M** into the $x'y'$ plane, while a 180° or π *pulse* nutates **M** to point along the $-z$ axis. **M** nutates about the rotating x' axis. Shifting the phase of B_1 changes the axis in the $x'y'$ plane about which **M** nutates. It may seem strange that an oscillating magnetic field pointing along an axis fixed in the laboratory frame causes rotation about an axis in the rotating frame. The reason is that B_1 is also oscillating at the Larmor frequency, so that its amplitude changes in just the right way to cause this behavior of **M**. Figures 18.7 and 18.8 show this nutation in both the rotating frame and the laboratory frame for a $\pi/2$ pulse and a π pulse.

Figure 18.7c emphasizes the difference between nutation and relaxation by plotting M_z vs. the projection of **M** in the $x'y'$ plane. For nutation the components of **M** are given by

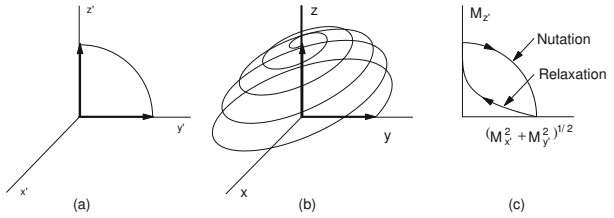


Fig. 18.7 The locus of the tip of the magnetization **M** when an oscillating magnetic field B_1 is applied for a time t such that $\omega_1 t = \pi/2$. This is often called a "$\pi/2$" pulse. **a** The rotating frame. **b** The laboratory frame. **c** Plots of $M_{z'}$ vs $(M_{x'}^2 + M_{y'}^2)^{1/2}$ showing the difference between nutation and relaxation

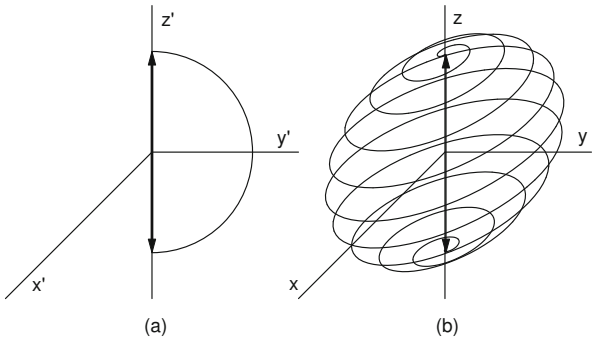


Fig. 18.8 A "π pulse" B_1 is applied for a time $t = \pi/\omega_1$ and rotates **M** to point along the $-z$ axis. **a** The rotating frame. **b** The laboratory frame

Eqs. 18.30, the magnitude of **M** is unchanged, and the locus is a circle. For relaxation the components are given by Eqs. 18.16.

Another interesting solution is one for which the initial value of **M** is

$$M_{x'}(0) = M_0 \cos\alpha,$$

$$M_{y'}(0) = M_0 \sin\alpha,$$

$$M_{z'}(0) = 0.$$

This corresponds to an **M** that has already been nutated into the $x'y'$ plane. Substituting these values in Eqs. 18.28 and 18.29 shows that $b = 0$ and $a = M_0 \sin\alpha$. Then the solution is

$$M_{x'}(t) = M_0 \cos\alpha,$$

$$M_{y'}(t) = M_0 \sin\alpha \cos(\omega_1 t), \tag{18.31}$$

$$M_{z'}(t) = -M_0 \sin\alpha \sin(\omega_1 t).$$

This solution is plotted in Fig. 18.9 in both the rotating frame and the laboratory frame for the case of a π pulse (a pulse of duration π/ω_1). The effect is to nutate **M** about the x' axis in the rotating coordinate system. We will see later that this is a very useful pulse.

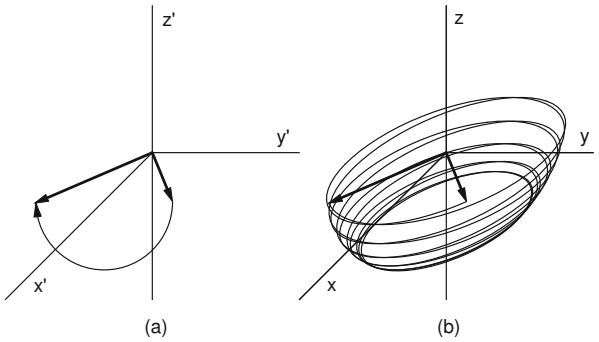


Fig. 18.9 A magnetic field B_1 pointing along the laboratory x axis and oscillating at the Larmor frequency causes nutation of **M** through an angle π around the rotating x' axis. In this case **M** was initially in the $x'y'$ plane. The motion shown here is plotted from Eqs. 18.30 in the rotating (**a**) and the laboratory (**b**) frames

18.6 Relaxation Times

Since longitudinal relaxation changes the value of M_z and hence $\boldsymbol{\mu} \cdot \mathbf{B}$, it is associated with a change of energy of the nucleus. The principal force that can do work on the nuclear spin and change its energy arises from the fact that the nucleus is in a fluctuating magnetic field due to neighboring nuclei and the electrons in paramagnetic atoms.

One way to analyze the effect of this magnetic field is to say that the change of spin energy E is accompanied by the emission or absorption of a photon of frequency $\nu_{\text{photon}} = E/h$, or $\omega_{\text{photon}} = \omega_0$. An increase of spin energy requires the absorption of a photon at the Larmor frequency (*stimulated absorption*). This will have a high probability if the fluctuating magnetic field has a large Fourier component at the Larmor frequency. A decrease of spin energy is accompanied by the emission of a photon. This can happen spontaneously in a vacuum (*spontaneous emission*), or it can be stimulated by the presence of other photons at the Larmor frequency (*stimulated emission*). These relative probabilities can be calculated using quantum mechanics. Stimulated emission or absorption is much more probable than is spontaneous emission.

The random magnetic field at a nucleus fluctuates because of the movement of the nucleus in the magnetic field of nearby atoms and nuclei. If the field changes rapidly enough, it will have Fourier components at the Larmor frequency that can induce transitions that cause M_z to change by absorption or emission. To get an idea of the strength of the field involved, consider the field at one hydrogen nucleus in a water molecule due to the other hydrogen nucleus. The field due to a magnetic dipole is given by

$$B_r = \frac{\mu_0}{4\pi} \frac{2\mu}{r^3} \cos\theta,$$

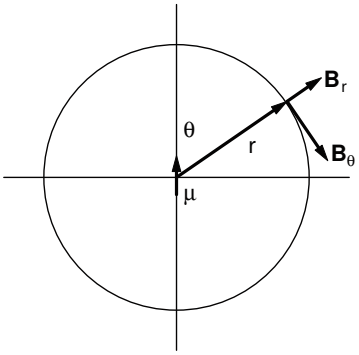


Fig. 18.10 The magnetic field components of a dipole in spherical coordinates point in the directions shown

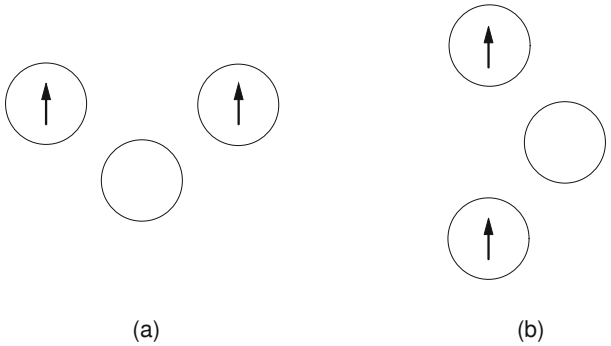


Fig. 18.11 The z components of the magnetic moments of two protons in a water molecule are shown for two different molecular orientations, **a** and **b**. When the water molecule is fixed in space, as in ice, the magnetic field that one proton produces in the neighborhood of the other is static. When the water molecule tumbles, as in a liquid or gas, the field that one proton produces at the other changes with time

$$B_\theta = \frac{\mu_0}{4\pi}\frac{\mu}{r^3}\sin\theta, \qquad (18.32)$$
$$B_\phi = 0,$$

where angle θ is defined in Fig. 18.10. (The factor $\mu_0/4\pi \equiv 10^{-7}$ T m A^{-1} is required in SI units.) The magnetic field at one hydrogen nucleus in a water molecule due to the other hydrogen nucleus is about 4×10^{-4} T (see Problem 14). Consider the water molecule shown in Fig. 18.11. We refer to each hydrogen nucleus as a proton. The z components of the proton magnetic moments are shown. If the water molecule is oriented as in Fig. 18.11a, the field at one proton due to the other has a certain value. If the water molecule remains fixed in space, as in ice, the field is constant with time. If the molecule is tumbling, as in liquid water, the orientation changes as in Fig. 18.11b, and the field changes with time.

When the molecules are moving randomly, the fluctuating magnetic field components are best described by their autocorrelation functions. The simplest assumption one can make[4] is that the autocorrelation function ϕ_{11} of each magnetic field component is exponential and that each field component has the same correlation time τ_C:

$$\phi_{11}(\tau) \propto e^{-|\tau|/\tau_C}. \qquad (18.33)$$

The Fourier transform of the autocorrelation function gives the power at different frequencies. It has only cosine terms because the autocorrelation is even. Comparison with the Fourier transform pair of Eq. 11.101 shows that the power at frequency ω is proportional to $\tau_C/(1+\omega^2\tau_C^2)$. With the assumption that the transition rate, which is $1/T_1$, is proportional to the power at the Larmor frequency, we have (see also Slichter 1990 or Levitt 2008)

$$\frac{1}{T_1} = \frac{C\tau_C}{1+\omega_0^2\tau_C^2}, \qquad (18.34)$$

where C is the proportionality constant.

The correlation time in a solid is much longer than in a liquid. For example, in liquid water at 20 °C it is about 3.5×10^{-12} s; in ice it is about 2×10^{-6} s. Figure 18.12 shows the behavior of T_1 as a function of correlation time, plotted from Eq. 18.34 with $C = 5.43\times10^{10}$ s^{-2}. For short correlation times T_1 does not depend on the Larmor frequency. At long correlation times T_1 is proportional to the square of the Larmor frequency, as can be seen from Eq. 18.34. The minimum in T_1 occurs when $\omega_0 = 1/\tau_C$ in this model.

Table 18.2 shows some typical values of the relaxation times for a Larmor frequency of 20 MHz. Neighboring paramagnetic atoms reduce the relaxation time by causing a fluctuating magnetic field. For example, adding 20 ppm of Fe^{3+} to pure water reduces T_1 from 3000 to 20 ms.

Differences in relaxation time are easily detected in an image. Different tissues have different relaxation times. A contrast agent containing gadolinium (Gd^{3+}), which is strongly paramagnetic, is often used in magnetic resonance imaging. It is combined with many of the same pharmaceuticals used with ^{99m}Tc, and it reduces the relaxation time of nearby nuclei. Gadolinium has been used to assess ischemic myocardium (Sakuma 2007). Iron oxide nanoparticles are sometimes used to create contrast in magnetic resonance images (Kim et al. 2011).

The hemoglobin that carries oxygen in the blood exists in two forms: oxyhemoglobin and deoxyhemoglobin. The former is diamagnetic and the latter is paramagnetic, so the relaxation time in blood depends on the amount of oxygen in the hemoglobin. The imaging technique that exploits this is called BOLD (blood oxygen level dependent).

[4] A more complete model recognizes that different atoms experience fluctuating fields with different correlation times and that frequency components at twice the Larmor frequency also contribute.

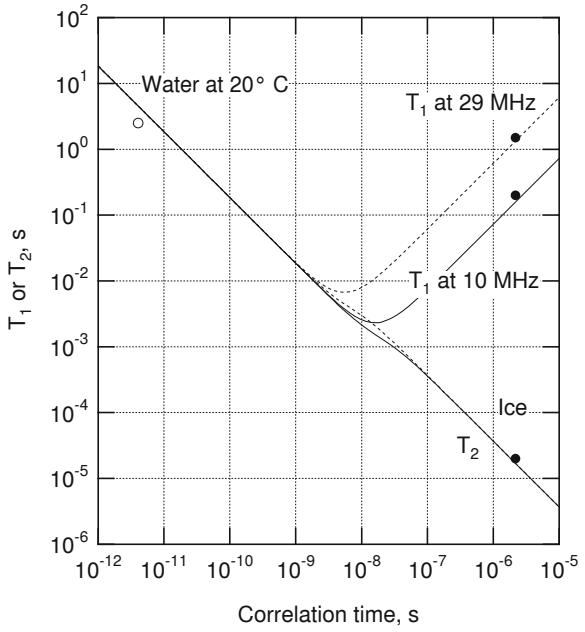


Fig. 18.12 Plot of T_1 and T_2 vs correlation time of the fluctuating magnetic field at the nucleus. The dashed lines are for a Larmor frequency of 29 MHz; the solid lines are for 10 MHz. Experimental points are shown for water (open dot) and ice (solid dots)

Table 18.2 Approximate relaxation times at 20 MHz

	T_1 (ms)	T_2 (ms)
Whole blood	900	200
Muscle	500	35
Fat	200	60
Water	3000	3000

The same model for the fluctuating fields which led to Eq. 18.34 gives an expression for T_2:

$$\frac{1}{T_2} = \frac{C\tau_C}{2} + \frac{1}{2T_1}, \tag{18.35}$$

$$T_2 = \frac{2}{C\tau_C}\left[\frac{1+\omega_0^2\tau_C^2}{2+\omega_0^2\tau_C^2}\right].$$

There is a slight frequency dependency to T_2 for values of the correlation time close to the reciprocal of the Larmor frequency.

Another effect that causes the magnetization to rapidly decrease is *dephasing*. Dephasing across the sample occurs because of inhomogeneities in the externally applied field. Suppose that the spread in Larmor frequency and the transverse relaxation time are related by $T_2\Delta\omega = K$. (Usually K is taken to be 2.) The spread in Larmor frequencies $\Delta\omega$ is due to a spread in magnetic field ΔB experienced by the nuclear spins in different atoms. The total variation in B is due to fluctuations caused by the magnetic field of neighbors and to variation in the applied magnetic field across the sample:

$$\Delta B_{\text{tot}} = \Delta B_{\text{internal}} + \Delta B_{\text{external}}.$$

Therefore

$$\Delta\omega_{\text{tot}} = \Delta\omega_{\text{internal}} + \Delta\omega_{\text{external}}.$$

The total spread is associated with the experimental relaxation time, $T_2^* = K/\Delta\omega_{\text{tot}}$. The *true* or *nonrecoverable* relaxation time $T_2 = K/\Delta\omega_{\text{internal}}$ is due to the fluctuations in the magnetic field intrinsic to the sample. Therefore

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{\gamma\Delta B_{\text{external}}}{K}. \tag{18.36}$$

T_2 is called the nonrecoverable relaxation time because various experimental techniques can be used to compensate for the external inhomogeneities, but not the internal atomic ones.

18.7 Detecting the Magnetic Resonance Signal

We have now seen that a sample of nuclear spins in a strong magnetic field has an induced magnetic moment; that it is possible to apply a sinusoidally varying magnetic field and nutate the magnetic moment to precess at any arbitrary angle with respect to the static field; and that the magnetization then relaxes or returns to its original state with two characteristic time constants, the longitudinal and transverse relaxation times. We next consider how a useful signal can be obtained from these spins. This is done by measuring the weak magnetic field generated by the magnetization as it precesses in the xy plane.

Suppose that a sample is at the origin. The motions plotted in Fig. 18.7 suggest that one way to produce a magnetization rotating in the xy plane is to have a static field along the z axis, combined with a coil in the yz plane (perpendicular to the x axis) connected to a generator of alternating current at frequency ω_0. Turning on the generator for a time $\Delta t = \pi/2\omega_1 = \pi/\gamma B_1$ rotates the magnetization into the xy plane (a 90° or $\pi/2$ pulse). If the generator is then turned off, the same coil can be used to detect the changing magnetic flux due to the rotating magnetic moments. The resulting signal, an exponentially damped sine wave, is called the *free induction decay* (FID).

To estimate the size of the signal induced in the coil, imagine a magnetic moment $\boldsymbol{\mu} = \mathbf{M}\Delta V$ rotating in the xy plane as shown in Fig. 18.13. The voltage v induced in a one-turn coil in the yz plane is the rate of change of the magnetic flux through the coil:

$$v = -\frac{\partial\Phi}{\partial t} = -\frac{\partial}{\partial t}\iint \mathbf{B}\cdot d\mathbf{S}.$$

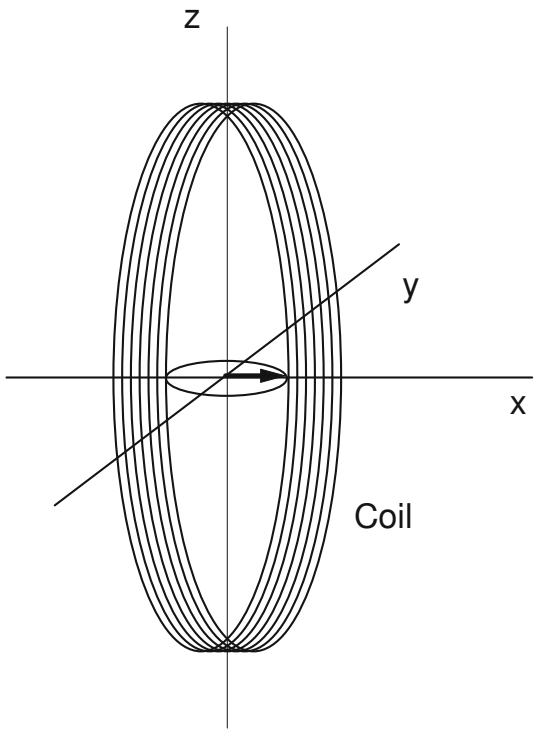


Fig. 18.13 A magnetic moment rotating in the xy plane induces a voltage in a pickup coil in the yz plane. The coil is viewed from slightly to the right of the coil

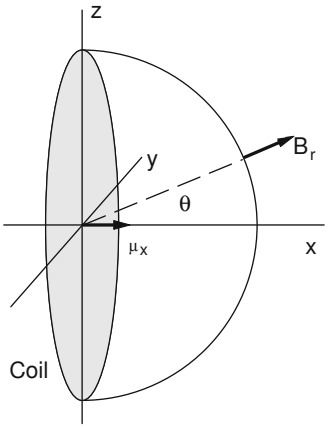


Fig. 18.14 A dipole along the x axis generates a flux through the shaded circle in the yz plane that is equal and opposite to that through the hemispherical cap. The drawing is viewed from slightly to the right of the yz plane

The magnetic field far from a magnetic dipole can be written most simply in spherical coordinates (Eqs. 18.32). We need the flux through the coil of radius a in the yz plane. However, Eqs. 18.32 are not valid close to the dipole. Since a fundamental property of the magnetic field is that for a closed surface $\iint \mathbf{B}\cdot d\mathbf{S} = 0$, the flux through the coil in Fig. 18.13 is the negative of the flux Φ through the hemispherical cap in Fig. 18.14:

$$\Phi = -\int B_r 2\pi a^2 \sin\theta\, d\theta = -\frac{\mu_0}{4\pi}\frac{4\pi\mu_x}{a}\int_0^{\pi/2} \cos\theta\, \sin\theta\, d\theta$$
$$= -\frac{\mu_0}{4\pi}\frac{2\pi\mu_x}{a}. \qquad (18.37)$$

At any instant $\boldsymbol{\mu}$ can be resolved into components along x and y. The component pointing along y contributes no net flux through the spherical cap of Fig. 18.14. Therefore, the flux for a magnetic moment $\boldsymbol{\mu} = \mathbf{M}\Delta V$, where $\mathbf{M}$ is given by Eqs. 18.16, is

$$\Phi = -\frac{\mu_0}{4\pi}\frac{2\pi M_0 \Delta V}{a} e^{-t/T_2}\cos(-\omega_0 t).$$

The induced voltage is $-\partial\Phi/\partial t$:

$$v = \frac{\mu_0}{4\pi}\frac{2\pi M_0 \Delta V}{a} e^{-t/T_2}\left(-\frac{1}{T_2}\cos(-\omega_0 t) + \omega_0 \sin(-\omega_0 t)\right).$$

Since $1/T_2 \ll \omega_0$, this can be simplified to

$$v = -\frac{\mu_0}{4\pi}\frac{\omega_0}{a} 2\pi M_0 \Delta V\, e^{-t/T_2}\sin(-\omega_0 t).$$

If the value of M_z which exists at thermal equilibrium has been nutated into the xy plane, then M_0 is given by the M_z of Eq. 18.10. For a spin-$\frac{1}{2}$ particle (and using the fact that $\omega_0 = \gamma B_0$) we obtain

$$v = -\frac{\mu_0}{4\pi}\frac{\pi N\, \Delta V\, \gamma^3 \hbar^2 B_0^2}{2k_B T a} e^{-t/T_2}\sin(-\omega_0 t). \qquad (18.38)$$

Here $N\,\Delta V$ is the total number of nuclear spins involved, B_0 is the field along the z axis, and a is the radius of the coil that detects the free-induction-decay signal. For a volume element of fixed size, as in magnetic resonance imaging, the sensitivity is inversely proportional to the coil radius. If the sample fills the coil, as in most laboratory spectrometers, then $\Delta V \propto a^2$ and the sensitivity is proportional to a. Because the signal in Eq. 18.38 is proportional to the square of the magnetic field B_0, there has been a push for higher and higher magnetic field strengths; 7 T is typical for high B_0 studies (Robitaille and Berliner 2006).

18.8 Some Useful Pulse Sequences

Many different ways of applying radio-frequency pulses to generate B_1 have been developed by nuclear magnetic resonance spectroscopists for measuring relaxation times. There are five "classic" sequences, which also form the basis for magnetic resonance imaging.

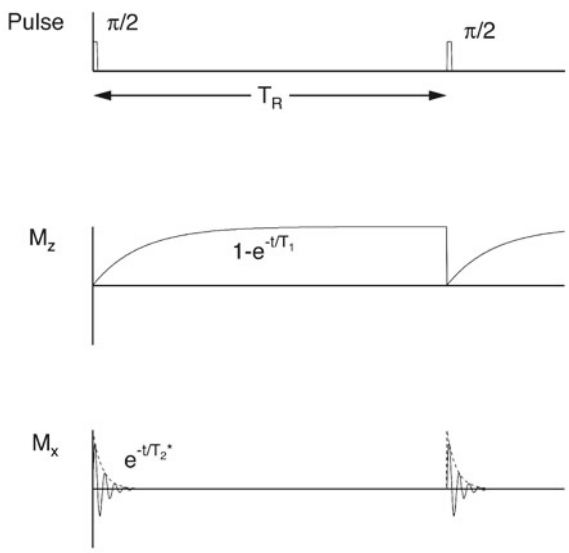


Fig. 18.15 Pulse sequence and signal for a free-induction-decay measurement

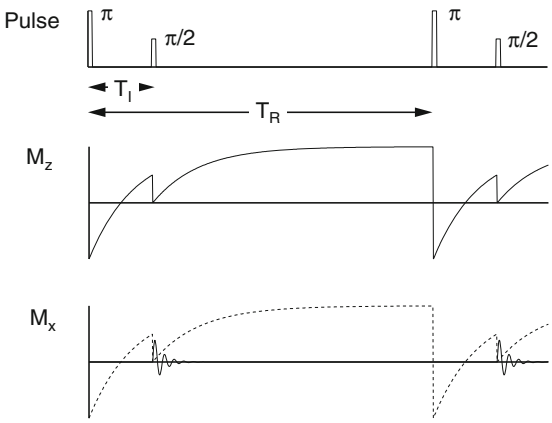


Fig. 18.16 The inversion recovery sequence allows determination of T_1 by making successive measurements at various values of the interrogation time T_I

18.8.1 Free-Induction-Decay (FID) Sequence

Free induction decay was described in Sect. 18.7. A $\pi/2$ pulse nutates **M** into the xy plane, where its precession induces a signal in a pickup coil. The signal is of the form $e^{-(t/T_2^*)}\sin(-\omega_0 t)$, where T_2^* is the experimental transverse relaxation time, including magnetic field inhomogeneities due to the apparatus as well as those intrinsic to the sample. Figure 18.15 shows the pulse sequence, the value of M_x, and the value of M_z. The signal is proportional to M_x. The pulses can be repeated after time T_R for signal averaging. It is necessary for T_R to be greater than, say, $5T_1$ in order for M_z to return nearly to its equilibrium value between pulses.

18.8.2 Inversion-Recovery (IR) Sequence

The inversion-recovery sequence allows measurement of T_1. A π pulse causes **M** to point along the $-z$ axis. There is not yet any signal at this time. M_z returns to equilibrium according to $M_z = M_0\left[1-2e^{-(t/T_1)}\right]$. A $\pi/2$ interrogation pulse at time T_I rotates the instantaneous value of M_z into the xy plane, thereby giving a signal proportional to $M_0\left[1-2e^{-(T_I/T_1)}\right]$, as shown in Fig. 18.16. The process can be repeated; again the repeat time must exceed $5T_1$.

You can see from Fig. 18.16 that there will be no signal at all if $T_I/T_1 = 0.693$. If T_I is less than this, the M_x signal will be inverted (negative). Unless special detector circuits are used which allow one to determine that M_x is negative, the results can be confusing.

Inversion recovery images take a long time to acquire and there is ambiguity in the sign of the signal. There are also problems with the use of a π pulse for slice selection (defined in Sect. 18.9; the details of the problems are found in Joseph et al. 1984).

18.8.3 Spin-Echo (SE) Sequence

The pulse sequence shown in Fig. 18.17 can be used to determine T_2 rather than T_2^*. Initially a $\pi/2$ pulse nutates **M** about the x' axis so that all spins lie along the rotating y' axis. Figure 18.17a shows two such spins. Spin **a** continues to precess at the same frequency as the rotating coordinate system; spin **b** is subject to a slightly smaller magnetic field and precesses at a slightly lower frequency, so that at time $T_E/2$ it has moved clockwise in the rotating frame by angle θ, as shown in Fig. 18.17b. At this time a π pulse is applied that rotates all spins around the x' axis. Spin **a** then points along the $-y'$ axis; spin **b** rotates to the angle shown in Fig. 18.17c. If spin **b** still experiences the smaller magnetic field, it continues to precess clockwise in the rotating frame. At time T_E both spins are in phase again, pointing along $-y'$ as shown in Fig. 18.17d. The resulting signal is called an *echo*, and the process for producing it is called a *spin-echo sequence*. The formation of an echo depends only on the fact that the magnetic field at the nucleus remained the same before and after the π pulse; it does not depend on the specific value of the dephasing angle. Therefore all of the spin dephasing that has been caused by a time-independent magnetic field is reversed in this process. There remains only the dephasing caused by fluctuating magnetic fields. Figure 18.18 shows the pulse sequence and signal.

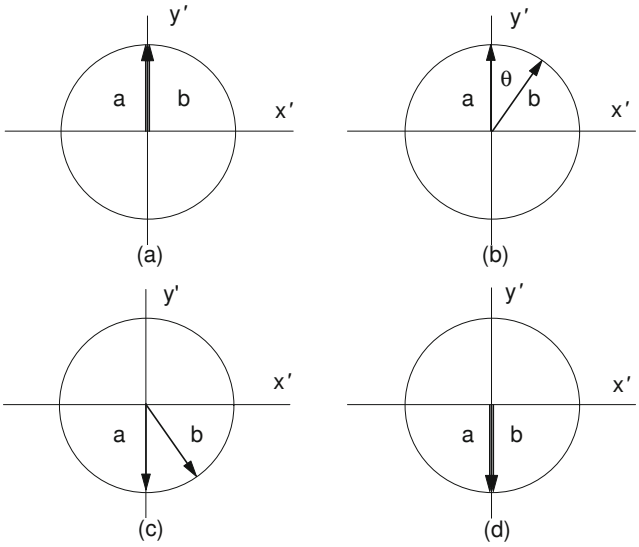


Fig. 18.17 Two magnetic moments are shown in the $x'y'$ plane in the rotating coordinate system. Moment **a** rotates at the Larmor frequency and remains aligned along the y' axis. Moment **b** rotates clockwise with respect to moment **a**. **a** Both moments are initially in phase. **b** After time $T_E/2$ moment **b** is clockwise from moment **a**. **c** A π pulse nutates both moments about the x' axis. **d** At time T_E both moments are in phase again

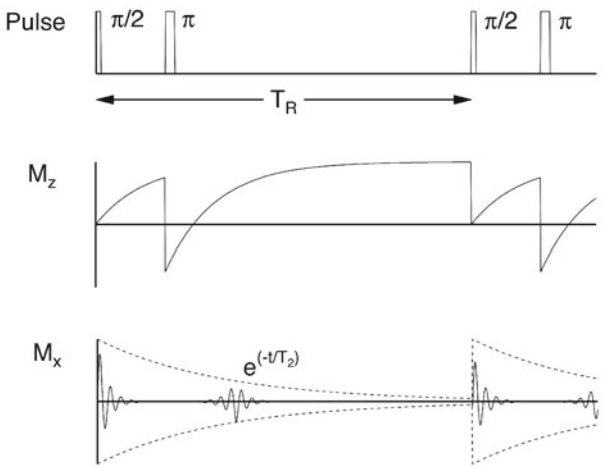


Fig. 18.18 The pulse sequence and magnetization components for a spin-echo sequence

18.8.4 Carr–Purcell (CP) Sequence

When a sequence of π pulses that nutate **M** about the x' axis are applied at $T_E/2$, $3T_E/2$, $5T_E/2$, etc., a sequence of echoes are formed, the amplitudes of which decay with relaxation time T_2. This is shown in Fig. 18.19. Referring to Fig. 18.17, one can see that the echoes are aligned alternately

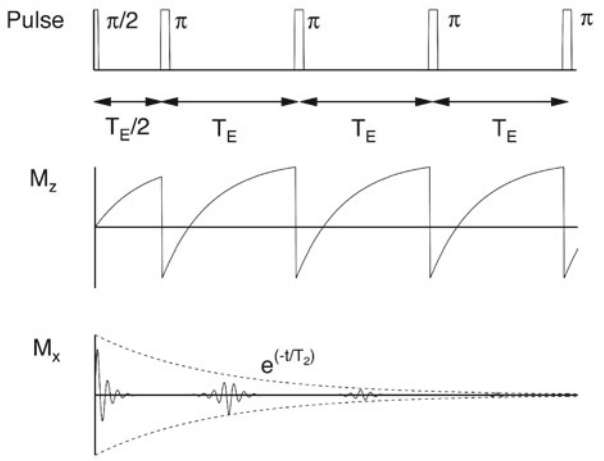


Fig. 18.19 The Carr–Purcell pulse sequence. All pulses nutate about the x' axis. Echoes alternate sign. The envelope of echoes decays as e^{-t/T_2}, where T_2 is the unrecoverable transverse relaxation time

along the $-y'$ and $+y'$ axes. One advantage of the Carr–Purcell sequence is that it allows one to determine rapidly many points on the decay curve. Another advantage relates to diffusion. The molecules that contain the excited nuclei may diffuse. If the external magnetic field B_0 is not uniform, the molecules can diffuse to another region where the magnetic field is slightly different. As a result the rephasing after a pulse does not completely cancel the initial dephasing. This effect is reduced by the Carr–Purcell sequence (see Problem 47).

18.8.5 Carr–Purcell–Meiboom–Gill (CPMG) Sequence

One disadvantage of the CP sequence is that the π pulse must be very accurate or a cumulative error builds up in successive pulses. The Carr–Purcell–Meiboom–Gill sequence overcomes this problem. The initial $\pi/2$ pulse nutates **M** about the x' axis as before, but the subsequent pulses are shifted a quarter cycle in time, which causes them to rotate about the y' axis. This is shown in Fig. 18.20. To see why this pulse sequence (Fig. 18.21) is less sensitive to errors in the duration of the π pulses, consider moment **a**. In the CP sequence, Fig. 18.17, a π pulse that is too long will nutate **a** too far, and it will have a smaller component in the $x'y'$ plane. The next pulse will nutate it even further. In Fig. 18.20, the π pulses will not affect moment **a** at all. This is explored further in Problems 29 and 30.

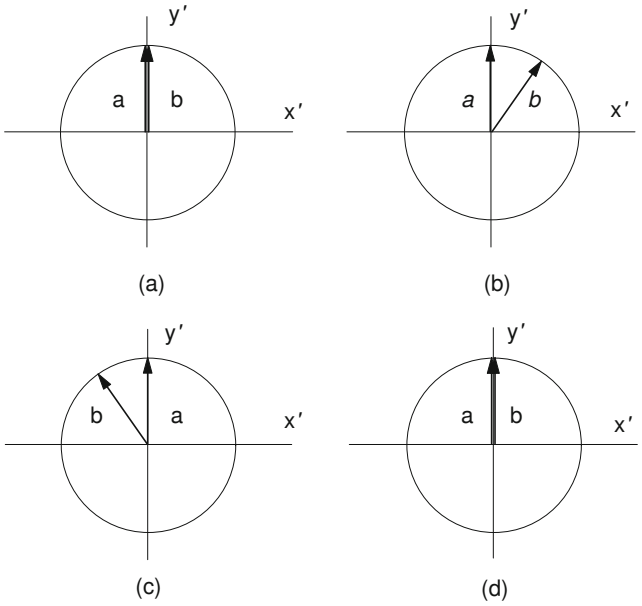


Fig. 18.20 The effect of the Carr–Purcell–Meiboom–Gill pulse sequence on the magnetization. This is similar to Fig. 18.17 except that the π pulses rotate around the y' axis. Moment **b** rotates clockwise in the $x'y'$ plane. **a** Both moments are initially in phase. **b** After time $T_E/2$ moment **b** is clockwise from moment **a**. **c** A π pulse rotates both moments about the y' axis. **d** At time T_E both moments are in phase again

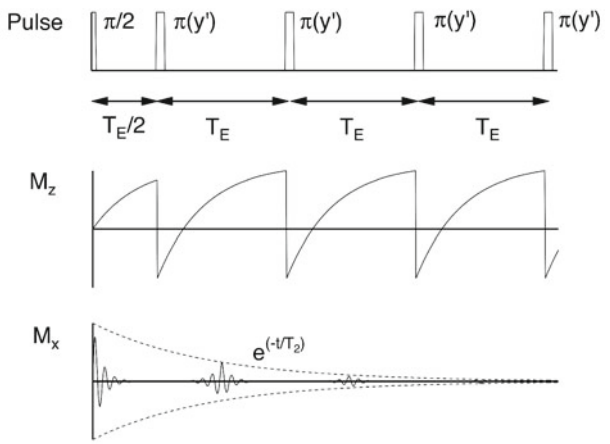


Fig. 18.21 The CPMG pulse sequence

18.9 Imaging

Many more techniques are available for imaging with magnetic resonance than for x-ray computed tomography. They are described by Brown et al. (1994), by Choet al. (1993), by Vlaardingerbroek and den Boer (2004), and by Liang and Lauterbur (2000). One of these authors, Paul C. Lauterbur, shared with Sir Peter Mansfield the 2003 Nobel Prize in physiology or medicine for the invention of magnetic resonance imaging.

Creation of the images requires the application of gradients in the static magnetic field B_z which cause the Larmor frequency to vary with position. The first gradient is applied in the z direction during the $\pi/2$ pulse so that only the spins in a slice in the patient are *selected* (nutated into the xy plane). Slice selection is followed by gradients of B_z in the x and y directions. These also change the Larmor frequency. If the gradient is applied during the readout, the Larmor frequency of the signal varies as B_z varies with position. If the gradient is applied before the readout, it causes a position-dependent phase shift in the signal which can be detected.

We discuss several reconstruction methods here. *Projection reconstruction* is similar to CT reconstruction, but it is slow and rarely used. A two-dimensional Fourier technique known as *spin warp* or *phase encoding* forms the basis of the techniques used in most machines. We also describe briefly some techniques that are even faster. Finally, we discuss how the image contrast can be modified by changing the pulse sequence parameters.

Our initial discussion is based on a spin-echo pulse sequence, repeated with a repetition time T_R as shown in Fig. 18.18.

18.9.1 Slice Selection

First, suppose we simply apply a $\pi/2$ pulse to the entire sample in a 1.5-T machine ($\omega_0 = 401 \times 10^6$ s^{-1}; $\nu_0 = 63.9$ MHz). If the duration of this pulse is to be, say, 5 ms, it requires a constant amplitude of the radio-frequency magnetic field

$$B_1 = \pi/\gamma\,\Delta t = 2.35 \times 10^{-6}\ \mathrm{T}. \tag{18.39}$$

The pulse lasts for 3×10^5 cycles at the Larmor frequency. The frequency spread of the pulse is about 200 Hz. This excites all the proton spins in the entire sample.

For MR imaging, we want to select a thin slice in the sample. In order to select a thin slice (say $\Delta z = 1$ cm) we apply a magnetic field gradient in the z direction while applying a specially shaped B_1 signal. In a static magnetic field B_0, the field lines are parallel. The field strength is proportional to the number of lines per unit area and does not change. With the gradient applied in the volume of interest, the field lines converge, and the field increases linearly with z as shown in Fig. 18.22a, b:

$$B_z(z) = B_0 + G_z z. \tag{18.40}$$

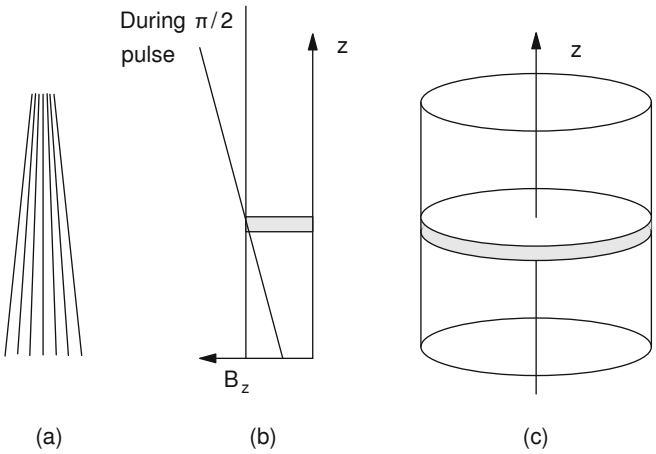


Fig. 18.22 **a** Magnetic field lines for a magnetic field that increases in the z direction. **b** A plot of B_z vs z with and without a gradient. **c** After application of a field gradient in the z direction during the specially shaped rf pulse, all of the spins in the shaded *slice* are excited, that is, they are precessing in the xy plane

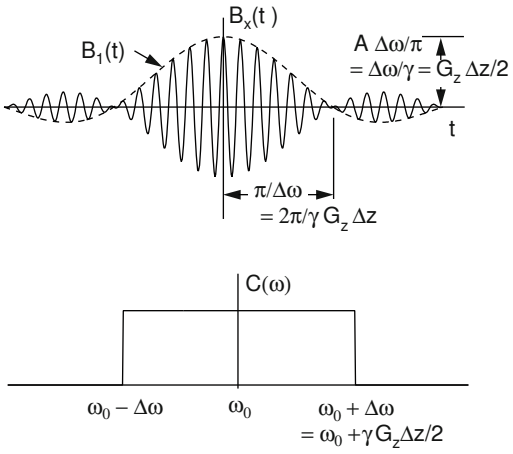


Fig. 18.23 **a** The $B_x(t)$ signal shown is used to selectively excite a slice. It consists of $\cos(\omega_0 t)$ modulated by a $\mathrm{sinc}(x)$ or $\sin(x)/x$ pulse $B_1(t)$. **b** The frequency spectrum contains a uniform distribution of frequencies

We adopt a notation in which **G** represents a partial derivative of the z component of the magnetic field:

$$\mathbf{G} = \begin{pmatrix} G_x \\ G_y \\ G_z \end{pmatrix} = \begin{pmatrix} \partial B_z/\partial x \\ \partial B_z/\partial y \\ \partial B_z/\partial z \end{pmatrix}. \tag{18.41}$$

In a typical machine, $G_z = 5\times10^{-3}$ T m^{-1}. For a slice thickness $\Delta z = 0.01$ m, the Larmor frequency across the slice varies from $\omega_0 - \Delta\omega$ to $\omega_0 + \Delta\omega$, where $\Delta\omega = \gamma G_z \Delta z/2 = 6.68 \times 10^3$ s^{-1} ($\Delta f = 1.064$ kHz).

It is possible to make the signal $B_x(t)$ consist of a uniform distribution of frequencies between $\omega_0 - \Delta\omega$ and $\omega_0 + \Delta\omega$, so that all protons are excited in a slice of thickness Δz from $-\Delta z/2$ to $+\Delta z/2$. Let the amplitude of B_x in the interval $(\omega, d\omega)$ be A. Using Eq. 11.57, $B_x(t)$ is given by

$$\begin{aligned} B_x(t) &= \frac{A}{2\pi}\int_{\omega_0-\Delta\omega}^{\omega_0+\Delta\omega} \cos(\omega t)\, d\omega \\ &= \frac{A\Delta\omega}{\pi}\frac{\sin(\Delta\omega t)}{\Delta\omega t}\cos(\omega_0 t). \end{aligned} \tag{18.42}$$

This has the form $B_1(t)\cos(\omega_0 t)$, where $B_1(t) = (A\Delta\omega/\pi)\sin(\Delta\omega t)/(\Delta\omega t)$. The function $\sin(x)/x$ has its maximum value of 1 at $x = 0$. It is also called the $\mathrm{sinc}(x)$ function. The angle ϕ through which the spins are nutated is

$$\begin{aligned} \phi &= \int_{-\infty}^{\infty} \omega_1(t)\, dt = \frac{\gamma}{2}\int_{-\infty}^{\infty} B_1(t)\, dt \\ &= \frac{\gamma A\Delta\omega}{2\pi}\int_{-\infty}^{\infty}\frac{\sin(\Delta\omega t)}{\Delta\omega t}\, dt \\ &= \frac{\gamma A}{2}. \end{aligned}$$

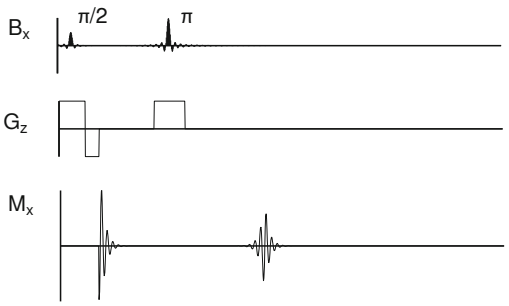


Fig. 18.24 A slice selection pulse sequence. While a gradient G_z is applied, a $\pi/2$ B_x (rf) pulse nutates the spins in a slice of thickness Δz into the xy plane. A negative G_z gradient restores the phase of the precessing spins. The echo after the π pulse is from the entire slice

For a $\pi/2$ pulse, $A = \pi/\gamma$. The maximum value of B_1 is therefore $\Delta\omega/\gamma = G_z\Delta z/2$, as shown in Fig. 18.23. The B_x pulse does not have an abrupt beginning; it grows and decays as shown. In practice, it is truncated at some distance from the peak where the lobes are small.

While the gradient is applied, the transverse components of spins at different values of z precess at different rates (see Problem 35). Therefore it is necessary to apply a gradient G_z of opposite sign after the $\pi/2$ pulse is finished in order to bring the spins back to the phase they had at the peak of the slice selection signal. The gradient is removed when all of the spins in the slice shown in Fig. 18.22c are back in phase. They then continue to precess in the xy plane at the Larmor frequency. This gives the first M_x pulse in Fig. 18.24. The complicated behavior of M_x during the slice selection gradient is not shown. This initial free-induction-decay pulse is not used for imaging.

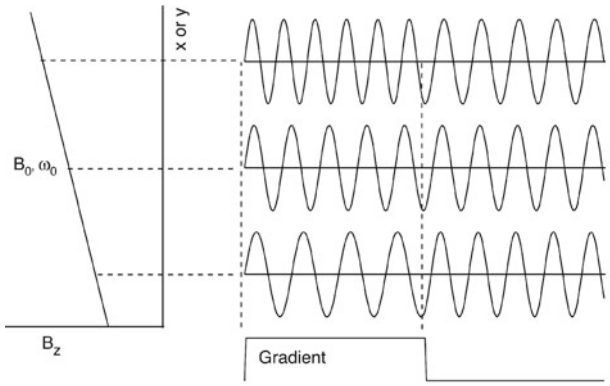


Fig. 18.25 A gradient in B_z causes the Larmor frequency to vary with position. If the signal is measured while the gradient is applied, the Larmor frequency varies with position. If the signal is measured after the gradient has been applied and removed, a position-dependent phase shift remains

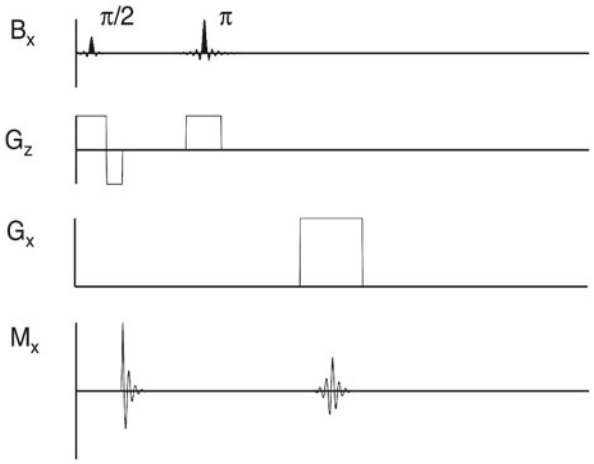


Fig. 18.26 A gradient G_x is applied during x readout. The echo signal between ω and $\omega + d\omega$ is proportional to the magnetization in a strip between x and $x + dx$, integrated over all values of y

The voltage induced in the pickup coil surrounding the sample is proportional to the free induction decay of **M** in the entire slice. That is, the voltage signal induced in the pickup coil is proportional to $\int M(x, y, z) \cos(-\omega_0 t) f(t)\, dV$, where $M(x, y, z)$ is the magnetization per unit volume that was nutated into the xy plane, $\cos(-\omega_0 t)$ represents the change in signal as **M** rotates in the xy plane at the Larmor frequency, and $f(t)$ represents relaxation, signal buildup during an echo, and so on. Figure 18.24 shows the echo after a subsequent π pulse, applied with G_z on, and which also has the form of a sinc function to affect only those spins in the slice of interest.

We assume that changes in $f(t)$ are slow compared to the Larmor frequency and neglect them here. Then the signal from an element $dxdy$ in the slice is

$$v(t) = A\, dx\, dy\, \Delta z\, M(x, y, z) \cos(-\omega_0 t). \qquad (18.43)$$

Constant A includes all the details of the detecting coils and receiver.

18.9.2 Readout in the Direction

We now need to extract x and y position information from $v(t)$. This is done by creating gradients of B_z in the x or y directions. As shown in Fig. 18.25, if the signal is measured while a gradient is applied, the Larmor frequency varies with position. Suppose that B_z is given a gradient G_x in the x direction during the echo signal readout, as shown in Fig. 18.26. G_x is called the *readout* or *frequency encoding* gradient. The spins that echo in the shaded slice between x and $x + dx$ in Fig. 18.27 will be precessing with a Larmor frequency between ω and $\omega + d\omega$, where $\omega = \omega_0 + \gamma G_x x$.

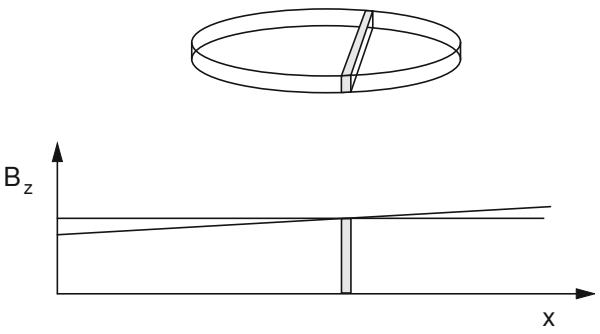


Fig. 18.27 Because the gradient G_x is applied during readout, the Larmor frequency of all spins in the shaded slice is between ω and $\omega + d\omega$

The signal from the entire slice is

$$v(t) = A\Delta z \int dx \left(\int dy\, M(x, y, z) \right) \cos[-\omega(x)t]. \qquad (18.44)$$

We use the fact that $\omega(x) = \omega_0 + \gamma G_x x$ to write the signal as

$$v(t) = A\Delta z \int dx \Big[\left(\int dy\, M(x, y, z) \right) \cos(-\omega_0 t - \gamma G_x x t) \Big]. \qquad (18.45)$$

Since the z slice has already been selected, let us simplify the notation by dropping the z dependence of **M**. The electronics in the detector multiply $v(t)$ by $\cos(\omega_0 t)$ or $\sin(\omega_0 t)$ and average over many cycles at the Larmor frequency.

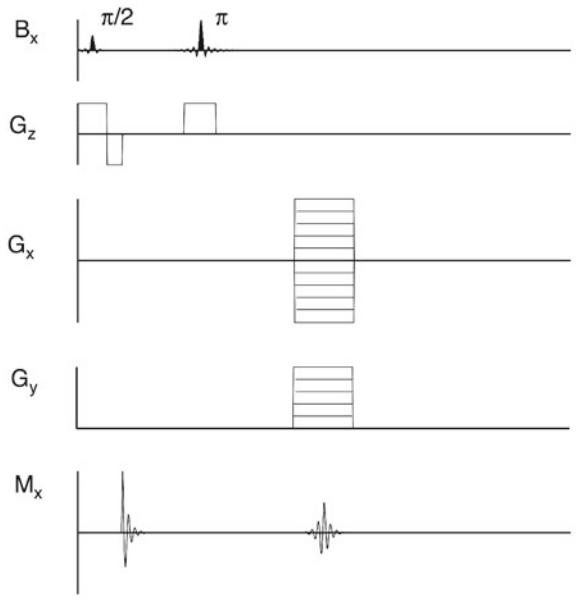


Fig. 18.28 Projection reconstruction techniques can be used to form an image. A series of measurements are taken, each with simultaneous gradients G_x and G_y

The results are two signals that form the basis for constructing the image:

$$s_c(t) = \overline{v(t)\cos(\omega_0 t)} \propto \iint dx\,dy\,M(x,y)\cos(-\gamma G_x x t),$$
$$s_s(t) = \overline{v(t)\sin(\omega_0 t)} \propto \iint dx\,dy\,M(x,y)\sin(-\gamma G_x x t). \tag{18.46}$$

The time average is over many cycles at the Larmor frequency but a time short compared to $2\pi/\gamma G_x x_{\max}$.

18.9.3 Projection Reconstruction

By inspecting Eq. 18.46 and remembering the relationship between ω and x, we see that the Fourier transforms of $s_c(t)$ and $s_s(t)$ are both proportional to $\int M(x,y)\,dy$. (Of course, the signals are digitized and one actually deals with discrete transforms.) This means that s_c or s_s can be Fourier analyzed to determine the amount of signal in the frequency interval $(\omega, d\omega)$ corresponding to (x, dx), which is proportional to the projection $\int M(x,y)\,dy$ along the shaded strip. In Sect. 12.5 we learned how to reconstruct an image from a set of projections. The entire readout process can therefore be repeated with the gradient rotated slightly in the xy plane (that is, with a combination of G_x and G_y during readout). This is indicated in Fig. 18.28, which indicates many scans, with different values of G_x and G_y, related by $G_y/G_x = \tan\theta$, where θ is the angle between the projection and the x axis. All of the techniques for reconstruction from projections that were developed for computed tomography can be used to reconstruct $M(x,y)$. Sending the proper combination of currents through the x and y gradient coils rotates the gradient; no rotating mechanical components are needed.

18.9.4 Phase Encoding

Techniques are available for magnetic resonance imaging that are not available for computed tomography. They are based on determining directly the Fourier coefficients in two or three dimensions. The basic technique is called *spin warp* or *phase encoding*. We saw in Fig. 18.25 that if a gradient is applied after the $\pi/2$ slice-selection pulse, a position-dependent phase shift remains even after the gradient is turned off. Let us make this quantitative. We wish to construct an image of $M(x,y)$, modified by the function $f(t)$ that accounts for relaxation, etc. For simplicity of notation we again assume f is unity and suppress the z dependence, since slice selection has already been done. We will construct $M(x,y)$ from its Fourier transform. The Fourier transform of $M(x,y)$ is given by Eqs. 12.9:

$$M(x,y) = \left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty} dk_x \int_{-\infty}^{\infty} dk_y [C(k_x,k_y)\cos(k_x x + k_y y) + S(k_x,k_y)\sin(k_x x + k_y y)]. \tag{18.47a}$$

with the coefficients given by

$$C(k_x,k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\,M(x,y)\cos(k_x x + k_y y), \tag{18.47b}$$

$$S(k_x,k_y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\,M(x,y)\sin(k_x x + k_y y). \tag{18.47c}$$

Our problem is to determine C and S and from them construct the image.

The information from the x readout gives us $C(k_x,0)$ and $S(k_x,0)$ directly. We show this for the cosine transform. From Eq. 18.47b

$$C(k_x,0) = \int_{-\infty}^{\infty} dx \left(\int_{-\infty}^{\infty} dy\,M(x,y)\right)\cos(k_x x). \tag{18.48}$$

Comparing this to the expression for $s_c(t)$ in Eq. 18.46, we see that

$$C(k_x,0) \propto s_c(k_x/\gamma G_x). \tag{18.49a}$$

Similarly,

$$S(k_x,0) \propto s_s(k_x/\gamma G_x). \tag{18.49b}$$

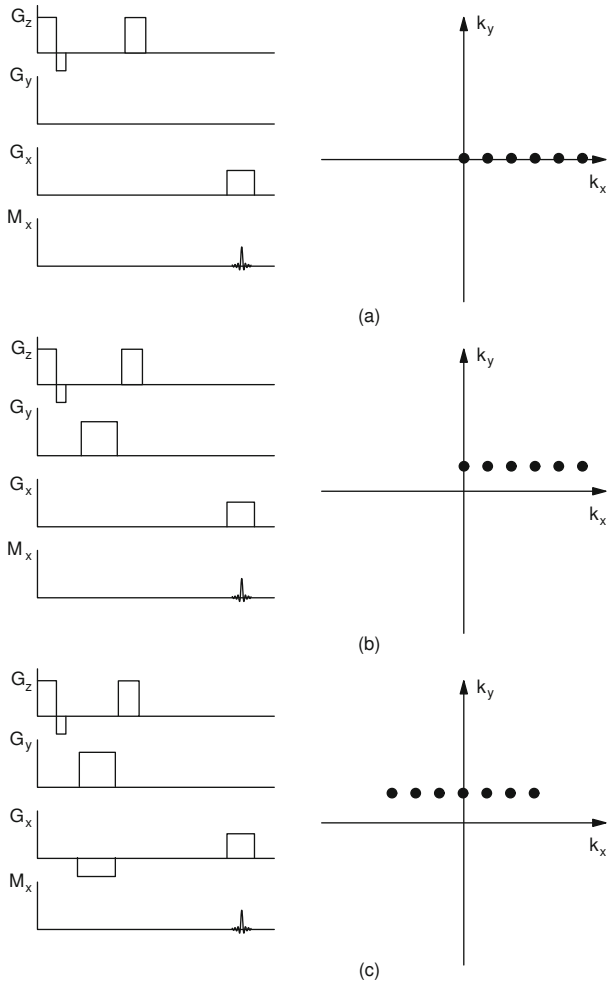


Fig. 18.29 **a** The signal measured while the x gradient is applied gives the spatial Fourier transform of the image along the k_x axis. **b** The addition of a phase-encoding gradient sets a nonzero value for k_y so that the readout determines the spatial Fourier transform along a line parallel to the k_x axis. **c** Phase encoding along the x axis as well shifts the line along which the coefficients are determined

The times at which s_c and s_s are measured and therefore the values of k_x are, of course, discrete. The discussion in Sect. 12.3 shows that the values of k_x are multiples of the lowest spatial frequency: $k_x = m\,\Delta k = mk_0 = 2\pi m/L_x$. The corresponding times to measure the signal are $t_m = 2\pi m/L_x\gamma G_x$. The spatial extent of the image in the x direction or *field of view* L_x determines the spacing Δk_x. The desired pixel size determines the maximum value of k_x or m: $\Delta x = \pi/k_{\max} = L_x/2m_{\max}$. The discrete values of k_x are shown in Fig. 18.29a.

The next problem is to make a similar determination for nonzero values of k_y. To do so, a gradient $G_y = \partial B_z/\partial y$ is applied at some time between slice selection and readout. This makes the Larmor frequency vary in the y direction. If the phase-encoding pulse is due to a uniform gradient that

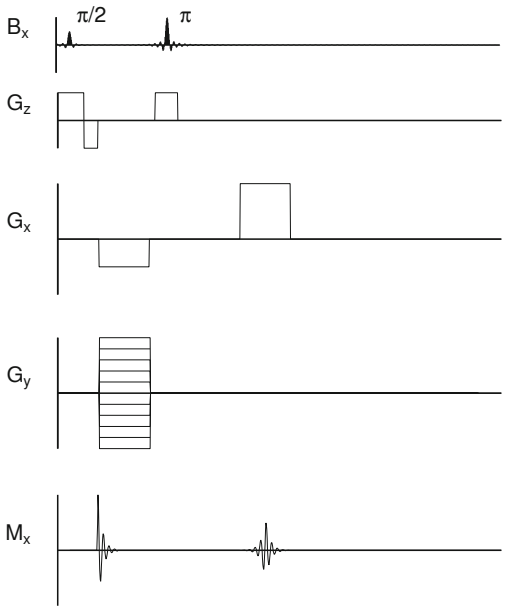


Fig. 18.30 The signals in a standard phase encoding. The pulse sequence is repeated for each value of k_y

lasts for a time T_p, the total phase change is

$$\Delta\phi = \int \omega(t)\,dt = \gamma G_y T_p y = k_y y. \tag{18.50}$$

The readout signal, Eq. 18.44, is replaced by

$$v(t) = A\,\Delta z \int dx \int dy\, M(x, y)\cos[\omega(x)t + k_y y]. \tag{18.51}$$

Note that the added phase does not depend on t because G_y is not on during readout. However, the cosine term must now be included in both the x and y integrals. Carrying through the mathematics of the detection process shows that temporal Fourier transformation of the signals determines $C(k_x, k_y)$ and $S(k_x, k_y)$ for all values of k_x and for the particular value of k_y determined by the G_y phase selection pulse. Different values of the G_y pulse give the coefficients for different values of k_y, as shown in Fig. 18.29. Both positive and negative gradients are used to give both positive and negative values of k_y. Application of a gradient G_x during the phase-encoding time (in addition to the readout gradient) changes the starting value of k_x. This allows one to determine the coefficients for negative values of k_x. This figure has been drawn without taking into account that the application of a π pulse changes k_x to $-k_x$ and k_y to $-k_y$. The gradients and signals for this spin-echo determination are shown in Fig. 18.30. The coefficients are substituted in Eq. 18.47a to reconstruct $M(x, y, z)$ for the z slice in question.

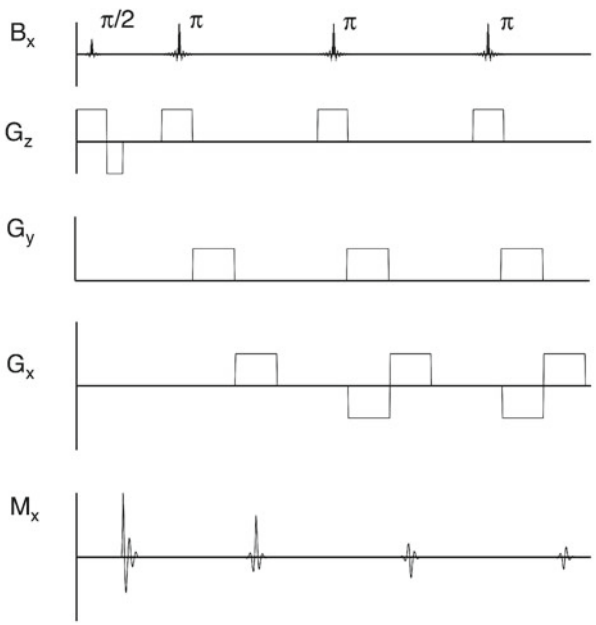


Fig. 18.31 A fast spin-echo sequence uses a single $\pi/2$ slice selection pulse followed by multiple echo rephasing pulses. A correction must be made for the transverse decay

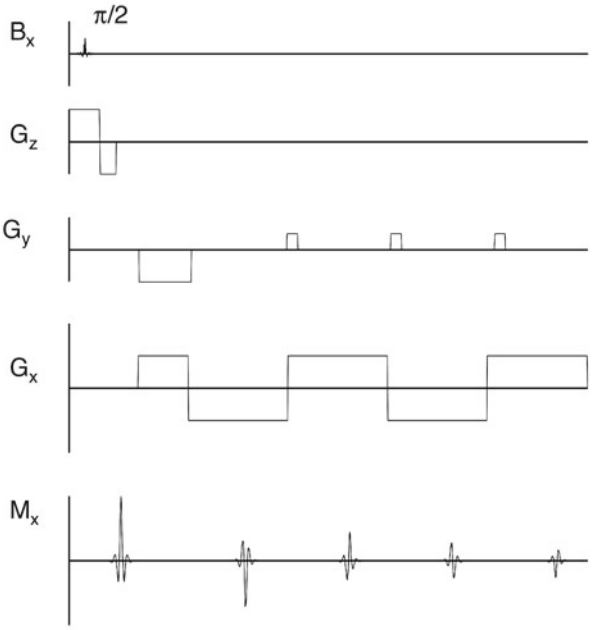


Fig. 18.32 Echo planar imaging uses a very uniform magnet and eliminates the rephasing π pulses. In Fig. 18.31 the decay of the individual echoes was determined by T_2^*, and the slower decay of the amplitude of the subsequent echoes was determined by T_2. In this figure the decay of the individual echoes is determined by the size of the gradient and the slower decay of the amplitude of the subsequent echoes is determined by T_2^*

18.9.5 Other Pulse Sequences

Dozens of other pulse sequences have been invented, all of which are based on the fundamentals presented here. We mention only a few, and there are many variations of these. For details, see Bernstein et al. (2004).

Fast spin echo or *turbo spin echo* uses a single $\pi/2$ pulse, followed by a series of π pulses, as shown in Fig. 18.31. Each π pulse produces an echo, though the echo amplitudes decay and a correction for this must be made in the image reconstruction. Each G_y pulse increments or *winds* the phase by a fixed amount. A negative G_x pulse resets the positions of the k_x values. Faster image acquisition sequences not only save time, but they may allow the image to be obtained while the patient's breath is held, thereby eliminating motion artifacts.

The major problem with conventional spin echo (Fig. 18.30) is that one must wait a time $T_R \gg T_1$ between measurements for different values of k_y. One way to speed things up is to use the intervening time to make measurements in a slice at a different value of z. Another technique is to use a flip angle smaller than $\pi/2$. Suppose the flip angle is $\alpha = 20\,^\circ$. This gives a transverse magnetization proportional to $\sin 20\,^\circ = 0.34$ while reducing the longitudinal magnetization only slightly, to $\cos 20\,^\circ = 0.94$. Thus, k space can be sampled until the transverse signal has decayed and another α flip pulse can immediately be applied to restore the transverse magnetization.

Echo-planar imaging (EPI) eliminates the π pulses. It requires a magnet with a very uniform magnetic field, so that T_2 (in the absence of a gradient) is only slightly greater than T_2^*. The gradient fields are larger, and the gradient pulse durations shorter, than in conventional imaging. The goal is to complete all the k-space measurements in a time comparable to T_2^*. In EPI the echoes are not created using π pulses. Instead, they are created by dephasing the spins at different positions along the x axis using a G_x gradient, and then reversing that gradient to rephase the spins, as shown in Fig. 18.32. Whenever the integral of $G_x(t)$ is zero, the spins are all in phase and the signal appears. A large negative G_y pulse sets the initial value of k_y to be negative; small positive G_y pulses ("blips") then increase the value of k_y for each successive k_x readout. Echo-planar imaging requires strong gradients—at least five times those for normal studies—so that the data can be acquired quickly. Moreover, the rise- and fall-times of these pulses are short, which induces large voltages in the coils. Eddy currents are also induced in the patient, and it is necessary to keep these below the threshold for neural activation. These problems can be reduced by using sinusoidally-varying gradient currents. The engineering problems are discussed in Schmitt et al. (1998); in Vlaardingerbroek and den Boer (2004); and in Bernsteinet al. (2004).

High spatial frequencies give the sharp edge detail in an image; the lowest spatial frequencies give the overall contrast. (We saw this in Figs. 12.9 and 12.10.) Changing the order of sampling points in k space can be useful. For example, when the image may be distorted by blood flow (see Sect. 18.11), it is possible to change the gradients in such a way that the values of k near zero are measured right after the excitation. This gives the proper signal within the volume of the vessel. The higher spatial frequencies, which show vessel edges, are less sensitive to blood flow and are acquired later.

A three-dimensional Fourier transform of the image can be obtained by selecting the entire sample and then phase encoding in both the y and z directions while doing frequency readout along x. One must step through all values of k_y for each value of k_z. This is one way to image very small samples with very high resolution (MRI microscopy) (Callaghan 1994).

18.9.6 Image Contrast and the Pulse Parameters

The appearance of an MR image can be changed drastically by adjusting the repetition time and the echo time. Problem 27 derives a general expression for the amplitude of the echo signal when a series of $\pi/2$ pulses are repeated every T_R seconds. The magnetic moment in the sample at the time of the measurement, considering both longitudinal and transverse relaxation, is

$$M(T_R, T_E) \tag{18.52}$$
$$= M_0 \left(1 - 2e^{-T_R/T_1 + T_E/2T_1} + e^{-T_R/T_1}\right) e^{-T_E/T_2},$$

where M_0 is proportional to the number of proton spins per unit volume N, as shown in Eq. 18.10. If $T_R \gg T_E$, this simplifies to

$$M(T_R, T_E) = M_0(1 - e^{-T_R/T_1})e^{-T_E/T_2}, \tag{18.53}$$

We consider an example that compares muscle ($M_0 = 1.02$ in arbitrary units, $T_1 = 500$ ms, and $T_2 = 35$ ms) with fat ($M_0 = 1.24$, $T_1 = 200$ ms, and $T_2 = 60$ ms).

Figure 18.33 shows two examples where T_R is relatively long and M_0 returns nearly to its initial value between pulses. If the echo time is short, then the image is nearly independent of both T_1 and T_2 and it is called a *density-weighted image*. If T_E is longer, then the transverse decay term dominates and it is called a *T_2-weighted image*. The signal is often weak and therefore noisy because there has been so much decay.

Figure 18.34 shows what happens if the repetition time is made small compared to T_1. This is a *T_1-weighted image* because the differences in T_1 are responsible for most of the difference in signal intensity. Notice also that the very first pulse nutates the full M_0 into the transverse plane, so an echo after the first pulse would give an anomalous reading. Echoes are measured only for the second and later pulses.

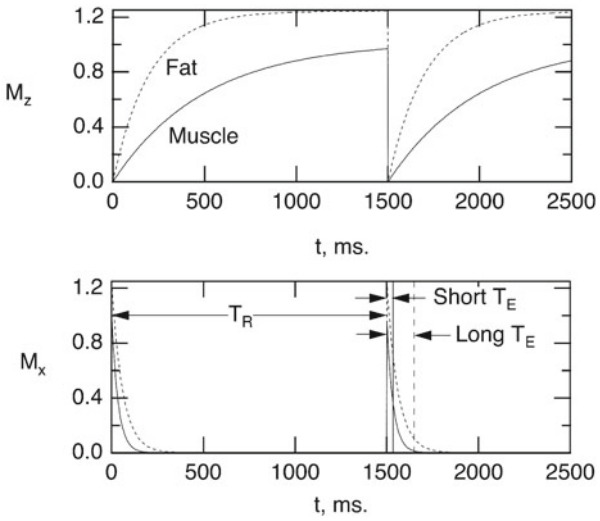


Fig. 18.33 The intensity of the signal from different tissues depends on the relationship between the repetition time and echo times of the pulse sequence, and the relaxation times of the tissues being imaged. This figure and the next show the magnetization curves for two tissues: muscle (relative proton density 1.02, $T_1 = 500$ ms, $T_2 = 35$ ms) and fat (relative proton density 1.24, $T_1 = 200$ ms, $T_2 = 60$ ms). The repetition time is 1500 ms, which is long compared to the longitudinal relaxation times. A long echo time gives an image density that is very sensitive to T_2 values. A short echo time (even shorter than shown) gives an image that depends primarily on the spin density

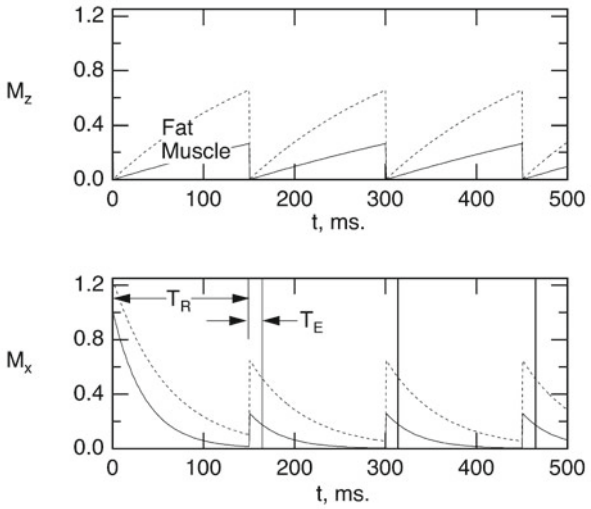


Fig. 18.34 The tissue parameters are the same as in Fig. 18.33. The repetition time is short compared to the longitudinal relaxation time. As a result, the first echo must be ignored. With a short T_E, the image density depends strongly on the value of T_1

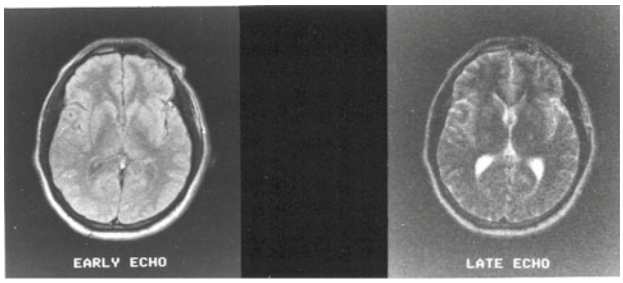


Fig. 18.35 Spin-echo images taken with short and long values of T_E, showing the difference in T_2 values for different parts of the brain. (Photograph courtesy of R. Morin, Ph.D., Department of Diagnostic Radiology, University of Minnesota)

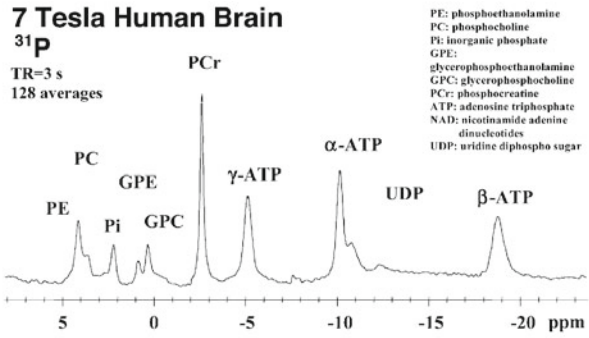


Fig. 18.36 A chemical shift spectrum for ^{31}P taken from the visual cortex at the back of the brain using a 7-tesla machine. (From Lei et al. 2003. Used by permission. Image courtesy of Prof. Kamil Ugurbil)

Suppose that the value of T_2 for fat had been shorter than the value for muscle. Then there would have been a value of T_E for which the two transverse magnetization curves crossed, and the two tissues would have been indistinguishable in the image. At larger values of T_E, their relative brightnesses would have been reversed. Figure 18.35 shows spin-echo images taken with two different values of T_E, for which the relative brightnesses are quite different.

18.9.7 Safety

Safety issues in MRI include forces on magnetic objects in and around the patient such as aneurysm clips, hairpins, pacemakers, wheel chairs, and gas cylinders (Kanal et al. 2007), absorbed radio-frequency energy (Problem 21), and induced currents from rapidly-changing magnetic field gradients. The rapid changes of magnetic field can stimulate nerves and muscles, cause heating in electrical leads and certain tattoos, and possibly induce ventricular fibrillation. Induced fields are reviewed by Schaefer et al. (2000). Cardiac pacemakers are being designed to be immune to the strong—and rapidly varying—magnetic and rf fields (Santini et al. 2013).

18.10 Chemical Shift

If the external magnetic field is very homogeneous, it is possible to detect a shift of the Larmor frequency due to a reduction of the magnetic field at the nucleus because of diamagnetic shielding by the surrounding electron cloud. The modified Larmor frequency can be written as

$$\omega = \gamma B_0(1 - \sigma). \tag{18.54}$$

Typical values of σ are in the range 10^{-5} to 10^{-6}. They are independent of B_0, as expected for a diamagnetic effect proportional to B_0. Measurements are made by Fourier transformation of the free-induction-decay signal, averaged over many repetitions if necessary to provide the resolution required to detect the shift.

A great deal of work has been done with ^{31}P, because of its presence in adenosine triphosphate and adenosine diphosphate (ATP and ADP). Free energy is supplied for many processes in the body by the conversion of ATP to ADP. Fig. 18.36 shows a very high resolution chemical shift spectrum from the human visual cortex taken with a 7-tesla machine.

It is also possible to make chemical shift images. Figure 18.37 shows a series of ^{31}P spectra from the brain. An image of the slice from which these data are obtained is shown below the spectra. The slice on the left cuts through the cerebellum and temporal lobes of the brain. It also includes some skeletal muscle. The slice on the right is through brain only.

18.11 Flow Effects

Flow effects can distort a magnetic resonance image. Spins initially prepared with one value of **M** can flow out of a slice before the echo and be replaced by spins that had a different initial value of **M**. This is called *washout*. Spins that have been shifted in phase by a field gradient can flow to another location before the readout pulse is applied. This causes artifacts and can also be used to measure blood flow (Axel 1984; Battocletti et al. 1981).

To understand the washout effect consider a simple model in which a blood vessel is perpendicular to the slice, as shown in Fig. 18.38. To simplify further, assume that all the blood

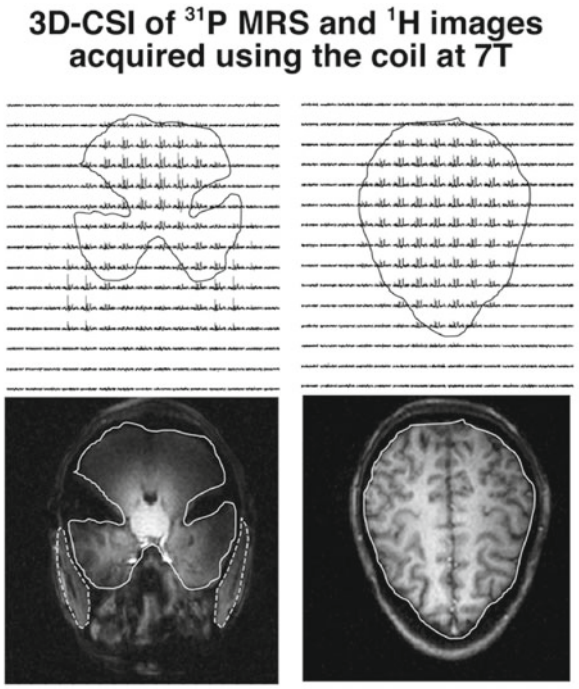


Fig. 18.37 The image on the left displays ^{31}P chemical shift data as spectra from individual voxels. The image of the slice from which these data are obtained is shown below the spectra. The slice cuts through the cerebellum and temporal lobes of the brain (solid outline). The dashed lines mark skeletal muscle which also contains phosphorylated metabolites, with a higher creatine phosphate level (PCr) compared to brain. The slice on the right is through the brain only. (Image courtesy of Prof. Kamil Ugurbil, University of Minnesota)

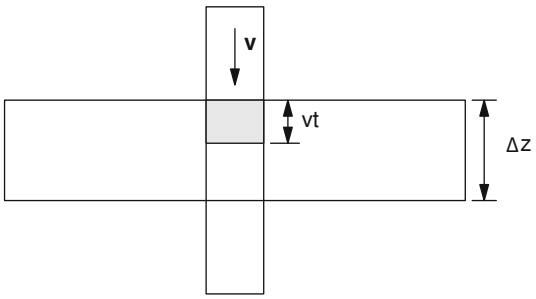


Fig. 18.38 A blood vessel is perpendicular to the slice. The model developed in the text assumes plug flow, that is, all of the blood is flowing with the same speed v

flows with the same speed v, independent of where it is in the vessel. This is called *plug flow*.

First consider washout of the excited spins. Suppose that at time $T_E/2$ a π pulse is applied to the slice in Fig. 18.38 and that the echo is measured at time T_E. The shaded area in the vessel represents new blood that flows in during time t. If the flow velocity is zero, no new blood flows in, all of the blood in the slice was excited, and the signal has full strength. If the velocity is greater than $2\Delta z/T_E$, all of the spins that were flipped by the pulse will leave the sensitive region by the time of the echo, and there will be no signal. Because we assume plug flow, the fraction washed out is a linear function of velocity up to the critical value of v. The fraction of excited spins remaining at T_E is given by

$$f = \begin{cases} 1 - \dfrac{vT_E}{2\Delta z}, & v < 2\Delta z/T_E \\ 0, & v \geq 2\Delta z/T_E. \end{cases} \tag{18.55}$$

Now consider washout of spins between pulses. We saw that the effect of repetition and echo times on the MRI signal is given by Eq. 18.52, which, if $T_R \gg T_E$, simplifies to Eq. 18.53. For low velocities ($v < \Delta z/T_R$) there is an enhancement of the signal because blood with a larger value of M_z flows into the sensitive region. For $vT_R < \Delta z$, the factor in parentheses in Eq. 18.53 is replaced by

$$\frac{vT_R}{\Delta z} + \left(1 - \frac{vT_R}{\Delta z}\right)\left(1 - e^{-T_R/T_1}\right).$$

The first term represents spins that flow in and the second those that still remain and that are still affected by the previous pulse. This can be rearranged as

$$\left(1 - e^{-T_R/T_1}\right) + \frac{vT_R}{\Delta z} e^{-T_R/T_1}. \tag{18.56}$$

This factor has the value $1 - e^{-T_R/T_1}$ for small v and is proportional to v when $v \gg \Delta z/T_R$. More complicated models can be developed. Phase changes because the blood flows through magnetic field gradients are also important.

Blood perfusion in the brain can be monitored using *arterial spin labeling* (Wolf and Detre 2007). A π pulse inverts the spins in a slice just upstream of the region of interest. Blood flow carries these labeled spins into the slice to be imaged. A second image of the slice is acquired without labeling the spins. The difference between the two images provides information about perfusion.

In addition to blood flow, MRI can also be used to image motion of the tissue. In *magnetic resonance elastography*, an acoustic signal is applied to the tissue (typically 0.1-1 kHz), creating a shear wave (Chap. 13). A magnetic resonance image is then obtained, using a magnetic field gradient that oscillates at the same frequency as the acoustic wave. In stationary tissue, the positive and negative phases produced by an oscillating gradient cancel to produce no net phase change, but in the oscillating tissue the phase shifts accumulate. Thus, information about the amplitude of the tissue motion is encoded in the phase of the magnetic resonance signal. When the applied signal and tissue response are both known, the shear modulus (Chap. 1) can be determined. If, for instance, a tumor is stiffer than the surrounding tissue, it will be imaged as a region of high shear modulus (Mariappan et al. 2010).

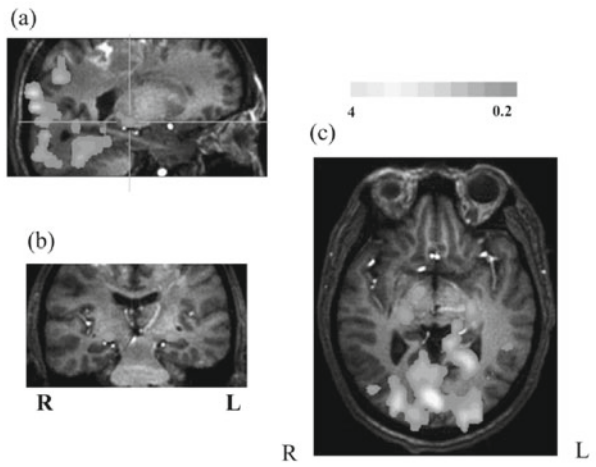


Fig. 18.39 Functional MRI in three planes: **a** sagittal (side) view; **b** coronal (front) view; **c** axial view (from below). The layers viewed in **b** and **c** are indicated by the lines in **a**. Bright spots superimposed on the image show activity in the visual cortex and in some structures between the eye and the visual cortex. The magnetic field is 4 T. (Adapted from Chen et al. 1998. Image supplied by Prof. Kamil Ugurbil)

18.12 Functional MRI

Magnetic resonance imaging provides excellent structural information. Various contrast agents can provide information about physiologic function. For example, contrast agents containing gadolinium are injected intravenously (Hao et al. 2012). They leak through a damaged blood-tissue barrier and accumulate in the damaged region. At small concentrations T_1 is shortened. One can also inject a contrast agent and watch its first pass through the circulatory system. Such an agent typically changes the magnetic susceptibility and shortens T_2.

The term *functional magnetic resonance imaging* (fMRI) usually refers to a technique developed in the 1990s that allows one to study structure and function simultaneously. The basis for fMRI is inhomogeneities in the magnetic field caused by the differences in the magnetic properties of oxygenated and deoxygenated hemoglobin. No external contrast agent is required. Oxygenated hemoglobin is less paramagnetic than deoxyhemoglobin. If we make images before and after a change in the blood flow to a small region of tissue (perhaps caused by a change in its metabolic activity), the difference between the two images is due mainly to changes in the blood oxygenation. One usually sees an increase in blood flow to a region of the brain when that region is active. This BOLD contrast in the two images provides information about the metabolic state of the tissue, and therefore about the tissue function (Ogawa et al. 1990; Kwong et al. 1992).

An image of the brain during visual stimulation is shown in Fig. 18.39. In addition to the visual cortex in part c, activity is seen in the lateral geniculate nucleus (parts b and c), which is on the pathway from the eye to the visual cortex. Functional MRI provides functional information similar to that from PET (Sect. 17.10), but without the need for radionuclides.

Other contrast agents, usually a complex molecule shielding a gadolinium atom, are being developed to measure pH, ions such as zinc, calcium, and copper, and certain enzymes (Louie 2013). The stable isotope ^{19}F is being tested as an alternative to gadolinium (Ahrens and Zhong 2013).

Another recent technique that can be classified as functional is the detection of prostate cancer that has metastasized to a lymph node when the metastasis is not yet apparent by other imaging techniques. Monocrystalline iron oxide particles injected in the blood will be taken up by normal lymph nodes but not those with metastases. The technique is effective for lymph nodes larger than 5 mm (Harisinghani et al. 2003; see also the commentary by Koh et al. 2003).

Much recent research has focused on using MRI to image neural activity directly, rather than through changes in blood flow (Bandettini et al. 2005). Two methods have been proposed to do this. In one, the biomagnetic field produced by neural activity (Chap. 8) acts as the contrast agent, perturbing the magnetic resonance signal. Images with and without the biomagnetic field present provide information about the distribution of neural action currents. In an alternative method, the Lorentz force (Eq. 8.2) acting on the action currents in the presence of a magnetic field causes the nerve to move slightly. If a magnetic field gradient is also present, the nerve may move into a region having a different Larmor frequency. Again, images taken with and without the action currents present provide information about neural activity. Unfortunately, both the biomagnetic field and the displacement caused by the Lorentz force are tiny, and neither of these methods has yet proved useful for neural imaging. However, if these methods could be developed, they would provide information about brain activity similar to that from the magnetoencephalogram, but without requiring the solution of an ill-posed inverse problem that makes the MEG so difficult to interpret.

18.13 Diffusion and Diffusion Tensor MRI

Our analysis of MRI so far assumes that the nuclei are stationary except for the rotation of their spin axis or their motion with the blood to create flow effects. In practice, these nuclei are also free to diffuse throughout the tissue (Chap. 4). The magnetization $\mathbf{M}$ depends on the total number of particles per unit volume with average spin components $\langle \mu_x \rangle$,

$\langle \mu_y \rangle$, and $\langle \mu_z \rangle$. In the rotating coordinate system there is no precession. In the absence of relaxation effects $\langle \boldsymbol{\mu} \rangle$ does not change. In that case changes in $\mathbf{M}$ depend on changes in the concentration of particles with particular components of $\langle \boldsymbol{\mu} \rangle$, so the rate of change of each component of $\langle \boldsymbol{\mu} \rangle$ is given by a diffusion equation. For example, for M_x,

$$\frac{\partial M_x}{\partial t} = D\nabla^2 M_x.$$

If the processes are linear, this diffusion term can be added to the other terms in the Bloch equations. The details are explored in Problem 47.

In a spin–echo pulse sequence, the amplitude of the echo will be smaller if the spins have diffused to different locations within the tissue between the time of the excitation pulse and the echo. This artifact degrades the signal during traditional MRI, but can be valuable if one wants to measure the diffusion constant. The rate of diffusion depends sensitively on temperature, so measurements of the diffusion constant provide a way to monitor internal temperatures noninvasively (Delannoy et al. 1991). Moseley et al. (1990) showed that diffusion MRI is valuable for detecting regional cerebral ischemia, and it has become a useful tool in stroke research.

Diffusion can be monitored during a spin-echo sequence by applying magnetic field gradients of the same magnitude and duration before and after the π pulse, as shown in Fig. 18.40. If a spin is stationary, these gradients have no effect: they shift the phase of the spins in one direction before the π pulse, but shift the phase in the other direction after the π pulse, restoring the original phase. However, for spins that diffuse from one location to another between the application of the gradients, the phase shift of the first gradient is not cancelled by an opposite phase shift in the second, so the gradients introduce a net phase shift. This shift lowers the echo amplitude, with the reduction depending on the square of the gradient and the diffusion constant (Prob. 47).

In some tissues diffusion is anisotropic, meaning that the diffusion constant depends on direction. In such cases the effect of diffusion depends on the direction of the magnetic field gradient. Basser et al. (1994) extended diffusion MRI so that the entire diffusion tensor is measured. The diffusion tensor (or matrix) is similar to the conductivity tensor discussed in Sect. 7.9. Using matrix notation, the fluence rate of diffusing particles with aligned nuclear spins is related to the particle concentration by

$$\begin{pmatrix} j_x \\ j_y \\ j_z \end{pmatrix} = - \begin{pmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{pmatrix} \begin{pmatrix} \dfrac{\partial C}{\partial x} \\ \dfrac{\partial C}{\partial y} \\ \dfrac{\partial C}{\partial z} \end{pmatrix}. \qquad (18.57)$$

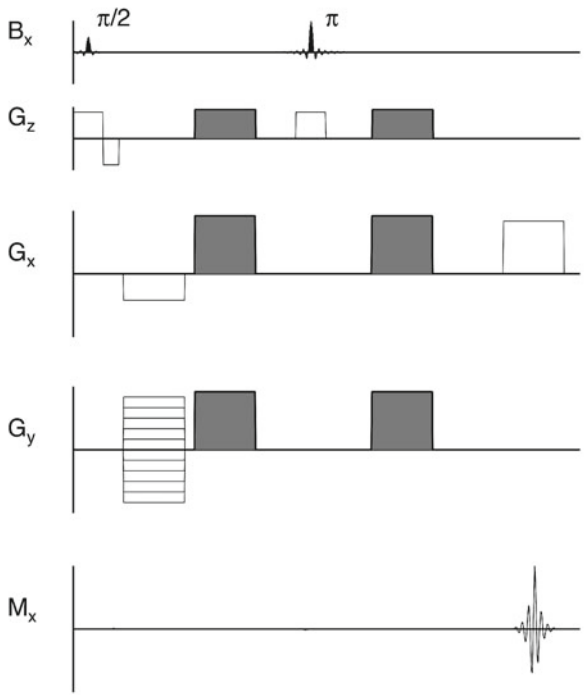


Fig. 18.40 A silmplified pulse sequence for diffusion tensor imaging. The sequence is similar to that shown in Fig. 18.30 for two-dimensional imaging using phase encoding. The M_x signal is shown only during readout. The diffusion gradients, shown in gray, are applied before and after the π pulse. For stationary spins, any phase shift produced by the first diffusion gradient is canceled by an opposite phase shift produced by the second diffusion gradient. Spins that diffuse during this pulse sequence are affected differently by the first and second diffusion gradients, which affects the signal. For diffusion tensor imaging the gradients must be applied in all three directions. For more details, see Mattiello et al. (1994)

One can show that the diffusion matrix is always symmetric: $D_{yx} = D_{xy}$, etc.

Diffusion is usually greater along the direction of the nerve or muscle fibers. Since the orientation of the fibers changes throughout the body, the elements of the diffusion tensor vary as well. However, some features of the diffusion tensor, such as the trace (see Prob. 49), are independent of the fiber direction, and are particularly useful when monitoring diffusion in anisotropic tissue, such as the white matter of the brain. In addition, the diffusion tensor contains information about the fiber direction, allowing one to map fiber tract trajectories noninvasively using MRI (Basser et al. 2000). See also the review by Thomas et al. (2000).

18.14 Hyperpolarized MRI of the Lung

The lung is difficult to image using MRI because of its low proton density. A new way to monitor lung function is to image the isotope ^{129}Xe inhaled into the lungs. The density

of ^{129}Xe is small, but the magnetization can be increased dramatically using the technique of *hyperpolarization* (Mugler and Altes 2013). In this two-step process, a laser is used to generate electron-spin polarization in a vapor of alkali metal such as rubidium subject to a magnetic field. Then collisions with the rubidium molecules transfer the polarization to the ^{129}Xe. This technique increases the polarization by a factor of about 1000 beyond what it would be in thermal equilibrium.

Symbols Used in Chapter 18

Symbol	Use	Units	First used page
a	Loop radius	m	545
a, b	Constants	J T^{-1} m^{-3}	541
f	Fraction		556
h	Planck's constant	J s	542
$\hbar$	Planck's constant (reduced)	J s	537
i	Current	A	536
j_x, j_y, j_z	fluence rate	m^{-2} s^{-1}	558
k_B	Boltzmann constant	J K^{-1}	537
k_x, k_y, k_z	Spatial frequency	m^{-1}	551
m	Mass	kg	537
m	Azimuthal quantum number		537
m	Integer		552
q	Electric charge	C	537
r	Radius	m	537
s	Signal	V	551
t	Time	s	536
v	Velocity	m s^{-1}	537
v	Voltage difference	V	544
x	Dimensionless variable		549
x, y, z	Axes	m	538
x', y', z'	Axes (rotating)	m	540
Δz	Slice thickness	m	549
A	Amplitude	T s	550
A	Constant	V T J^{-1}	550
B, $\mathbf{B}$	Magnetic field	T	535
B_1	Oscillating magnetic field	T	540
C	Constant in expression for relaxation time	s^{-2}	543
$C(k), S(k)$	Fourier transforms	J T^{-1} m^{-1}	551
C	Concentration	m^{-3}	558
D_{xx}, etc.	Components of diffusion tensor	m^2 s^{-1}	558
E	Energy	J	542
G_x, G_y, G_z	Gradient of B_z in the x, y, or z direction	T m^{-1}	549
$\mathbf{I}$	Nuclear angular momentum	kg m^2 s^{-1}	537
I	Nuclear angular momentum quantum number		537
K	Constant		544
L, $\mathbf{L}$	Orbital angular momentum	kg m^2 s^{-1}	536
M, $\mathbf{M}$	Magnetization	J T^{-1} m^{-3}	537
N	Number of spins per unit volume	m^{-3}	537
$\mathbf{R}$	Rotation matrix		540
S	Area	m^2	536
$\mathbf{S}$	Spin angular momentum	kg m^2 s^{-1}	537
T	Temperature	K	537
T	Period	s	541
T_E	Time of echo	s	547
T_I	Interrogation time	s	546
T_R	Repetition time between pulse sequences	s	548
T_1	Longitudinal relaxation time	s	538
T_2	Transverse relaxation time	s	538
T_2^*	Experimental transverse relaxation time	s	544
T_p	Length of gradient pulse	s	552
U	Potential energy	J	536
V	Volume	m^3	545
α	Arbitrary angle		542
γ	Gyromagnetic ratio	T^{-1} s^{-1}	536
μ, $\boldsymbol{\mu}$	Magnetic moment	J T^{-1}	535
μ_0	Magnetic permeability of space	T m A^{-1}	535
ν	Frequency	Hz	537
θ	Angle		536
σ	Chemical shift factor		555
τ, $\boldsymbol{\tau}$	Torque	N m	535
τ	Shift time for autocorrelation	s	543
τ_C	Correlation time	s	543
ω	Angular frequency	s^{-1}	537
ω_1	Angular frequency for B_1 rotation	s^{-1}	541
ω_0	Larmor angular frequency	s^{-1}	539
ϕ	Azimuthal angle		543
ϕ	Phase		552
ϕ_{11}	Autocorrelation function		543
Φ	Magnetic flux	weber (T m^2)	545
$\boldsymbol{\Omega}$	Angular velocity vector	s^{-1}	540

Problems

Section 18.1

Problem 1. Show that for a particle of mass m located at position $\mathbf{r}$ with respect to the origin, the torque about the origin is the rate of change of the angular momentum about the origin.

Section 18.2

Problem 2. Show that the units of γ are T^{-1} s^{-1}.

Problem 3. Find the ratio of the gyromagnetic ratio in Table 18.1 to the value $q/2m$ for the electron and proton.

Section 18.3

Problem 4. Evaluate the quantity $\gamma m\hbar B/k_BT$ and the Larmor frequency for electron spins and proton spins in magnetic fields of 0.5 and 4.0 T at body temperature (310 K).

Problem 5. Verify that $\sum 1 = 2I + 1$, $\sum m = 0$, and $\sum m^2 = I(I+1)(2I+1)/3$, when the sums are taken from $-I$ to I, in the cases that $I = \frac{1}{2}$, 1, and $\frac{3}{2}$.
Problem 6. Obtain an expression for the magnetization analogous to Eq. 18.10 in the case $I = \frac{1}{2}$ when one cannot make the assumption $\gamma\hbar B/k_B T \ll 1$.
Problem 7. Calculate the coefficient of B in Eq. 18.10 for a collection of hydrogen nuclei at 310 K when the number of hydrogen nuclei per unit volume is the same as in water.

Section 18.4

Problem 8. Verify that Eqs. 18.16 are a solution of Eqs. 18.15.
Problem 9. Calculate the value of $M_x^2 + M_y^2 + M_z^2$ for relaxation Eqs. 18.16 when $T_1 = T_2$.
Problem 10. Equations 18.16 correspond to a solution of the Bloch equations in the presence of a static field B for one initial condition: $M_x = M_0$, $M_y = 0$, $M_z = 0$. Solve the Bloch equations for a different initial condition: $M_x = 0$, $M_y = 0$, and $M_z = -M_0$.

Section 18.5

Problem 11. (a) Use Fig. 18.6 to derive Eq. 18.18.
(b) Show that

$$M_{x'} = M_x \cos\theta + M_y \sin\theta,$$

$$M_{y'} = -M_x \sin\theta + M_y \cos\theta.$$

(c) Combine these equations with the equations for M_x and M_y to show that the application of both transformations brings one back to the starting point.
Problem 12. Equation 18.17 shows how to transform the components of a vector in the primed system (rotated an angle θ clockwise from the unprimed system) into the unprimed system. Use the arguments of Section 18.5 to derive the following transformation matrices for counterclockwise rotations.
(a) Angle α about the x axis:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}$$

(b) Angle β about the y axis:

$$\begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix},$$

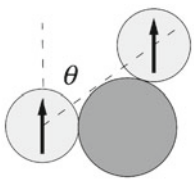
(c) Angle θ about the z axis:

$$\begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Why are the minus signs different from those in Eq. 18.17b?
Problem 13. Calculate $M^2 = M_x^2 + M_y^2 + M_z^2$ for the solution of Eqs. 18.30 and compare it to the results of Problem 9.

Section 18.6

Problem 14. Use Eqs. 18.32 to find the magnetic field at one proton due to the other proton in a water molecule when both proton spins are parallel to each other and perpendicular to the line between the protons. The two protons form an angle of 104.5° and are each 96.5×10^{-12} m from the oxygen.
Problem 15. The magnetic field at a distance of 0.15 nm from a proton is 4×10^{-4} T. What change in Larmor frequency does this ΔB cause? How long will it take for a phase difference of π radians to occur between a precessing spin feeling this extra field and one that is not?
Problem 16. Consider a collection of spins that are aligned along the x axis at $t = 0$. They precess in the xy plane with different angular frequencies spread uniformly between $\omega - \Delta\omega/2$ and $\omega + \Delta\omega/2$. If the total magnetic moment per unit volume is M_0 at $t = 0$, show that at time $T = 4/\Delta\omega$ it is $M_0 \sin(2)/2 = 0.455 M_0$.
Problem 17. What is the contribution to the transverse relaxation time for a magnetic field of 1.5 T with a uniformity of 1 ppm? The nonrecoverable relaxation time of brain is about 2.5 ms. What dominates the measured transverse relaxation in brain?
Problem 18. Suppose the two dipoles of the water molecule shown below point in the z direction while the line between them makes an angle θ with the x axis. Determine the angle θ for which the magnetic field of one dipole is perpendicular to the dipole moment of the other. For this angle the interaction energy is zero. This θ is called the *magic angle* and is used when studying anisotropic tissue such as cartilage (Xia 2000).



Problem 19. Using Eq. 18.34, determine the value of the minimum correlation time as a function of the Larmor frequency ω_0.
Problem 20. Redraw the plot in Fig. 18.12, assuming protons and your static magnetic field is either 1.5 or 4 T. If the

correlation time is approximately 3 ns, estimate T_1 and T_2. Explain how T_1 and T_2 depend on the magnetic field.

Section 18.7

Problem 21. In solving this problem, you will develop a simple model for estimating the radio-frequency energy absorption in a patient undergoing an MRI procedure.

(a) Consider a uniform conductor with electrical conductivity σ. If it is subject to a changing magnetic field $B_1(t) = B_1 \cos(\omega_0 t)$, apply Eq. 8.21 to a circular path of radius R at right angles to the field to show that the electric field at radius R has amplitude $E_0 = R\omega_0 B_1/2$. (Because this is proportional to R, the model gives the skin dose, along the path for which R is largest.)

(b) Use Ohm's law in the form $j = \sigma E$ to show that the time average power dissipated per unit volume of material is $p = \sigma E_0^2/2 = \sigma R^2 \omega_0^2 B_1^2/8$ and that if the mass density of the material is ρ, the *specific absorption rate* (SAR) or dose rate is $SAR = \sigma R^2 \omega_0^2 B_1^2/8\rho$.

(c) If the radio-frequency signal is not continuous but is pulsed, show that this must be modified by the *duty cycle* factor $\Delta t/T_R$, where Δt is the pulse duration and T_R is the repetition period.

(d) Combine these results with the fact that rotation through an angle θ (usually π or $\pi/2$) in time Δt requires $B_1 = 2\theta/\gamma\Delta t$ and that $\omega_0 = \gamma B_0$, to obtain $SAR = (1/T_R\Delta t)(\sigma/2\rho)R^2 B_0^2 \theta^2$.

(e) Use typical values for the human body—$R = 0.17$ m, $\sigma = 0.3$ S m^{-1}—to evaluate this expression for a $\pi/2$ pulse.

(f) For $B_0 = 0.5$ T and $SAR < 0.4$ W kg^{-1} determine the minimum value of Δt for $T_R = 1$ s. Also find B_1.

(g) For π pulses, what is the dose in Gy? (This should not be compared to an x-ray dose because this is nonionizing radiation.)

Problem 22. Use Eq. 18.38 to calculate the initial amplitude of a signal induced in a one-turn coil of radius 0.5 m for protons in a 1-mm cube of water at 310 K in a magnetic field of 1.0 T. (The answer will be a signal too small to be useful; multiple-turn coils must be used.)

Problem 23. Consider increasing B_0 from 4 T to 7 T. Discuss what changes this will make in

(a) The frequency of the RF pulse,
(b) the signal recorded by the detection coil,
(c) the specific absorption rate (see Problem 21),
(d) the skin depth for magnetic field penetration (see Chap. 8, Problem 29), and
(e) the values of T_1 and T_2.

Section 18.8

Problem 24. Plot the maximum amplitude of an inversion recovery signal vs the interrogation time if the detector is sensitive to the sign of the signal and if it is not.

Problem 25. (a) Obtain an analytic expression for the maximum value of the first and second echo amplitudes in a Carr–Purcell pulse sequence in terms of T_2 and T_E.
(b) Repeat for a CPMG pulse sequence.

Problem 26. Consider the behavior of M_z in Figs. 18.19 and 18.21. The general equation for M_z is $M_z = M_0 + Ae^{-t/T_1}$. After several π pulses, the value of M_z is flipping from $-b$ to b. Find the value of b.

Problem 27. Consider a spin–echo pulse sequence (Fig. 18.18). Find

(a) M_z just before the π pulse at $T_E/2$,
(b) M_z just after the π pulse at $T_E/2$,
(c) M_z just before the $\pi/2$ pulse at T_R, and
(d) the first and second echo amplitudes as a function of T_E, T_R, T_1 and T_2. (The second amplitude is the same as all subsequent amplitudes.)

Problem 28. This problem uses matrices to analyze the spin-echo pulse sequence. Use the rotation matrices given in Problem 12. Start with $\mathbf{M} = (0, 0, M_0)$. Rotate $\mathbf{M}$ about x' by $\pi/2$, then about z' by θ, then about x' by π, and finally about z' by θ. What are the final components of $\mathbf{M}$? Identify what pulse sequence or physical process corresponds to each rotation. Why would θ be nonzero in the rotating reference frame? What would be the significance if the final $\mathbf{M}$ is independent of θ?

Problem 29. (a) Make a three-dimensional sketch of Fig. 18.17. Assume spin a is initially aligned with the y' axis and spin b is 30° clockwise from spin a. Then make similar sketches for a Carr–Purcell sequence that rotates the spins about the x' axis at the following times: just before the π pulse at $T_E/2$, just after the π pulse at $T_E/2$, at T_E, just before the π pulse at $3T_E/2$, just after the π pulse at $3T_E/2$, and at $2T_E$. Assume that the π pulse rotates the spins exactly 180°. Then make sketches when the π pulses rotate the spins by 185°.

(b) Repeat for a CPMG pulse sequence that rotates spin a and spin b around the y' axis. Again, consider two cases: the π pulses rotate by 180° and 185°. Your sketches will show the advantage of the CPMG pulse sequence when there is an error in the duration of the π pulse.

Problem 30. This problem uses the matrices introduced in Problem 12 to examine the difference between the Carr–Purcell and the Carr–Purcell–Meiboom–Gill pulse sequences.

(a) Start with $\mathbf{M} = (0, 0, M_0)$. Rotate about x' by $\pi/2$, about z' by θ, about x' by π, about z' by 2θ, about x' by π, and about z' by θ. What is the final result? This

process corresponds to the first two echoes produced by a Carr–Purcell pulse sequence.

(b) Repeat the analysis of part (a), but change the two π rotations about x' to two $\pi + \delta$ rotations about x'. Assume $\delta \ll \pi$ and use the approximations $\cos(\pi + \delta) = -\cos\delta \approx -1$ and $\sin(\pi + \delta) = -\sin\delta \approx -\delta$ to simplify your result. Keep only terms in order δ. What is your final result? This process corresponds to the first two echoes produced by a Carr–Purcell pulse sequence in which the π pulses have slightly wrong amplitudes.

(c) Repeat the analysis of part (b) but change the rotations about x' to be rotations about y'. What are the differences between the CP and CPMG pulse sequences? Explain why the CPMG pulse sequence is superior to the CP pulse sequence.

Section 18.9

Problem 31. Show that an alternative expression for the field amplitude required for a $\pi/2$ pulse is $B_1 = B_0\pi/\omega_0\Delta t = B_0/2\nu\Delta t$.

Problem 32. A certain MRI machine has a static magnetic field of 1.0 T. Spins are excited while applying a field gradient of 3 mT m^{-1}. If the slice is to be 5 mm thick, what is the Larmor frequency and the spread in frequencies that is required?

Problem 33. Consider a pair of gradient coils of radius a perpendicular to the z axis and located at $z = \pm\sqrt{3}a/2$. The current flows in the opposite direction in each single-turn coil.

(a) Use the results of Problem 8.10 to obtain an expression for B_z along the z axis.

(b) For a gradient of 5 mT m^{-1} at the origin and $a = 10$ cm, find the current required in a single-turn coil.

Problem 34. Find a linear approximation for Eq. 18.53 for very small values of T_E and T_R, and discuss why it is called a T_1-weighted image.

Problem 35. The slice selection gradient G_z must be applied for a time τ which is at least as long as the duration of the B_1 pulse. Suppose that $\tau = 6 \times 2\pi/(\gamma G_z \Delta z)$ (see Fig. 18.23). How much has the phase at the top of the slice ($z = \Delta z$) changed with respect to the middle of the slice ($z = 0$)?

Problem 36. Relate the resolution in the y direction to G_y and T_p.

Problem 37. Discuss the length of time required to obtain a 256×256 image in terms of T_R and T_E. The field of view is 15 cm square. Consider both projection reconstruction and spin warp images. Introduce any other parameters you need.

Problem 38. The limiting noise in a well-designed machine is due to thermal currents in the body. The noise is proportional to B_0 and the volume V_n sampled by the radio-frequency pickup coil. The noise is also proportional to $T^{-1/2}$, where T is the time it takes to acquire the image. Show that the signal-to-noise ratio is proportional to $B_0 T^{1/2} V_v / V_n$, where V_v is the volume of the picture element.

Problem 39. Explain in words why in Fig. 18.24 a negative lobe for G_z to eliminate unwanted phase shifts is not needed following the π pulse, although it is needed following the initial $\pi/2$ pulse.

Problem 40. The readout gradient G_x shown in Fig. 18.26 not only resolves the echo into its frequency components, but also introduces a phase shift. In more detailed analyses the readout gradient consists of two parts: a *prephasing lobe* and a *readout lobe*. Modify Fig. 18.26 to include a prephasing lobe in G_x between the $\pi/2$ and π pulses, so that the net phase shit at the peak of the echo caused by G_x is zero. Pay attention to the amplitude, duration, and polarity of the pulse.

Problem 41. In this book, gradient pulses are drawn as rectangles when showing a pulse sequence. However, there is often a limit, called the *maximum slew rate*, to how fast a gradient can change. Consider a trapezoidal pulse (linear rise, then constant, then linear fall). What is the shortest rise time of the pulse if it has a peak gradient of 30 mT m^{-1} and a maximum slew rate of 100 T m^{-1} s^{-1}?

Problem 42. Suppose one is imaging using the projection reconstruction algorithm shown in Fig. 18.28. After the echo from the initial gradient, when one is ready to repeat the sequence using a different gradient, there may be some residual transverse magnetization that could affect the subsequent signal. Explain why a large G_x gradient, called a *spoiler gradient*, applied after the echo in Fig. 18.28 would eliminate any remaining transverse magnetization.

Section 18.10

Problem 43. The chemical shift difference between water and fat is $\Delta\sigma = 3.5$ ppm. This can cause a spatial shift of the images from fat and water if the readout gradients are large. Estimate this shift for a 1.5 T machine and a gradient of 5 mT m^{-1}.

Section 18.11

Problem 44. Use the model of Sect. 18.11 to plot the flow correction as a function of velocity for $T_E = 10$ ms, $T_1 = 900$ ms, and $T_2 = 400$ ms, when (a) $T_R = 50$ ms, (b) $T_R = 200$ ms.

Problem 45. Excite the spins using a sinc $\pi/2$ pulse and a G_{z1} gradient so that spins in slice $z1$ are in resonance. Then

apply a π pulse, but with a different gradient G_{z2} so the spins are flipped in a different, nonoverlapping slice $z2$.

(a) If the spins are stationary, what signal do you observe?
(b) If the spins move (for instance are carried by flowing blood) from $z1$ to $z2$ during the time between the two RF pulses, what signal do you see?
(c) Design a pulse sequence for performing this experiment.

Problem 46. Suppose your median nerve (the primary nerve in your arm) carries a current I along its length L.

(a) You are having a magnetic resonance image taken, and the steady uniform magnetic field B is directed perpendicular to the nerve. Derive an expression for F, the magnitude of the magnetic force on the nerve. Draw a picture showing the directions of I, **B**, and **F**.
(b) Assume this nerve is held in position by an elastic force per unit length with magnitude equal to kr, where k is the spring constant per unit length and r is the distance the nerve is displaced from its equilibrium position. Find an expression for the displacement of the nerve.
(c) Assume that a magnetic field gradient G is present, so that when the nerve moves a distance r it leaves a region with magnetic field strength B and enters a region of magnetic field strength $B+Gr$. Derive an expression for the change in resonance angular frequency $\Delta\omega$ caused by the displacement, in terms of G, B, I, k, and the gyromagnetic ratio of the proton, γ. (Hint: $\Delta\omega = \gamma\Delta B$). If the gradient and current last for time T, what is the change in phase of the MRI signal?
(d) Calculate the distance that the axon moves if $B_0 = 4$ T, $I = 0.1$ mA, and $k = 40,000$ N m^{-2}. Calculate the resulting phase shift (in degrees) if $G = 36$ mT m^{-1}, $T = 10$ ms, and $\gamma = 2.68 \times 10^8$ rad s^{-1} T^{-1}.

Section 18.13

Problem 47. This problem shows how to extend the Bloch equations to include the effect of diffusion of the molecules containing the nuclear spins in an inhomogeneous external magnetic field. Since **M** is the magnetization per unit volume, it depends on the total number of particles per unit volume with average spin components $\langle\mu_x\rangle$, $\langle\mu_y\rangle$, and $\langle\mu_z\rangle$. In the rotating coordinate system there is no precession. In the absence of relaxation effects $\langle\boldsymbol{\mu}\rangle$ does not change. In that case changes in **M** depend on changes in the concentration of particles with particular components of $\langle\boldsymbol{\mu}\rangle$, so the rate of change of each component of $\langle\boldsymbol{\mu}\rangle$ is given by a diffusion equation. For example, for M_x,

$$\frac{\partial M_{x'}}{\partial t} = D\nabla^2 M_{x'}.$$

If the processes are linear this diffusion term can be added to the other terms in the Bloch equations. Suppose that there is a uniform gradient in B_z, G_z, and that the coordinate system rotates with the Larmor frequency for $z = 0$. When z is not zero, the rotation term does not quite cancel the $(\mathbf{M} \times \mathbf{B})_z$ term.

(a) Show that the x and y Bloch equations become

$$\frac{\partial M_{x'}}{\partial t} = +\gamma G_z z M_{y'} - \frac{M_{x'}}{T_2} + D\nabla^2 M_{x'},$$

$$\frac{\partial M_{y'}}{\partial t} = -\gamma G_z z M_{x'} - \frac{M_{y'}}{T_2} + D\nabla^2 M_{y'}.$$

(b) Show that in the absence of diffusion

$$M_{x'} = M(0)e^{-t/T_2}\cos(\gamma G_z z t),$$

$$M_{y'} = -M(0)e^{-t/T_2}\sin(\gamma G_z z t).$$

(c) Suppose that **M** is uniform in all directions. At $t = 0$ all spins are aligned. Spins that have been rotating faster in the plane at $z + \Delta z$ will diffuse into plane z. Equal numbers of slower spins will diffuse in from plane $z - \Delta z$. Show that this means that the phase of **M** will not change but the amplitude will.
(d) It is reasonable to assume that the amplitude of the diffusion-induced decay will not depend on z as long as we are far from boundaries. Therefore try a solution of the form

$$M_{x'} = M(0)e^{-t/T_2}\cos(\gamma G_z z t)\, A(t),$$

$$M_{y'} = M(0)e^{-t/T_2}\sin(\gamma G_z z t)\, A(t),$$

and show that A must obey the differential equation

$$\frac{1}{A}\frac{dA}{dt} = -D\gamma^2 G_z^2 t^2,$$

which has a solution $A(t) = \exp(-D\gamma^2 G_z^2 t^3/3)$.
(e) Show that if there is a rotation about y' at time $T_E/2$, then at time T_E, M_x is given by

$$M_x(T_E) = -M_0\exp(-T_E/T_2)\exp(-D\gamma^2 G_z^2 T_E^3/12).$$

Hint: This can be done formally from the differential equations. However it is much easier to think physically about the meaning of each factor in the expressions shown in (d) for $M_{x'}$ and $M_{y'}$. This result means that a CPMG sequence with short T_E intervals can reduce the effect of diffusion when there is an external gradient.

Problem 48. A commercial MRI machine is operated with a magnetic gradient of 3 mT m^{-1} while a slice is being selected. What is the effect of diffusion? Use the diffusion

constant for self-diffusion in water and the results of Problem 47. Compare the correction factor to $\exp(-T_E/T_2)$ when $T_2 = 75$ ms.

Problem 49. When a coordinate system is rotated as in Fig. 18.6, the diffusion tensor or diffusion matrix, which is always symmetric, transforms as

$$\begin{pmatrix} D_{x'x'} & D_{x'y'} \\ D_{x'y'} & D_{y'y'} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \times \begin{pmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

We have not proved this; note that the right-most matrix is the same one that would be seen if Eq. 18.17 were written in matrix form:

$$\begin{pmatrix} M_x \\ M_y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} M_{x'} \\ M_{y'} \end{pmatrix}.$$

(a) Perform the matrix multiplication and find expressions for $D_{x'x'}$, $D_{x'y'}$, and $D_{y'y'}$ in terms of D_{xx}, D_{xy}, D_{yy}, and θ.

(b) Find the angle θ such that $D_{x'y'}$ is zero (the diffusion tensor is diagonal). This is equivalent to finding the orientation of the fibers in the tissue.

(c) The *trace* of a matrix is the sum of its diagonal elements. Show that the trace of the diffusion matrix in the rotated coordinates, $D_{x'x'} + D_{y'y'}$, is equal to the trace of the diffusion matrix in the original coordinates, $D_{xx} + D_{yy}$. Thus, the trace of the diffusion tensor is independent of fiber direction.

References

Ahrens ET, Zhong J (2013) In vivo MRI cell tracking using perfluorocarbon probes and fluorine-19 detection. NMR Biomed 26: 860–871. doi:10.1002/nbm.2948

Axel L (1984) Blood flow effects in magnetic resonance imaging. Am J Roentgenol 143:1157–1166

Bandettini PA, Petridou N, Bodurka J (2005) Direct detection of neuronal activity with MRI: fantasy, possibility, or reality? Appl Magn Reson 29:65–88

Basser PJ, Mattiello J, LeBihan D (1994) MR diffusion tensor spectroscopy and imaging. Biophys J 66:259–267

Basser PJ, Pajevic S, Pierpaoli C, Duda J, Aldroubi A (2000) In vivo fiber tractography using DT-MRI data. Magn Reson Med 44: 625–632

Battocletti JH, Halbach RE, Salles-Cunha SX (1981) The NMR blood flow meter—theory and history. Med Phys 8:435–443

Bernstein MA, King KF, Joe Zhou X-h (2004). Handbook of MRI pulse sequences. Elsevier Academic Press, Amsterdam

Brown RW, Cheng Y-CN, Haacke EM, Thompson MR, Venkatesan R (2014) Magnetic resonance imaging: physical properties and sequence design. Wiley-Blackwell, Hoboken

Callaghan P (1994) Principles of nuclear magnetic resonance microscopy. Oxford University Press, Oxford

Chen W, Kato T, Zhu X-H, Strupp J, Ogawa S, Ugurbil K (1998) Mapping of lateral geniculate nucleus activation during visual stimulation in the human brain using fMRI. Magn Reson Med 39(1):89–96

Cho Z-H, Jones JP, Singh M (1993) Foundations of medical imaging. Wiley-Interscience, New York

Delannoy J, Chen CN, Turner R, Levin RL, LeBihan D (1991) Noninvasive temperature imaging using diffusion MRI. Magn Reson Med 19:333–339

Hao D, Ai T, Goerner F, Hu X, Runge VM, Tweedle M (2012) MRI contrast agents: Basic chemistry and safety. J Mag Res Imaging 36:1060–1071. doi:10:1002/jmri.23725

Harisinghani MG, Barentsz J, Hahn PF, Deserno WM, Tabatabaei S, van de Kaa CH, de la Rosette J, Weissleder R (2003) Noninvasive detection of clinically occult lymph-node metastases in prostate cancer. N Engl J Med 348(25):2491–2499

Joseph PM, Axel L, O'Donnell M (1984) Potential problems with selective pulses in NMR imaging systems. Med Phys 11(6):772–777

Kanal E et al (2007) ACR guidance document for safe MR practices: 2007. Am J Roentgenol 188:1447–1474. doi:10.2214/AJR.06.1616

Kim BH et al (2011) Large-scale synthesis of uniform and extremely small-sized iron oxide nanoparticles for high resolution T_1 magnetic resonance imaging contrast agents. J Am Chem Soc 133:12624–12631

Koh D-M, Cook GJR, Husband JE (2003) New horizons in oncologic imaging. N Engl J Med 348(25):2487–2488

Kwong KK, Belliveau JW, Chesler DA, Goldberg IE, Weisskoff RM, Poncelet BP, Kennedy DN, Hoppel BE, Cohen MS, Turner R, Cheng HM, Brady TJ, Rosen BR (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proc Nat Acad Sci U S A 89(12):5675–5679. doi:10.1073/pnas.89.12.5675

Lei H, Zhu X-H, Zhang X-L, Ugurbil K (2003) In vivo ^{31}P magnetic resonance spectroscopy of the human brain at 7 T: an initial experience. Magn Reson Med 49:199–205

Levitt MH (2008) Spin dynamics, 2nd edn. Wiley, New York

Liang Z-P, Lauterbur PC (2000) Principles of magnetic resonance imaging: a signal processing perspective. IEEE Press, New York

Louie A (2013) MRI biosensors: a short primer. J Mag Res Imaging 38:530–539. doi:10.1002/jmri.24298

Mariappan YK, Glaser KJ, Ehman RL (2010). Magnetic resonance elastography: a review. Clin Anat 23:497–511

Mattiello J, Basser PJ, LeBihan D (1994) Analytical expressions for the B-matrix in NMR diffusion imaging and spectroscopy. J Magn Reson Ser A 108:131–141

Moseley ME, Cohen Y, Mintorovitch J, Chileuitt L, Shimizu H, Kucharczyk J, Wendland MF, Weinstein PR (1990) Early detection of regional cerebral ischemia in cats: comparison of diffusion- and T_2-weighted MRI and spectroscopy. Magn Reson Med 14:330–346

Mugler JP, Altes TA (2013) Hyperpolarized ^{129}Xe MRI of the human lung. J Magn Reson Imag 37:313–331

Ogawa S, Lee TM, Kay AR, Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A 87:9868–9872

Robitaille P-M, Berliner LJ (2006) Ultra high field magnetic resonance imaging. Springer, New York

Sakuma H (2007) Magnetic resonance imaging for ischemic heart disease. J Magn Reson Imaging 26:3–13

Santini L, Forelo GB, Santini M (2013) Evaluating MRI-compatible pacemakers: patient data now paves the way to widespread clinical application? PACE 36:270–278. doi:10.1111/pace.12061

Schaefer DJ, Bourland JD, Nyenhuis JA (2000) Review of patient safety in time-varying gradient fields. J Magn Reson Imaging 12:20–29

Schmitt F, Stehling MK, Turner R (1998). Echo-planar imaging. Springer, Berlin

Slichter CP (1990) Principles of magnetic resonance, 3rd edn. Springer, New York

Thomas DL, Lythgoe MF, Pell GS, Calamante F, Ordidge RJ (2000) The measurement of diffusion and perfusion in biological systems using magnetic resonance imaging. Phys Med Biol 45:R97–R138

Vlaardingerbroek MT, den Boer JA (2004) Magnetic resonance imaging: Theory and practice, 3rd edn. Springer, Berlin

Wolf RL, Detre JA (2007) Clinical neuroimaging using arterial spin-labeled perfusion magnetic resonance imaging. Neurotherapeutics 4:346–359

Xia Y (2000) Magic angle effect in magnetic resonance imaging of articular cartilage: a review. Invest Radiol 35:602–621

Appendix A
Plane and Solid Angles

A.1 Plane Angles

The angle θ between two intersecting lines is shown in Fig. A.1. It is measured by drawing a circle centered on the vertex or point of intersection. The arc length s on that part of the circle contained between the lines measures the angle. In daily work, the angle is marked off in degrees.

In some cases, there are advantages to measuring the angle in *radians*. This is particularly true when trigonometric functions have to be differentiated or integrated. The angle in radians is defined by

$$\theta = \frac{s}{r}. \tag{A.1}$$

Since the circumference of a circle is $2\pi r$, the angle corresponding to a complete rotation of 360° is $2\pi r/r = 2\pi$. Other equivalences are

Degrees	Radians	
360	2π	
180	π	(A.2)
57.2958	1	
1	0.01745	

Since the angle in radians is the ratio of two distances, it is dimensionless. Nevertheless, it is sometimes useful to specify that something is measured in radians to avoid confusion.

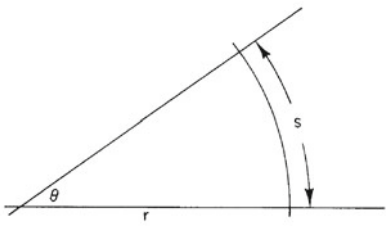


Fig. A.1 A plane angle θ is measured by the arc length s on a circle of radius r centered at the vertex of the lines defining the angle

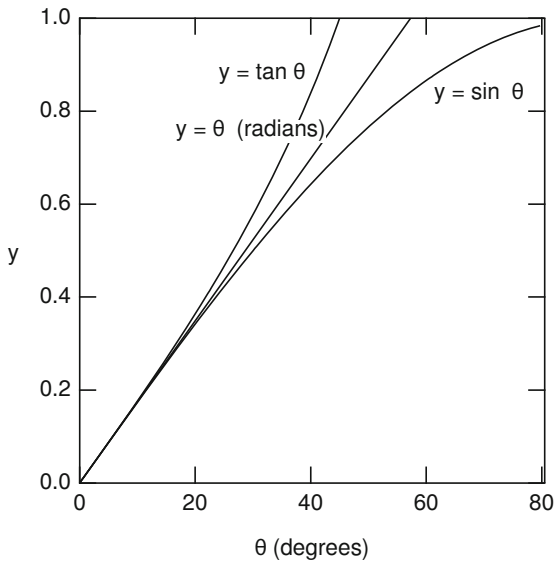


Fig. A.2 Comparison of $y = \tan\theta$, $y = \theta$ (radians), and $y = \sin\theta$

One of the advantages of radian measure can be seen in Fig. A.2. The functions $\sin\theta$, $\tan\theta$, and θ in radians are plotted vs. angle for angles less than 80°. For angles less than 15°, $y = \theta$ is a good approximation to both $y = \tan\theta$ (2.3 % error at 15°) and $y = \sin\theta$ (1.2 % error at 15°).

A.2 Solid Angles

A plane angle measures the diverging of two lines in two dimensions. *Solid angles* measure the diverging of a cone of lines in three dimensions. Figure A.3 shows a series of rays diverging from a point and forming a cone. The solid angle Ω is measured by constructing a sphere of radius r centered at the vertex and taking the ratio of the surface area S on the

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

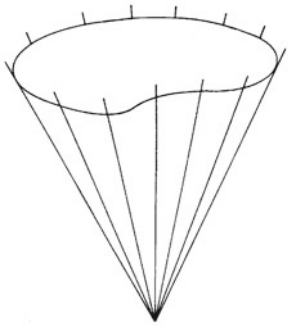


Fig. A.3 A cone of rays in three dimensions

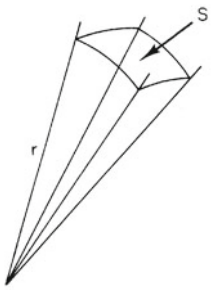


Fig. A.4 The solid angle of this cone is $\Omega = S/r^2$. S is the surface area on a sphere of radius r centered at the vertex

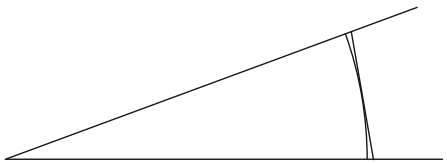


Fig. A.5 For small angles, the arc length is very nearly equal to the length of the tangent to the circle

sphere enclosed by the cone to r^2:

$$\Omega = \frac{S}{r^2}. \tag{A.3}$$

This is shown in Fig. A.4 for a cone consisting of the planes defined by adjacent pairs of the four rays shown. The unit of solid angle is the steradian (sr). A complete sphere subtends a solid angle of 4π steradians, since the surface area of a sphere is $4\pi r^2$.

When the included angle in the cone is small, the difference between the surface area of a plane tangent to the sphere and the sphere itself is small. (This is difficult to draw in three dimensions. Imagine that Fig. A.5 represents a slice through a cone; the difference in length between the circular arc and the tangent to it is small.) This approximation is often useful. A 3×5-in. card at a distance of 6 ft (72 in.) subtends a solid angle which is approximately

$$\frac{3 \times 5}{72^2} = 2.9 \times 10^{-3} \text{ sr}.$$

It is not necessary to calculate the surface area on a sphere of 72-in. radius.

Problems

Problem 1. Convert 0.1 radians to degrees. Convert 7.5 ° to radians.

Problem 2. Use the fact that $\sin\theta \approx \theta \approx \tan\theta$ to estimate the sine and tangent of 3 °. Look up the values in a table and see how accurate the approximation is.

Problem 3. What is the solid angle subtended by the pupil of the eye (radius = 3 mm) at a source of light 30 m away?

Problem 4. Figure A.2 suggests that $y = \theta$ is a better approximation to $\sin\theta$ than to $\tan\theta$ and that $y = \theta$ overestimates $\sin\theta$ and underestimates $\tan\theta$. Calculate (See Appendix D) or look up the Taylor expansions of $\sin\theta$ and $\tan\theta$ and use the first two nonzero terms in each expansion to verify this behavior.

Problem 5. What is the solid angle subtended by the "cap" of a sphere from the sphere center, where the "cap" is defined using spherical coordinates (Appendix L) as the surface of the sphere between $\theta = 0$ and $\theta = 30$ °. Hint: In spherical coordinates an element of surface area on a sphere is $dS = r^2 \sin\theta d\theta d\phi$.

Appendix B
Vectors; Displacement, Velocity, and Acceleration

B.1 Vectors and Vector Addition

A *displacement* describes how to get from one point to another. A displacement has a magnitude (how far point 2 is from point 1 in Fig. B.1) and a direction (the direction one has to go from point 1 to get to point 2). The displacement of point 2 from point 1 is labeled **A**. Displacements can be added: displacement **B** from point 2 puts an object at point 3. The displacement from point 1 to point 3 is **C** and is the sum of displacements **A** and **B**:

$$\mathbf{C} = \mathbf{A} + \mathbf{B}. \tag{B.1}$$

A displacement is a special example of a more general quantity called a *vector*. One often finds a vector defined as a quantity having a magnitude and a direction. However, the complete definition of a vector also includes the requirement that vectors add like displacements. The rule for adding two vectors is to place the tail of the second vector at the head of the first; the sum is the vector from the tail of the first to the head of the second.

A displacement is a change of position so far in such a direction. It is independent of the starting point. To know where an object is, it is necessary to specify the starting point as well as its displacement from that point.

Displacements can be added in any order. In Fig. B.2, either of the vectors **A** represents the same displacement.

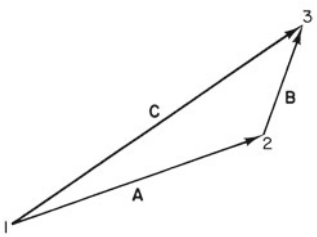


Fig. B.1 Displacement **C** is equivalent to displacement **A** followed by displacement **B**: $\mathbf{C} = \mathbf{A} + \mathbf{B}$

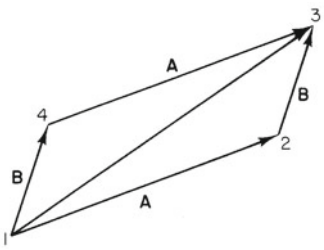


Fig. B.2 Vectors **A** and **B** can be added in either order

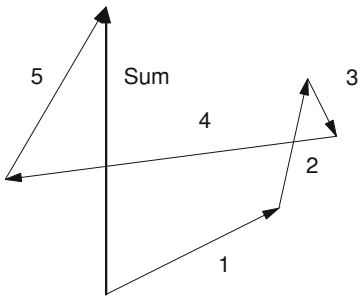


Fig. B.3 Addition of several vectors

Displacement **B** can first be made from point 1 to point 4, followed by displacement **A** from 4 to 3. The sum is still **C**:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}. \tag{B.2}$$

The sum of several vectors can be obtained by first adding two of them, then adding the third to that sum, and so forth. This is equivalent to placing the tail of each vector at the head of the previous one, as shown in Fig. B.3. The sum then goes from the tail of the first vector to the head of the last.

The negative of vector **A** is that vector which, added to **A**, yields zero:

$$\mathbf{A} + (-\mathbf{A}) = 0. \tag{B.3}$$

It has the same magnitude as **A** and points in the opposite direction.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

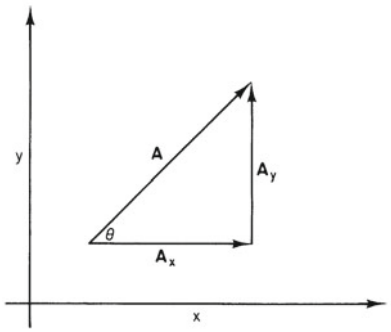


Fig. B.4 Vector **A** has components $\mathbf{A}_x$ and $\mathbf{A}_y$

Multiplying a vector **A** by a *scalar* (a number with no associated direction) multiplies the magnitude of vector **A** by that number and leaves its direction unchanged.

B.2 Components of Vectors

Consider a vector in a plane. If we set up two perpendicular axes, we can regard vector **A** as being the sum of vectors parallel to each of these axes. These vectors, $\mathbf{A}_x$ and $\mathbf{A}_y$ in Fig. B.4, are called the *components* of **A** along each axis[1]. If vector **A** makes an angle θ with the x axis and its magnitude is A, then the magnitudes of the components are

$$\begin{aligned} A_x &= A\cos\theta, \\ A_y &= A\sin\theta. \end{aligned} \tag{B.4}$$

The sum of the squares of the components is $A_x^2 + A_y^2 = A^2\cos^2\theta + A^2\sin^2\theta = A^2(\sin^2\theta + \cos^2\theta)$. Since, by Pythagoras' theorem, this must be A^2, we obtain the trigonometric identity

$$\cos^2\theta + \sin^2\theta = 1. \tag{B.5}$$

In three dimensions, $\mathbf{A} = \mathbf{A}_x + \mathbf{A}_y + \mathbf{A}_z$. The magnitudes can again be related using Pythagoras' theorem, as shown in Fig. B.5. From triangle OPQ, $A_{xy}^2 = A_x^2 + A_y^2$. From triangle OQR,

$$A^2 = A_{xy}^2 + A_z^2 = A_x^2 + A_y^2 + A_z^2. \tag{B.6}$$

In our notation, $\mathbf{A}_x$ means a vector pointing in the x direction, while A_x is the magnitude of that vector. It can become difficult to keep the distinction straight. Therefore, it is customary to write $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ to mean vectors of unit length

[1] Some texts define the component to be a scalar, the magnitude of the component defined here.

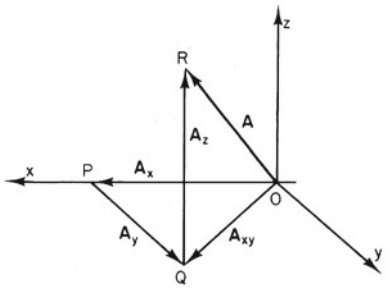


Fig. B.5 Addition of components in three dimensions

pointing in the x, y, and z directions. (In some books, the unit vectors are denoted by $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ instead of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.) With this notation, instead of $\mathbf{A}_x$, one would always write $A_x\hat{\mathbf{x}}$.

The addition of vectors is often made easier by using components. The sum $\mathbf{A} + \mathbf{B} = \mathbf{C}$ can be written as

$$\begin{aligned} &A_x\hat{\mathbf{x}} + A_y\hat{\mathbf{y}} + A_z\hat{\mathbf{z}} + B_x\hat{\mathbf{x}} + B_y\hat{\mathbf{y}} + B_z\hat{\mathbf{z}} \\ &\quad = C_x\hat{\mathbf{x}} + C_y\hat{\mathbf{y}} + C_z\hat{\mathbf{z}}. \end{aligned}$$

Like components can be grouped to give

$$\begin{aligned} &(A_x + B_x)\hat{\mathbf{x}} + (A_y + B_y)\hat{\mathbf{y}} + (A_z + B_z)\hat{\mathbf{z}} \\ &\quad = C_x\hat{\mathbf{x}} + C_y\hat{\mathbf{y}} + C_z\hat{\mathbf{z}}. \end{aligned}$$

Therefore, the magnitudes of the components can be added separately:

$$\begin{aligned} C_x &= A_x + B_x, \\ C_y &= A_y + B_y, \\ C_z &= A_z + B_z. \end{aligned} \tag{B.7}$$

B.3 Position, Velocity, and Acceleration

The *position* of an object at time t is defined by specifying its displacement from an agreed-upon origin:

$$\mathbf{R}(t) = x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}} + z(t)\hat{\mathbf{z}}.$$

The *average velocity* $\mathbf{v}_{\text{av}}(t_1, t_2)$ between times t_1 and t_2 is defined to be

$$\mathbf{v}_{\text{av}}(t_1, t_2) = \frac{\mathbf{R}(t_2) - \mathbf{R}(t_1)}{t_2 - t_1}.$$

This can be written in terms of the components as

$$\mathbf{v}_{\text{av}} = \left(\frac{x(t_2) - x(t_1)}{t_2 - t_1}\right)\hat{\mathbf{x}} + \left(\frac{y(t_2) - y(t_1)}{t_2 - t_1}\right)\hat{\mathbf{y}} + \left(\frac{z(t_2) - z(t_1)}{t_2 - t_1}\right)\hat{\mathbf{z}}.$$

The *instantaneous velocity* is

$$\mathbf{v}(t) = \frac{d\mathbf{R}}{dt} = \frac{dx}{dt}\hat{\mathbf{x}} + \frac{dy}{dt}\hat{\mathbf{y}} + \frac{dz}{dt}\hat{\mathbf{z}}$$

$$= v_x(t)\hat{\mathbf{x}} + v_y(t)\hat{\mathbf{y}} + v_z(t)\hat{\mathbf{z}}. \tag{B.8}$$

The x component of the velocity tells how rapidly the x component of the position is changing.

The *acceleration* is the rate of change of the velocity with time. The instantaneous acceleration is

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{dv_x}{dt}\hat{\mathbf{x}} + \frac{dv_y}{dt}\hat{\mathbf{y}} + \frac{dv_z}{dt}\hat{\mathbf{z}}.$$

Problems

Problem 1. At $t = 0$, the position of an object is given by $\mathbf{R} = 10\hat{\mathbf{x}} + 5\hat{\mathbf{y}}$, where R is in meters. At $t = 3$ s, the position is $\mathbf{R} = 16\hat{\mathbf{x}} - 10\hat{\mathbf{y}}$. What was the average velocity between $t = 0$ and 3 s?

Problem 2. The position of an object as a function of time is $\mathbf{R}(t) = (20 + 4t)\hat{\mathbf{x}} + (10 + 5t - 49t^2)\hat{\mathbf{y}}$. Determine the instantaneous velocity and acceleration as functions of time.

Problem 3. The electric field $\mathbf{E}$ is a vector (see Chap. 6). $\mathbf{E}_1$ has a magnitude of 30 V m^{-1} and is directed along the y axis. $\mathbf{E}_2$ has a magnitude of 15 V m^{-1} and is directed at an angle of $+30\,^\circ$ from the x axis. Calculate $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$. Express your answer in two ways: give the magnitude and direction of $\mathbf{E}$, and give E_x and E_y.

Appendix C
Properties of Exponents and Logarithms

In the expression a^m, a is called the *base* and m is called the *exponent*. Since $a^2 = a \times a$, $a^3 = a \times a \times a$, and

$$a^m = (\underbrace{a \times a \times a \times \cdots \times a}_{\text{m times}}),$$

it is easy to show that

$$a^m a^n = (\underbrace{a \times a \times a \times \cdots \times a}_{\text{m times}})(\underbrace{a \times a \times a \times \cdots \times a}_{\text{n times}}),$$

$$a^m a^n = a^{m+n}. \tag{C.1}$$

If $m > n$, the same technique can be used to show that

$$\frac{a^m}{a^n} = a^{m-n}. \tag{C.2}$$

If $m = n$, this gives

$$1 = \frac{a^m}{a^m} = a^{m-m} = a^0,$$

$$a^0 = 1. \tag{C.3}$$

The rules also work for $m < n$ and for negative exponents. For example,

$$(a^{-n})(a^n) = 1$$

so

$$a^{-n} = \frac{1}{a^n}. \tag{C.4}$$

Finally,

$$\left(a^m\right)^n = (\underbrace{a^m \times a^m \times a^m \times \cdots \times a^m}_{\text{n times}}),$$

$$(a^m)^n = a^{mn}. \tag{C.5}$$

If $y = a^x$, then by definition, x is the *logarithm* of y to the base a: $x = \log_a(y)$. If the base is 10, since $100 = 10^2$, $2 = \log_{10}(100)$. Similarly, $3 = \log_{10}(1000)$, $4 = \log_{10}(10,000)$, and so forth.

The most useful property of logarithms can be derived by letting

$$y = a^m,$$

$$z = a^n,$$

$$w = a^{m+n},$$

so that

$$m = \log_a y,$$

$$n = \log_a z,$$

$$m + n = \log_a w.$$

Then, since $a^{m+n} = a^m a^n$,

$$w = yz,$$

$$\log_a(yz) = \log_a w = \log_a y + \log_a z. \tag{C.6}$$

This result can be used to show that

$$\begin{aligned} \log(y^m) &= \log(y \times y \times y \times \cdots \times y) \\ &= \log(y) + \log(y) + \log(y) + \cdots + \log(y), \\ \log(y^m) &= m \log y. \end{aligned} \tag{C.7}$$

All logarithms in this book, unless labeled with a specific base, are to base e (see Chap. 2). These are the so-called natural logarithms. We will denote the *natural logarithm* by ln, using $\log_{10}$ when we want logarithms to the base 10.

Problems

Problem 1. What is $\log_2(8)$?

Problem 2. If $\log_{10}(2) = 0.3$, what is $\log_{10}(200)$? $\log_{10}(2 \times 10^5)$?

Problem 3. What is $\log_{10}(\sqrt{10})$?

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Appendix D
Taylor's Series

Consider the function $y(x)$ shown in Fig. D.1. The value of the function at x_1, $y_1 = y(x_1)$, is known. We wish to estimate $y(x_1 + \Delta x)$.

The simplest estimate, labeled approximation 0 in Fig. D.1, is to assume that y does not change: $y(x_1 + \Delta x) \approx y(x_1)$. A better estimate can be obtained if we assume that y changes everywhere at the same rate it does at x_1. Approximation 1 is

$$y(x_1 + \Delta x) \approx y(x_1) + \left.\frac{dy}{dx}\right|_{x_1} \Delta x.$$

The derivative is evaluated at point x_1.

An even better estimate is shown in Fig. D.2. Instead of fitting the curve by the straight line that has the proper first derivative at x_1, we fit it by a parabola that matches both the first and second derivatives. The approximation is

$$y(x_1 + \Delta x) \approx y(x_1) + \left.\frac{dy}{dx}\right|_{x_1} \Delta x + \frac{1}{2} \left.\frac{d^2y}{dx^2}\right|_{x_1} (\Delta x)^2.$$

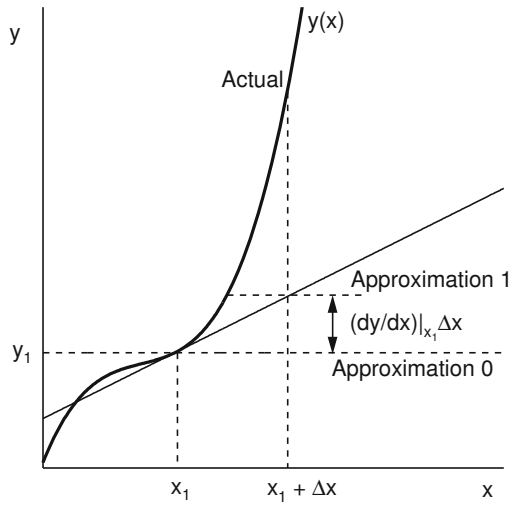


Fig. D.1 The zeroth-order and first-order approximations to $y(x)$

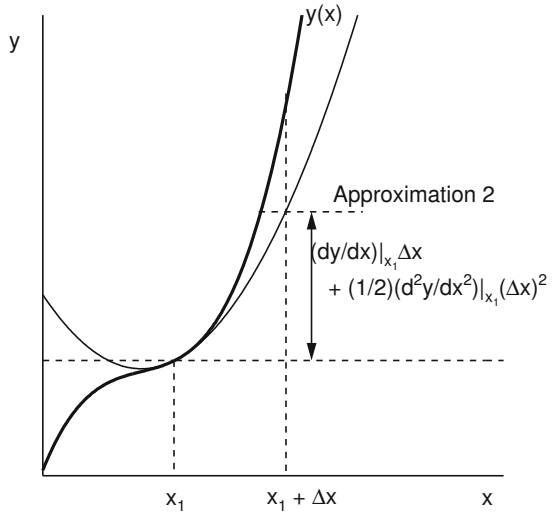


Fig. D.2 The second-order approximation fits $y(x)$ with a parabola

That this is the best approximation can be derived in the following way. Suppose the desired approximation is more general and uses terms up to $(\Delta x)^n = (x - x_1)^n$:

$$y_{\text{approx}} = A_0 + A_1(x - x_1) + A_2(x - x_1)^2 + \cdots + A_n(x - x_1)^n. \tag{D.1}$$

The constants $A_0, A_1, \ldots, A_n$ are determined by making the value of y_{approx} and its first n derivatives agree with the value of y and its first n derivatives at $x = x_1$. When $x = x_1$, all terms with $x - x_1$ in y_{approx} vanish, so that

$$y_{\text{approx}}(x_1) = A_0.$$

The first derivative of y_{approx} is

$$\frac{d(y_{\text{approx}})}{dx} = A_1 + 2A_2(x - x_1) + 3A_3(x - x_1)^2 + \cdots + nA_n(x - x_1)^{n-1}.$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Table D.1 $y = e^{2x}$ and its derivatives

Function or derivative	Value at $x_1 = 0$
$y = e^{2x}$	1
$\frac{dy}{dx} = 2e^{2x}$	2
$\frac{d^2y}{dx^2} = 4e^{2x}$	4
$\frac{d^3y}{dx^3} = 8e^{2x}$	8

Table D.2 Values of y and successive approximations

x	$y = e^{2x}$	$1 + 2x$	$1+2x+2x^2$	$1 + 2x +2x^2 + \frac{4}{3}x^3$
−2	0.0183	−3.0	5.0	−5.67
−1.5	0.0498	−2.0	2.5	−2.0
−1	0.1353	−1.0	1.0	−0.33
−0.4	0.4493	0.2000	0.5200	0.4347
−0.2	0.6703	0.6000	0.6800	0.6693
−0.1	0.8187	0.8000	0.8200	0.8187
0	1.0000	1.0000	1.0000	1.0000
0.1	1.2214	1.2000	1.2200	1.2213
0.2	1.4918	1.4000	1.4800	1.4907
0.4	2.2255	1.8000	2.1200	2.2053
1.0	7.389	3.0000	5.0000	6.33
2.0	54.60	5.0	13.0	23.67

The second derivative is

$$2A_2 + 3 \times 2A_3(x - x_1) + \cdots + n(n-1)A_n(x - x_1)^{n-2},$$

and the nth derivative is

$$n(n-1)(n-2)\cdots 2A_n = n!A_n.$$

Evaluating these at $x = x_1$ gives

$$\left.\frac{d(y_{\text{approx}})}{dx}\right|_{x_1} = A_1,$$

$$\left.\frac{d^2(y_{\text{approx}})}{dx^2}\right|_{x_1} = 2 \times 1 \times A_2,$$

$$\left.\frac{d^3(y_{\text{approx}})}{dx^3}\right|_{x_1} = 3 \times 2 \times 1 \times A_3,$$

$$\left.\frac{d^n(y_{\text{approx}})}{dx^n}\right|_{x_1} = n!A_n.$$

Combining these expressions for A_n with Eq. D.1, we get

$$y(x_1 + \Delta x) \approx y(x_1) + \sum_{n=1}^{N} \frac{1}{n!} \left.\frac{d^n y}{dx^n}\right|_{x_1} (\Delta x)^n. \qquad \text{(D.2)}$$

Tables D.1 and D.2 and Figs. D.3 and D.4 show how the Taylor's series approximation gets better over a larger and larger region about x_1 as more terms are added. The function being approximated is $y = e^{2x}$. The derivatives are given in Table D.1. The expansion is made about $x_1 = 0$.

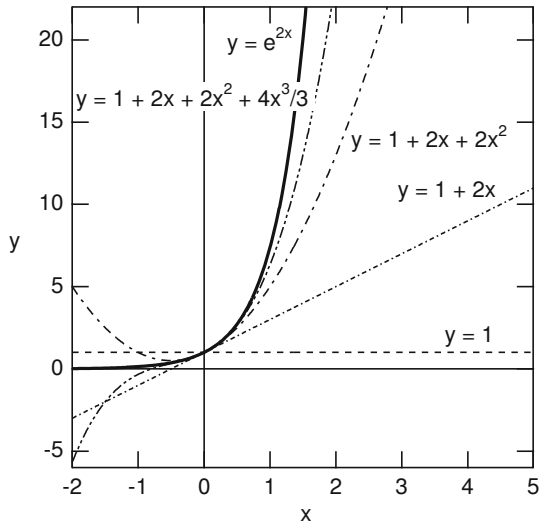


Fig. D.3 The function $y = e^{2x}$ with Taylor's series expansions about $x = 0$ of degree 0, 1, 2, and 3

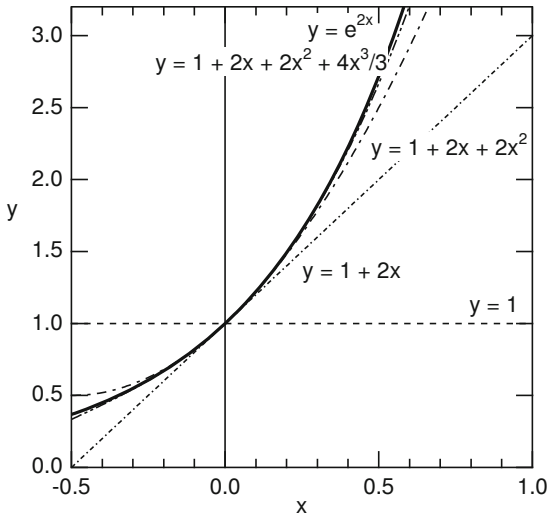


Fig. D.4 An enlargement of Fig. D.3 near $x = 0$

Finally, the Taylor's series expansion for $y = e^x$ about $x = 0$ is often useful. Since all derivatives of e^x are e^x, the value of y and each derivative at $x = 0$ is 1. The series is

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{m=0}^{\infty} \frac{x^m}{m!}. \qquad \text{(D.3)}$$

(Note that $0! = 1$ by definition.)

Problems

Problem 1. Make a Taylor's series expansion of $y = a + bx + cx^2$ about $x = 0$. Show that the expansion exactly reproduces the function.

Problem 2. Repeat the previous problem, making the expansion about $x = 1$.

Problem 3. (a) Make a Taylor's series expansion of the cosine function about $x = 0$. Remember that $d(\sin x)/dx = \cos x$ and $d(\cos x)/dx = -\sin x$.

(b) Make a Taylor's series expansion of the sine function.

Problem 4. The "sinc" function is defined as $\sin x/x$. Make a Taylor's series expansion of the sinc function about $x = 0$. Hint: first make a Taylor series expansion of $\sin x$ and then divide by x.

Problem 5. Derive a Taylor's series of $y = 1/(1 - x)$ about $x = 0$. Plot $y(x)$ vs x including approximation 0, approximation 1, and approximation 2 as in Figs. D.1 and D.2.

Problem 6. Derive a Taylor's series of $y(x) = \ln(1 + x)$ about $x = 0$. Plot $y(x)$ vs x, including approximation 0, approximation 1, and approximation 2, as in Figs. D.1 and D.2.

Appendix E
Some Integrals of Sines and Cosines

The average of a function of x with period T is defined to be

$$\langle f \rangle = \frac{1}{T} \int_{x'}^{x'+T} f(x)\, dx. \tag{E.1}$$

The sine function is plotted in Fig. E.1a. The integral over a period is zero, and its average value is zero. The area above the axis is equal to the area below the axis. Figure E.1b shows a plot of $\sin^2 x$. Since $\sin x$ varies between -1 and $+1$, $\sin^2 x$ varies between 0 (when $\sin x = 0$) and $+1$ (when $\sin x = \pm 1$). Its average value, from inspection of Fig. E.1b is $\frac{1}{2}$. If you do not want to trust the drawing to convince yourself of this, recall the identity $\sin^2\theta + \cos^2\theta = 1$. Since the sine function and the cosine function look the same, but are just shifted along the axis, their squares must also look similar. Therefore, $\sin^2\theta$ and $\cos^2\theta$ must have the same average. But if their sum is always 1, the sum of their averages must be 1. If the two averages are the same, then each must be $\frac{1}{2}$.

These same results could have been obtained analytically by using the trigonometric identity

$$\sin^2 x = \frac{1}{2} - \frac{1}{2}\cos 2x. \tag{E.2}$$

The integrals of $\sin x$ and $\cos x$ are

$$\begin{aligned} \int \sin ax\, dx &= -\frac{1}{a}\cos ax, \\ \int \cos ax\, dx &= \frac{1}{a}\sin ax. \end{aligned} \tag{E.3}$$

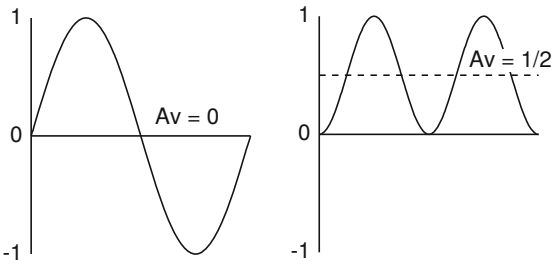


Fig. E.1 **a** Plot of $y = \sin x$. **b** Plot of $y = \sin^2 x$

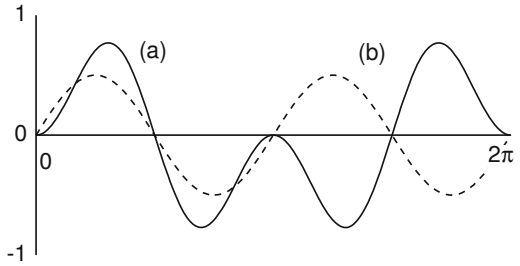


Fig. E.2 Plot of one period of **a** $y = \sin x \sin 2x$; **b** $y = \sin x \cos x$

These could be used to show that the average value of $\sin x$ or $\cos x$ is zero. Then Eq. E.2 could be used to show that the average of $\sin^2 x$ is $\frac{1}{2}$.

The integral of $\sin^2 x$ over a period is its average value times the length of the period:

$$\int_0^T \sin^2 x\, dx = \int_0^T \cos^2 x\, dx = \frac{T}{2}. \tag{E.4}$$

We will also encounter integrals like

$$\int_0^T \sin mx\, \sin nx\, dx, \quad m \neq n,$$

$$\int_0^T \cos mx\, \cos nx\, dx, \quad m \neq n, \tag{E.5}$$

$$\int_0^T \cos mx\, \sin nx\, dx, \quad m = n, \quad m \neq n.$$

All these integrals are zero. This can be shown using integral tables. Or, you can see why the integrals vanish by considering the specific examples plotted in Fig. E.2. Each integrand has equal positive and negative contributions to the total integral.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Problems

Problem 1. Plot the following functions over the range 0 to 2π as in Fig. E.2, and show by inspection that the negative and positive areas cancel, giving an integral of zero: $\cos\theta \sin 2\theta$, $\sin 2\theta \sin 3\theta$, $\cos\theta \cos 2\theta$, and $\cos 2\theta \sin 2\theta$.

Problem 2. Use the trigonometric relationship

$$\sin A \sin B = \frac{1}{2}[\cos(A - B) - \cos(A + B)],$$
$$\cos A \cos B = \frac{1}{2}[\cos(A - B) + \cos(A + B)],$$
$$\sin A \cos B = \frac{1}{2}[\sin(A - B) + \sin(A + B)],$$

to verify that all of the integrals in Eq. E.5 are zero.

Appendix F
Linear Differential Equations with Constant Coefficients

The equation

$$\frac{dy}{dt} + by = a \tag{F.1}$$

is called a *linear differential equation* because each term involves only y or its derivatives (not $y(dy/dt)$ or $(dy/dt)^2$, etc.). A more general equation of this kind has the form

$$\frac{d^N y}{dt^N} + b_{N-1}\frac{d^{N-1}y}{dt^{N-1}} + \cdots + b_1\frac{dy}{dt} + b_0 y = f(t). \tag{F.2}$$

The highest derivative is the Nth derivative, so the equation is of order N. It has been written in standard form by dividing through by any b_N that was there originally, so that the coefficient of the highest term is one. If all the bs are constants, this is a linear differential equation with constant coefficients. The right-hand side may be a function of the independent variable t, but *not* of y. If $f(t) = 0$, it is a *homogeneous* equation; if $f(t)$ is not zero, it is an *inhomogeneous* equation.

Consider first the homogeneous equation

$$\frac{d^N y}{dt^N} + b_{N-1}\frac{d^{N-1}y}{dt^{N-1}} + \cdots + b_1\frac{dy}{dt} + b_0 y = 0. \tag{F.3}$$

The exponential e^{st} (where s is a constant) has the property that $d(e^{st})/dt = se^{st}$, $d^2(e^{st})/dt^2 = s^2e^{st}$, $d^n(e^{st})/dt^n = s^n e^{st}$. The function $y = Ae^{st}$ satisfies Eq. F.3 for any value of A and certain values of s. The equation becomes

$$A\left(s^N e^{st} + b_{N-1}s^{N-1}e^{st} + \cdots + b_1 s e^{st} + b_0 e^{st}\right) = 0,$$

$$A\left(s^N + b_{N-1}s^{N-1} + \cdots + b_1 s + b_0\right)e^{st} = 0.$$

This equation is satisfied if the polynomial in parentheses is equal to zero. The equation

$$s^N + b_{N-1}s^{N-1} + \cdots + b_1 s + b_0 = 0 \tag{F.4}$$

is called the *characteristic equation* of this differential equation. It can be written in a much more compact form using summation notation:

$$\sum_{n=0}^{N} b_n s^n = 0, \tag{F.5}$$

with $b_N = 1$.

For Eq. F.1, the characteristic equation is $s + b = 0$ or $s = -b$, and a solution to the homogeneous equation is $y = Ae^{-bt}$.

If the characteristic equation is a polynomial, it can have up to N roots. For each distinct root s_n, $y = A_n e^{s_n t}$ is a solution to the differential equation. (The question of solutions when there are not N distinct roots will be taken up below.) This is still not the solution to the inhomogeneous equation. However, one can prove[1] that the most general solution to the inhomogeneous equation is the sum of the homogeneous solution,

$$y = \sum_{n=1}^{N} A_n e^{s_n t},$$

and *any* solution to the inhomogeneous equation. The values of the arbitrary constants A_n are picked to satisfy some other conditions that are imposed on the problem. If we can guess the solution to the inhomogeneous equation, that is fine. However we get it, we need only one such solution to the inhomogeneous equation. We will not prove this assertion, but we will apply it to the first- and second-order equations and see how it works.

[1] See any calculus text.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

F.1 First-Order Equation

The homogeneous equation corresponding to Eq. F.1 has solution $y = Ae^{-bt}$. There is one solution to the inhomogeneous equation that is particularly easy to write down: when y is constant, with the value $y = a/b$, the time derivative vanishes and the inhomogeneous equation is satisfied. The most general solution is therefore of the form

$$y = Ae^{-bt} + \frac{a}{b}.$$

If the initial condition is $y(0) = 0$, then A can be determined from $0 = Ae^{-b0} + a/b$. Since $e^0 = 1$, this gives $A = -a/b$. Therefore,

$$y = \frac{a}{b}\left(1 - e^{-bt}\right). \tag{F.6}$$

A physical example of this is given in Sect. 2.8.

F.2 Second-Order Equation

The second-order equation

$$\frac{d^2y}{dt^2} + b_1\frac{dy}{dt} + b_0 y = 0 \tag{F.7}$$

has a characteristic equation $s^2 + b_1 s + b_0 = 0$ with roots

$$s = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_0}}{2}. \tag{F.8}$$

This equation may have zero, one, or two solutions.

If it has two solutions s_1 and s_2, then the general solution of the homogeneous equation is $y = A_1e^{s_1t} + A_2e^{s_2t}$.

If $b_1^2 - 4b_0$ is negative, there is no solution to the equation for a real value of s. However, a solution of the form $y = Ae^{-\alpha t}\sin(\omega t + \phi)$ will satisfy the equation. This can be seen by direct substitution. Differentiating this twice shows that

$$\frac{dy}{dt} = -\alpha Ae^{-\alpha t}\sin(\omega t + \phi) + \omega Ae^{-\alpha t}\cos(\omega t + \phi),$$

$$\frac{d^2y}{dt^2} = \alpha^2 Ae^{-\alpha t}\sin(\omega t + \phi) - 2\alpha\omega Ae^{-\alpha t}\cos(\omega t + \phi) - \omega^2 Ae^{-\alpha t}\sin(\omega t + \phi).$$

If these derivatives are substituted in Eq. F.7, one gets the following results. The terms are written in two columns. One column contains the coefficients of terms with $\sin(\omega t + \phi)$, and the other column contains the coefficients of terms with $\cos(\omega t + \phi)$. The rows are labeled on the left by which term of the differential equation they came from.

Term	Coefficients	
	$\sin(\omega t + \phi)$	$\cos(\omega t + \phi)$
d^2y/dt^2	$\alpha^2 - \omega^2$	$-2\alpha\omega$
$b_1(dy/dt)$	$-b_1\alpha$	$b_1\omega$
$b_0 y$	b_0	0

The only way that the equation can be satisfied for all times is if the coefficient of the $\sin(\omega t + \phi)$ term and the coefficient of the $\cos(\omega t + \phi)$ term separately are equal to zero. This means that we have two equations that must be satisfied (call $b_0 = \omega_0^2$):

$$2\alpha\omega = b_1\omega,$$

$$\alpha^2 - \omega^2 - b_1\alpha + \omega_0^2 = 0.$$

From the first equation $2\alpha = b_1$, while from this and the second, $\alpha^2 - \omega^2 - 2\alpha^2 + \omega_0^2 = 0$, or $\omega^2 = \omega_0^2 - \alpha^2$. Thus, the solution to the equation

$$\frac{d^2y}{dt^2} + 2\alpha\frac{dy}{dt} + \omega_0^2 y = 0 \tag{F.9}$$

is

$$y = Ae^{-\alpha t}\sin(\omega t + \phi) \tag{F.10a}$$

where

$$\omega^2 = \omega_0^2 - \alpha^2, \qquad \alpha < \omega_0. \tag{F.10b}$$

Solution F.10 is a decaying exponential multiplied by a sinusoidally varying term. The initial amplitude A and the phase angle ϕ are arbitrary and are determined by other conditions in the problem. The constant α is called the *damping*. Parameter ω_0 is the undamped frequency, the frequency of oscillation when $\alpha = 0$. ω is the damped frequency.

When the damping becomes so large that $\alpha = \omega_0$, then the solution given above does not work. In that case, the solution is given by

$$y = (A + Bt)e^{-\alpha t}, \qquad \alpha = \omega_0. \tag{F.11}$$

This case is called *critical damping* and represents the case in which y returns to zero most rapidly and without multiple oscillations. The solution can be verified by substitution.

If $\alpha > \omega_0$, then the solution is the sum of the two exponentials that satisfy Eq. F.8:

$$y = Ae^{-at} + Be^{-bt}, \tag{F.12a}$$

where

$$a = \alpha + \sqrt{\alpha^2 - \omega_0^2}, \tag{F.12b}$$

$$b = \alpha - \sqrt{\alpha^2 - \omega_0^2}. \tag{F.12c}$$

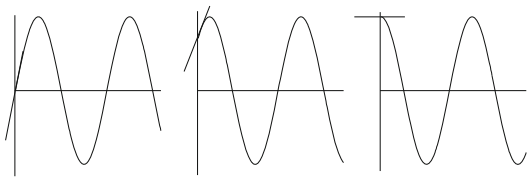


Fig. F.1 Different starting points on the sine wave give different combinations of the initial position and the initial velocity

When $\alpha = 0$, the equation is

$$\frac{d^2y}{dt^2} + \omega_0^2 y = 0. \tag{F.13}$$

The solution may be written either as

$$y = C\sin(\omega_0 t + \phi) \tag{F.14a}$$

or as

$$y = A\cos(\omega_0 t) + B\sin(\omega_0 t). \tag{F.14b}$$

The simplest physical example of this equation is a mass on a spring. There will be an equilibrium position of the mass ($y = 0$) at which there is no net force on the mass. If the mass is displaced toward either positive or negative y, a force back toward the origin results. The force is proportional to the displacement and is given by $F = -ky$. The proportionality constant k is called the *spring constant*. Newton's second law, $F = ma$, is $m(d^2y/dt^2) = -ky$ or, defining $\omega_0^2 = k/m$,

$$\frac{d^2y}{dt^2} + \omega_0^2 y = 0.$$

This (as well as the equation with $\alpha \neq 0$) is a second-order differential equation. Integrating it twice introduces two constants of integration: C and ϕ, or A and B. They are usually found from two initial conditions. For the mass on the spring, they are often the initial velocity and initial position of the mass.

The equivalence of the two solutions can be demonstrated by using Eqs. F.14a and a trigonometric identity to write $C\sin(\omega_0 t + \phi) = C[\sin\omega_0 t\cos\phi + \cos\omega_0 t\sin\phi]$. Comparison with Eq. F.14b shows that $B = C\cos\phi$, $A = C\sin\phi$. Squaring and adding these gives $C^2 = A^2 + B^2$, while dividing one by the other shows that $\tan\phi = A/B$.

Changing the initial phase angle changes the relative values of the initial position and velocity. This can be seen from the three plots of Fig. F.1, which show phase angles 0, $\pi/4$, and $\pi/2$. When $\phi = 0$, the initial position is zero, while the initial velocity has its maximum value. When $\phi = \pi/4$, the initial position has a positive value, and so does the initial velocity. When $\phi = \pi/2$, the initial position has its maximum value and the initial velocity is zero. The values of A and B are determined from the initial position and velocity. At $t = 0$, Eq. F.14b and its derivative give $y(0) = A$, $dy/dt(0) = \omega_0 B$.

The term in the differential equation equal to $2\alpha(dy/dt)$ corresponds to a drag force acting on the mass and damping the motion. Increasing the damping coefficient α increases the rate at which the oscillatory behavior decays. Figure F.2 shows plots of y and dy/dt for different values of α.

The second-order equation we have just studied is called the *harmonic oscillator* equation. Its solution is summarized in Table F.1.

Table F.1 Solutions of the harmonic oscillator equation

$$\frac{d^2y}{dt^2} + 2\alpha\frac{dy}{dt} + \omega_0^2 y = 0$$

Case	Criterion	Solution
Underdamped	$\alpha < \omega_0$	$y = Ae^{-\alpha t}\sin(\omega t + \phi)$ $\omega^2 = \omega_0^2 - \alpha^2$
Critically damped	$\alpha = \omega_0$	$y = (A + Bt)e^{-\alpha t}$
Overdamped	$\alpha > \omega_0$	$y = Ae^{-at} + Be^{-bt}$ $a = \alpha + (\alpha^2 - \omega_0^2)^{1/2}$ $b = \alpha - (\alpha^2 - \omega_0^2)^{1/2}$

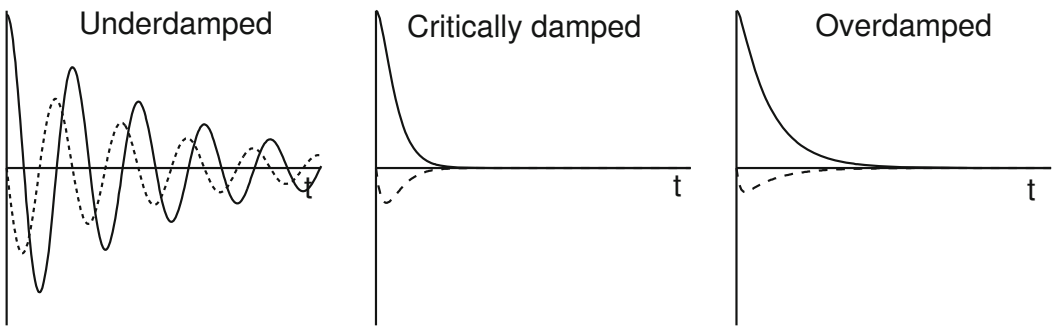


Fig. F.2 Plot of $y(t)$ (solid line) and dy/dt (dashed line) for different values of α

Problems

Problem 1. From Eq. F.14 with $\omega_0 = 10$, find A, B, C, and ϕ for the following cases:

(a) $y(0) = 5$, $(dy/dt)(0) = 0$.
(b) $y(0) = 5$, $(dy/dt)(0) = 5$.
(c) $y(0) = 0$, $(dy/dt)(0) = 50$.
(d) What values of A, B, and C would be needed to have the same ϕ as in case (b) and the same amplitude as in case (a)?

Problem 2. Verify Eq. F.11 in the critically damped case.

Problem 3. Find the general solution of the equation

$$\frac{d^2y}{dt^2} + 2\alpha\frac{dy}{dt} + \omega_0^2 y = \begin{cases} 0, & t \le 0 \\ \omega_0^2 y_0, & t \ge 0 \end{cases}$$

subject to the initial conditions $y(0) = 0, \quad (dy/dt)(0) = 0$

(a) for critical damping, $\alpha = \omega_0$,
(b) for no damping, and
(c) for overdamping, $\alpha = 2\omega_0$.

Problem 4. Show using numerical examples or physical arguments that the overdamped and critically damped solutions can cross the $y = 0$ axis at most once. Draw a plot of one such case.

Problem 5. Start with Eq. F.9. Add the function $f(t) = \sin\omega_1 t$ to the right-hand side so you have an inhomogeneous equation. Search for a solution to the inhomogeneous equation (sometimes called a *particular solution*) by guessing that $y(t) = A\sin\omega_1 t + B\cos\omega_1 t$. Put this back in the differential equation and find values of A and B that satisfy the equation. For what values of ω_1 will A and B be largest? This is an example of *resonance*: when the system is driven at its natural frequency, the response is largest.

Appendix G
The Mean and Standard Deviation

In many measurements in physics or biology there may be several possible outcomes to the measurement. Different values are obtained when the measurement is repeated. For example, the measurement might be the number of red cells in a certain small volume of blood, whether a person is right handed or left handed, the number of radioactive disintegrations of a certain sample during a 5-min interval, or the scores on a test.

Table G.1 gives the scores on an examination administered to 30 people. These results are also plotted as a histogram in Fig. G.1.

Table G.1 Quiz scores

Student No.	Score	Student No.	Score
1	80	16	71
2	68	17	83
3	90	18	88
4	72	19	75
5	65	20	69
6	81	21	50
7	85	22	81
8	93	23	94
9	76	24	73
10	86	25	79
11	80	26	82
12	88	27	78
13	81	28	84
14	72	29	74
15	67	30	70

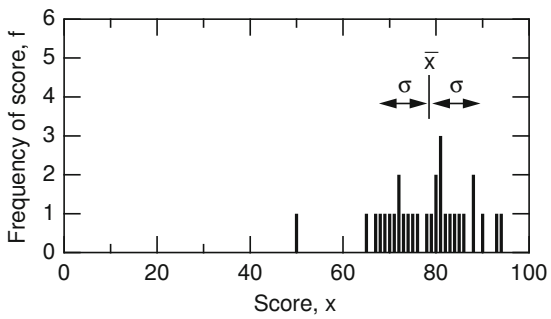


Fig. G.1 Histogram of the quiz scores in Table G.1

The table and the histogram give all the information that there is to know about the experiment unless the result depends on some variable that was not recorded, such as the age of the student or where the student was sitting during the test.

In many cases the frequency distribution gives more information than we need. It is convenient to invent some quantities that will answer the questions: Around what values do the results cluster? How wide is the distribution of results? Many different quantities have been invented for answering these questions. Some are easier to calculate or have more useful properties than others.

The *mean* or *average* shows where the distribution is centered. It is familiar to everyone: add up all the scores and divide by the number of students. For the data given above, the mean is $\overline{x} = 77.8$.

It is often convenient to group the data by the value obtained, along with the frequency of that value. The data of Table G.1 are grouped this way in Table G.2. The mean is calculated as

$$\overline{x} = \frac{1}{N}\sum_i f_i x_i = \frac{\sum_i f_i x_i}{\sum_i f_i},$$

where the sum is over the different values of the test scores that occur. For the example in Table G.2, the sums are $\sum_i f_i = 30$, $\sum_i f_i x_i = 2335$, so $\overline{x} = 2335/30 = 77.8$. If a large number of trials are made, f_i/N can be called the *probability* p_i of getting result x_i. Then

$$\overline{x} = \sum_i x_i p_i. \tag{G.1}$$

Note that $\sum p_i = 1$.

The average of some function of x is

$$\overline{g(x)} = \sum_i g(x_i) p_i. \tag{G.2}$$

For example,

$$\overline{x^2} = \sum_i (x_i)^2 p_i.$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Table G.2 Quiz scores grouped by score

Score number i	Score x_i	Frequency of score, f_i	$f_i x_i$
1	50	1	50
2	65	1	65
3	67	1	67
4	68	1	68
5	69	1	69
6	70	1	70
7	71	1	71
8	72	2	144
9	73	1	73
10	74	1	74
11	75	1	75
12	76	1	76
13	78	1	78
14	79	1	79
15	80	2	160
16	81	3	243
17	82	1	82
18	83	1	83
19	84	1	84
20	85	1	85
21	86	1	86
22	88	2	176
23	90	1	90
24	93	1	93
25	94	1	94

The width of the distribution is often characterized by the *dispersion* or *variance*:

$$\overline{(\Delta x)^2} = \overline{(x - \overline{x})^2} = \sum_i p_i (x_i - \overline{x})^2. \tag{G.3}$$

This is also sometimes called the mean square variation: the mean of the square of the variation of x from the mean. A measure of the width is the square root of this, which is called the *standard deviation* σ. The need for taking the square root is easy to see since x may have units associated with it. If x is in meters, then the variance has the units of square meters. The width of the distribution in x must be in meters.

A very useful result is

$$\overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2.$$

To prove this, note that $(x_i - \overline{x})^2 = x_i^2 - 2x_i\overline{x} + \overline{x}^2$. The variance is then

$$\overline{(\Delta x)^2} = \sum_i p_i x_i^2 - 2\sum_i x_i\, \overline{x}\, p_i + \sum_i p_i \overline{x}^2.$$

The first sum is the definition of $\overline{x^2}$. The second sum has a number $\overline{x}$ in every term. It can be factored in front of the sum, to make the second term $-2\overline{x}\sum x_i\, p_i$, which is just $-2(\overline{x})^2$. The last term is $(\overline{x})^2 \sum p_i = (\overline{x})^2$. Combining all three sums gives Eq. G.4. In summary,

$$\sigma = \sqrt{\overline{(\Delta x)^2}},$$
$$\sigma^2 = \overline{(\Delta x)^2} = \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2. \tag{G.4}$$

This equation is true as long as the p_is are accurately known. If the p_is have only been estimated from N experimental observations, the best estimate of σ^2 is $N/(N-1)$ times the value calculated from Eq. G.4.

For the data of Fig. G.1, $\sigma = 9.4$. This width is shown along with the mean at the top of the figure.

Problems

Problem 1. Calculate the variance and standard deviation for the data in Table G.2.

Problem 2. Use the data in Table G.1 to calculate the mean of the squares and then verify that $\sigma^2 = \overline{x^2} - \overline{x}^2$.

Problem 3. Another way to characterize the distribution of values is the *mode*: the most common value recorded, or the one with the highest probability. Find the mode of the data in Table G.1.

Problem 4. Still another way to describe a distribution is the median: line up all the values in order from the smallest to largest, and find the middle value. (If you have an even number of values, average the middle two.) Find the median of the data in Table G.1.

Problem 5. Find the mean and standard deviation of the following data: 14, 8, 12, 13, 7, 7, 11 and 9. You do not have enough data to know the probabilities accurately, so use the factor $N/(N-1)$ to calculate the variance.

Problem 6. Imagine that the data in Table G.1 represent 30 measurements of some quantity. The measurements contain errors, which explain why the values are not all the same. One property of the mean and standard deviation is that approximately two-third of the measurements should fall within the range $\overline{x} \pm \sigma$. (This is true for a Gaussian distribution of data and is approximately true for many others.) Check whether this is approximately true for the data in Table G.1.

Problem 7. Suppose that you make a set of measurements that have mean $\overline{x}$ and standard deviation σ. You now repeat this set of measurements N times, so that you have N mean values. These mean values will have a distribution that is narrower than the distribution of the values in a single set of measurements. The *standard deviation of the mean* is denoted by $\sigma_{mean} = \sigma/\sqrt{N}$. Calculate the standard deviation of the mean for the data in Table G.1.

Appendix H
The Binomial Probability Distribution

Consider an experiment with two mutually exclusive outcomes, which is repeated N times, with each repetition being independent of every other. One of the outcomes is labeled "success"; the other is called "failure." The experiment could be throwing a die with success being a three, flipping a coin with success being a head, or placing a particle in a box with success being that the particle is located in a subvolume v.

In a single try, call the probability of success p and the probability of failure q. Since one outcome must occur and both cannot occur at the same time,

$$p + q = 1. \tag{H.1}$$

Suppose that the experiment is repeated N times. The probability of n successes out of N tries is given by the *binomial probability distribution*, which is stated here without proof.[1] We can call the probability $P(n; N)$, since it is a function of n and depends on the parameter N. Strictly speaking, it depends on two parameters, N and p: $P(n; N, p)$. It is[2]

$$P(n; N) = P(n; N, p) = \left(\frac{N!}{n!(N-n)!}\right) p^n (1-p)^{N-n}. \tag{H.2}$$

The factor $N!/[n!(N-n)!]$ counts the number of different ways that one can get n successful outcomes; the probability of each of these ways is $p^n(1-p)^{N-n}$. In the example of three particles in Sect. 3.1, there are three ways to have one particle in the left-hand side. The particle can be either particle a or particle b or particle c. The factor gives directly

$$\left(\frac{N!}{n!(N-n)!}\right) = \frac{3!}{1!2!} = \frac{3 \times 2 \times 1}{(1)(2 \times 1)} = \frac{6}{2} = 3.$$

The remaining factor, $p^n(1-p)^{N-n}$, is the probability of taking n tries in a row and having success and taking $N-n$ tries in a row and having failure.

The binomial distribution applies if each "try" is independent of every other try. Such processes are called *Bernoulli processes* (and the binomial distribution is often called the Bernoulli distribution). In contrast, if the probability of an outcome depends on the results of the previous try, the random process is called a *Markov process*. Although such processes are important, they are more difficult to deal with and are not discussed here.

Some examples of the use of the binomial distribution are given in Chap. 3. As another example, consider the problem of performing several laboratory tests on a patient. In the 1970s it became common to use automated machines for blood-chemistry evaluations of patients; such machines automatically performed (and reported) 6, 12, 20, or more tests on one small sample of a patient's blood serum, for less cost than doing just one or two of the tests. But this meant that the physician got a large number of results—many more than would have been asked for if the tests were done one at a time. When such test batteries were first done, physicians were surprised to find that patients had many more abnormal tests than they expected. This was in part because some tests were not known to be abnormal in certain diseases, because no one had ever looked at them in that disease. But there still was a problem that some tests were abnormal in patients who appeared to be perfectly healthy.

We can understand why by considering the following idealized situation. Suppose that we do N independent tests, and suppose that in healthy people, the probability that each test is abnormal is p. (In our vocabulary, having an abnormal test is "success"!). The probability of not having the test abnormal is $q = 1 - p$. In a perfect test, p would be 0 for healthy people and would be 1 in sick people; however, very few tests are that discriminating. The definition of normal vs abnormal involves a compromise between false positives (abnormal test results in healthy people) and false negatives (normal test results in sick people). Good reviews of this

[1] A detailed proof can be found in many places. See, for example, F. Reif (1964). *Statistical Physics*. Berkeley Physics Course, Vol. 5. New York, McGraw-Hill, p. 67.

[2] $N!$ is N factorial and is $N(N-1)(N-2)\cdots 1$. By definition, $0! = 1$.

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

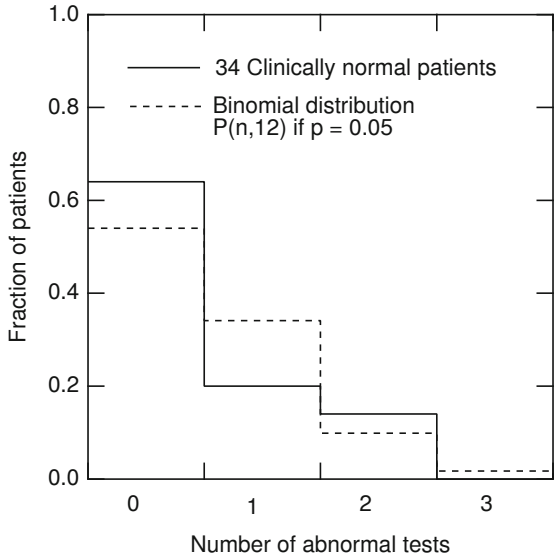


Fig. H.1 Measurement of the probability that a clinically normal patient having a battery of 12 tests done has n abnormal tests (*solid line*) and a calculation based on the binomial distribution (*dashed line*). The calculation assumes that $p = 0.05$ and that all 12 tests are independent. Several of the tests in this battery are not independent, but the general features are reproduced

problem have been written by Murphy and Abbey[3] and by Feinstein.[4] In many cases, p is about 0.05. Now suppose that p is the same for all the tests and that the tests are independent. Neither of these assumptions is very good, but they will show what the basic problem is. Then, the probability for all of the N tests to be normal in a healthy patient is given by the binomial probability distribution:

$$P(0; N, p) = \frac{N!}{0!N!} p^0 q^N = q^N.$$

If $p = 0.05$, then $q = 0.95$, and $P(0; N, p) = 0.95^N$. Typical values are $P(0; 12) = 0.54$ and $P(0; 20) = 0.36$. If the assumptions about p and independence are right, then only 36 % of healthy patients will have all their tests normal if 20 tests are done.

Figure H.1 shows a plot of the number of patients in a series who were clinically normal but who had abnormal tests. The data have the general features predicted by this simple model.

We can derive simple expressions to give the mean and standard deviation if the probability distribution is binomial. The mean value of n is defined to be

$$\overline{n} = \sum_{n=0}^{N} n P(n; N) = \sum_{n=0}^{N} \frac{N!n}{n!(N-n)!} p^n (1-p)^{N-n}.$$

The first term of each sum is for $n = 0$. Since each term is multiplied by n, the first term vanishes, and the limits of the sum can be rewritten as

$$\sum_{n=1}^{N} \frac{N!n}{n!(N-n)!} p^n (1-p)^{N-n}.$$

To evaluate this sum, we use a trick. Let $m = n - 1$ and $M = N - 1$. Then we can rewrite various parts of this expression as follows:

$$\frac{n}{n!} = \frac{1}{(n-1)!} = \frac{1}{m!},$$

$$p^n = pp^m,$$

$$N! = (N)(N-1)!,$$

$$(N-n)! = [N-1-(n-1)]! = (M-m)!.$$

The limits of summation are $n = 1$ or $m = 0$, and $n = N$ or $m = M$. With these substitutions

$$\overline{n} = Np \sum_{m=0}^{M} \frac{M!}{m!(M-m)!} p^m (1-p)^{M-m}.$$

This sum is exactly the sum of a binomial distribution over all possible values of m and is equal to one. We have the result that, for a binomial distribution,

$$\overline{n} = Np. \tag{H.3}$$

This says that the average number of successes is the total number of tries times the probability of a success on each try. If 100 particles are placed in a box and we look at half the box so that $p = \frac{1}{2}$, the average number of particles in that half is $100 \times \frac{1}{2} = 50$. If we put 500 particles in the box and look at $\frac{1}{10}$ of the box, the average number of particles in the volume is also 50. If we have 100,000 particles and $v/V = p = 1/2000$, the average number is still 50.

For the binomial distribution, the variance σ^2 can be expressed in terms of N and p using Eq. G.4. The average of n^2 is

$$\overline{n^2} = \sum_{n} P(n; N) n^2 = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} n^2 p^n (1-p)^{N-n}.$$

The trick to evaluate this is to write $n^2 = n(n-1) + n$. With this substitution we get two sums:

$$\overline{n^2} = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} n(n-1) p^n q^{N-n}$$

$$+ \sum_{n=0}^{N} \frac{N!n}{n!(N-n)!} p^n q^{N-n}.$$

[3] E. A. Murphy and H. Abbey (1967). The normal range—a common misuse. *J Chronic Dis* **20**: 79.

[4] A. R. Feinstein (1975). Clinical biostatistics XXVII. The derangements of the normal range. *Clin Pharmacol Therap* **15**: 528.

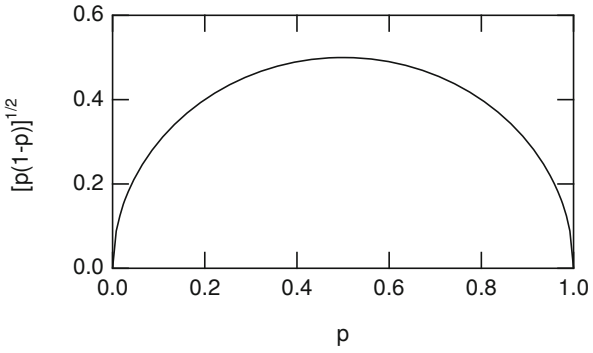


Fig. H.2 Plot of $\left[p(1-p)\right]^{1/2}$

The second sum is $\overline{n} = Np$. The first sum is rewritten by noticing that the terms for $n = 0$ and $n = 1$ both vanish. Let $m = n - 2$ and $M = N - 2$:

$$\overline{n^2} = Np + N(N-1)\sum_{m=0}^{M} \frac{M!}{m!(M-m)!} p^2 p^m q^{M-m}$$
$$= Np + N(N-1)p^2 = Np + N^2p^2 - Np^2.$$

Therefore,

$$\overline{(\Delta n)^2} = \overline{n^2} - \overline{n}^2 = Np - Np^2 = Np(1-p) = Npq.$$

For the binomial distribution, then,

$$\sigma = \sqrt{Npq} = \sqrt{\overline{n}q}.$$

The standard deviation for the binomial distribution for fixed p goes as $N^{1/2}$. For fixed N, it is proportional to $\sqrt{p(1-p)}$, which is plotted in Fig. H.2. The maximum value of σ occurs when $p = q = \frac{1}{2}$. If p is very small, the event happens rarely; if p is close to 1, the event nearly always happens. In either case, the variation is reduced. On the other hand, if N becomes large while p becomes small in such a way as to keep $\overline{n}$ fixed, then σ increases to a maximum value of $\sqrt{\overline{n}}$. This variation of σ with N and p is demonstrated in Fig. H.3. Figure H.3a–c shows how σ changes as N is held fixed and p is varied. For $N = 100$, p is 0.05, 0.5, and 0.95. Both the mean and σ change. Comparing Fig. H.3b with H.3d shows two different cases where $\overline{n} = 50$. When p is very small because N is very large in Fig. H.3d, σ is larger than in Fig. H.3b.

Problems

Problem 1. Calculate the probability of throwing $0, 1, \ldots, 9$ heads out of a total of nine throws of a coin.

Problem 2. Assume that males and females are born with equal probability. What is the probability that a couple will have four children, all of whom are girls? The couple has had three girls. What is the probability that they will have a fourth girl? Why are these probabilities so different?

Problem 3. The Mayo Clinic reported that a single stool specimen in a patient known to have an intestinal parasite yields positive results only about 90 % of the time (R. B. Thomson, R. A. Haas, and J. H. Thompson, Jr. (1984). Intestinal parasites: The necessity of examining multiple stool specimens. *Mayo Clin Proc* **59**: 641–642). What is the probability of a false negative if two specimens are examined? Three?

Problem 4. The *Minneapolis Tribune* on October 31, 1974, listed the following incidence rates for cancer in the Twin Cities greater metropolitan area, which at that time had a total population of 1.4 million. These rates are compared to those in nine other areas of the country whose total population is 15 million. Assume that each study was for 1 year. Are the differences statistically significant? Show calculations to support your answer. How would your answer differ if the study were for several years?

Type of cancer	Incidence per 100,000 per year	
	Twin Cities	Other
Colon	35.6	30.9
Lung (women)	34.2	40.0
Lung (men)	63.6	72.0
Breast (women)	81.3	73.8
Prostate (men)	69.9	60.8
Overall	313.8	300.0

Problem 5. The probability that a patient with cystic fibrosis gets a bad lung illness is 0.5 % per day. With treatment, which is time consuming and not pleasant, the daily probability is ten times less.[5] Show that the probability of not having an illness in a year is 16 % without treatment and 83 % with treatment.

[5] These numbers are from W. Warwick, MD, private communication. See also A. Gawande, The bell curve. *The New Yorker,* December 6, 2004, pp. 82–91.

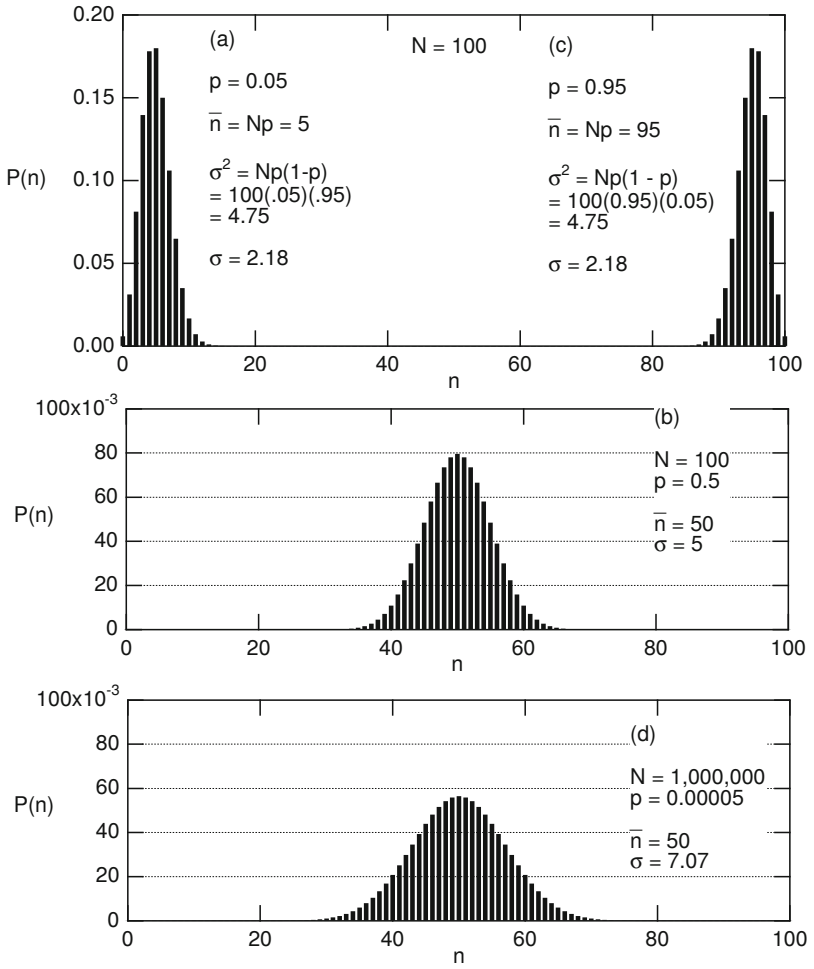


Fig. H.3 Examples of the variation of σ with N and p. (*a*), (*b*), and (*c*) show variations of σ with p when N is held fixed. The maximum value of σ occurs when $p = 0.5$. Note that (*a*) and (*c*) are both in the top panel. Comparison of (*b*) and (*d*) shows the variation of σ as p and N change together in such a way that $\overline{n}$ remains equal to 50

Appendix I
The Gaussian Probability Distribution

Appendix H considered a process that had two mutually exclusive outcomes and was repeated N times, with the probability of "success" on one try being p. If each try is independent, then the probability of n occurrences of success in N tries is

$$P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}. \qquad \text{(I.1)}$$

This probability distribution depends on two parameters N and p. We have seen two other parameters, the mean, which roughly locates the center of the distribution, and the standard deviation, which measures its width. These parameters, $\overline{n}$ and σ, are related to N and p by the equations

$$\overline{n} = Np,$$

$$\sigma^2 = Np(1-p).$$

It is possible to write the binomial distribution formula in terms of the new parameters instead of N and p. At best, however, it is cumbersome, because of the need to evaluate so many factorial functions. We will now develop an approximation that is valid when N is large and which allows the probability to be calculated more easily.

The procedure is to take the log of the probability, $y = \ln(P)$ and expand it in a Taylor's series (Appendix D) about some point. Since there is a value of n for which P has a maximum and since the logarithmic function is monotonic, y has a maximum for the same value of n. We will expand about that point; call it n_0. Then the form of y is

$$y = y(n_0) + \left.\frac{dy}{dn}\right|_{n_0} (n-n_0) + \frac{1}{2}\left.\frac{d^2y}{dn^2}\right|_{n_0} (n-n_0)^2 + \cdots.$$

Since y is a maximum at n_0, the first derivative vanishes and it is necessary to keep the quadratic term in the expansion.

To take the logarithm of Eq. I.1, we need a way to handle the factorials. There is a very useful approximation to the factorial, called *Stirling's approximation*:

$$\ln(n!) \approx n \ln n - n. \qquad \text{(I.2)}$$

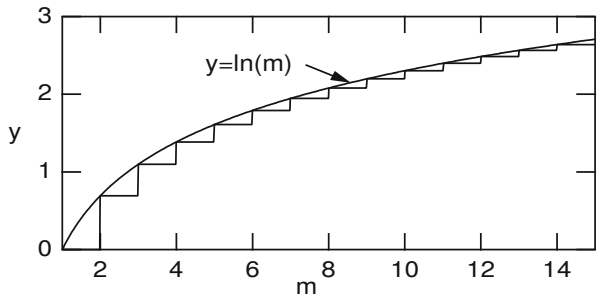


Fig. I.1 Plot of $y = \ln m$ used to derive Stirling's approximation

Table I.1 Accuracy of Stirling's approximation

n	$n!$	$\ln(n!)$	$n \ln n - n$	Error	% Error
5	120	4.7875	3.047	1.74	36
10	3.6×10^6	15.104	13.026	2.08	14
20	2.4×10^{18}	42.336	39.915	2.42	6
100	9.3×10^{157}	363.74	360.51	3.23	0.8

To derive it, write $\ln(n!)$ as

$$\ln(n!) = \ln 1 + \ln 2 + \cdots + \ln n = \sum_{m=1}^{n} \ln m.$$

The sum is the same as the total area of the rectangles in Fig. I.1, where the height of each rectangle is $\ln m$ and the width of the base is one. The area of all the rectangles is approximately the area under the smooth curve, which is a plot of $\ln m$. The area is approximately

$$\int_1^n \ln m \, dm = [m \ln m - m]_1^n = n \ln n - n + 1.$$

This completes the proof of Eq. I.2. Table I.1 shows values of $n!$ and Stirling's approximation for various values of n. The approximation is not too bad for $n > 100$.

We can now return to the task of deriving the binomial distribution. Taking logarithms of Eq. I.1, we get

$$y = \ln P = \ln(N!) - \ln(n!) - \ln(N-n)! + n \ln p + (N-n)\ln(1-p).$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

With Stirling's approximation, this becomes

$$y = N\ln N - n\ln n - N\ln(N-n) + n\ln(N-n) + n\ln p + (N-n)\ln(1-p). \quad \text{(I.3)}$$

The derivative with respect to n is

$$\frac{dy}{dn} = -\ln n + \ln(N-n) + \ln p - \ln(1-p).$$

The second derivative is

$$\frac{d^2y}{dn^2} = -\frac{1}{n} - \frac{1}{N-n}.$$

The point of expansion n_0 is found by making the first derivative vanish:

$$0 = \ln\frac{(N-n)p}{n(1-p)}.$$

Since $\ln 1 = 0$, this is equivalent to $(N-n_0)p = n_0(1-p)$ or $n_0 = Np$. The maximum of y occurs when n is equal to the mean. At $n = n_0$, the value of the second derivative is

$$\frac{d^2y}{dn^2} = -\frac{1}{Np} - \frac{1}{N(1-p)} = -\frac{1}{Np(1-p)}.$$

It is still necessary to evaluate $y_0 = y(n_0)$. If we try to do this by substitution of $n = n_0$ in Eq. I.3, we get zero. The reason is that the Stirling approximation we used is too crude for this purpose. (There are additional terms in Stirling's approximation that make it more accurate.) The easiest way to find $y(n_0)$ is to call it y_0 for now and determine it from the requirement that the probability be normalized. Therefore, we have

$$y = y_0 - \frac{1}{2Np(1-p)}(n-Np)^2$$

so that, in this approximation,

$$P(n) = e^y = e^{y_0}e^{-(n-Np)^2/[2Np(1-p)]}.$$

With $Np = \overline{n}$, $e^{y_0} = C_0$, and $Np(1-p) = \sigma^2$, this is

$$P(n) = C_0 e^{-(n-\overline{n})^2/2\sigma^2}.$$

To evaluate C_0, note that the sum of $P(n)$ for all n is the area of all the rectangles in Fig. I.2. This area is approximately the area under the smooth curve, so that

$$1 = C_0\int_{-\infty}^{\infty} e^{-(n-\overline{n})^2/2\sigma^2}\,dn.$$

It is shown in Appendix K that half of this integral is

$$\int_0^{\infty} dx\, e^{-bx^2} = \frac{1}{2}\sqrt{\frac{\pi}{b}}.$$

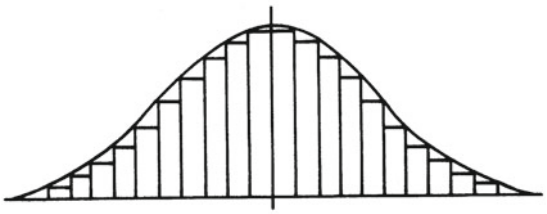


Fig. I.2 Evaluating the normalization constant

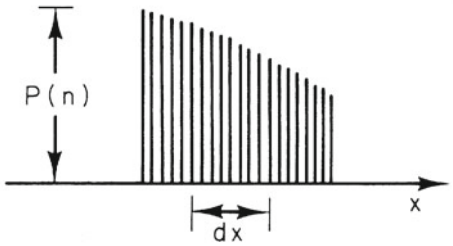


Fig. I.3 The allowed values of x are closely spaced in this case

Therefore the normalization integral is (letting $x = n - \overline{n}$)

$$\int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}dx = \sqrt{2\pi\sigma^2}.$$

The normalization constant is $C_0 = 1/\sqrt{2\pi\sigma^2}$, so that the Gaussian or normal probability distribution is

$$P(n) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(n-\overline{n})^2/2\sigma^2}. \quad \text{(I.4)}$$

It is possible, as in the case of the random-walk problem, that the measured quantity x is proportional to n with a very small proportionality constant, $x = kn$, so that the values of x appear to form a continuum. As shown in Fig. I.3, the number of different values of n (each with about the same value of $P(n)$) in the interval dx is proportional to dx. The easiest way to write down the Gaussian distribution in the continuous case is to recognize that the mean is $\overline{x} = k\overline{n}$, and the standard deviation is $\sigma_x^2 = \overline{(x-\overline{x})^2} = \overline{x^2} - \overline{x}^2 = k^2\overline{n^2} - k^2\overline{n}^2 = k^2\sigma^2$. The term $P(x)dx$ is given by $P(n)$ times the number of different values of n in dx. This number is dx/k. Therefore,

$$P(x)dx = P(n)\frac{dx}{k} = dx\frac{1}{k\sqrt{2\pi\sigma^2}}e^{-(x/k-\overline{x}/k)^2/2\sigma^2} = dx\frac{1}{\sqrt{2\pi}\sigma_x}e^{-(x-\overline{x})^2/2\sigma_x^2}. \quad \text{(I.5)}$$

To recapitulate: the binomial distribution in the case of large N can be approximated by Eq. I.4, the Gaussian or normal distribution, or Eq. I.5 for continuous variables. The original parameters N and p are replaced in these approximations by $\overline{n}$ (or $\overline{x}$) and σ.

Problems

Problem 1. An improved approximation to Stirling's formula[1] is

$$\ln n! \approx n \ln n - n + \frac{\ln(2\pi n)}{2}$$

Expand Table I.1 to include entries using this approximation.

Problem 2. Let $y = (x - \overline{x})/\sigma$. Express the Gaussian probability distribution as a function of y. Calculate the mean and standard deviation of this distribution.

[1] For more about Stirling's formula, see N. D. Mermin (1994) Stirling's formula! *Am J Phys* **52**: 362–365.

Appendix J
The Poisson Distribution

Appendix H discussed the binomial probability distribution. If an experiment is repeated N times, and has two possible outcomes, with "success" occurring with probability p in each try, the probability of getting that outcome x times in N tries is

$$P(x; N, p) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}.$$

The distribution of possible values of x is characterized by a mean value $\overline{x} = Np$ and a variance $\sigma^2 = Np(1-p)$. It is possible to specify $\overline{x}$ and σ^2 instead of N and p to define the distribution.

Appendix I showed that it is easier to work with the Gaussian or normal distribution when N is large. It is specified in terms of the parameters $\overline{x}$ and σ^2 instead of N and p:

$$P(x; \overline{x}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\overline{x})^2/2\sigma^2}.$$

The *Poisson distribution* is an approximation to the binomial distribution that is valid for large N and for small p (when N gets large and p gets small in such a way that their product remains finite). To derive it, rewrite the binomial probability in terms of $p = \overline{x}/N$:

$$\begin{aligned} P(x) &= \frac{N!}{x!(N-x)!} (\overline{x}/N)^x (1-\overline{x}/N)^{N-x} \\ &= \frac{N!}{x!(N-x)!} \frac{1}{N^x} \overline{x}^x \left(1 - \frac{\overline{x}}{N}\right)^N \left(1 - \frac{\overline{x}}{N}\right)^{-x}. \end{aligned} \tag{J.1}$$

It is necessary next to consider the behavior of some of these factors as N becomes very large. The factor $(1-\overline{x}/N)^N$ approaches $e^{-\overline{x}}$ as $N \to \infty$, by definition (see p. 34). The factor $N!/(N-x)!$ can be written out as

$$\frac{N(N-1)(N-2)\cdots 1}{(N-x)(N-x-1)\cdots 1} = N(N-1)(N-2)\cdots(N-x+1).$$

If these factors are multiplied out, the first term is N^x, followed by terms containing N^{x-1}, N^{x-2}, ..., down to N^1. But there is also a factor N^x in the *denominator* of the expression for P, which, combined with this gives

$$1 + (\text{something})N^{-1} + (\text{something})N^{-2} + \cdots.$$

As long as N is very large, all terms but the first can be neglected. With these substitutions, Eq. J.1 takes the form

$$P(x) = \frac{1}{x!} \overline{x}^x e^{-\overline{x}} \left(1 - \frac{\overline{x}}{N}\right)^{-x}. \tag{J.2}$$

The values of x for which $P(x)$ is not zero are near $\overline{x}$, which is much less than N. Therefore, the last term, which is really $\left[1/(1-p)\right]^x$, can be approximated by one, while such a term raised to the Nth power had to be approximated by e^{-x}. If this is difficult to understand, consider the following numerical example. Let N=10,000 and $p = 0.001$, so $\overline{x} = 10$. The two terms we are considering are $(1 - 10/10,000)^{10,000} = 4.517\times10^{-5}$, which is approximated by $e^{-10} = 4.54 \times 10^{-5}$, and terms like $(1 - 10/10,000)^{-10} = 1.001$, which are approximated by 1.

With these approximations, the probability is $P(x) = [(\overline{x})^x/x!]e^{-\overline{x}}$ or, calling $\overline{x} = m$,

$$P(x) = \frac{m^x}{x!} e^{-m}. \tag{J.3}$$

This is the Poisson distribution and is an approximation to the binomial distribution for large N and small p, such that the mean $\overline{x} = m = Np$ is defined (that is, it does not go to infinity or zero as N gets large and p gets small).

This probability, when summed over all values of x, should be unity. This is easily verified. Write

$$\sum_{x=0}^{\infty} P(x) = e^{-m} \sum_{x=0}^{\infty} \frac{m^x}{x!}.$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Table J.1 Comparison of the binomial, Gaussian, and Poisson distributions

Binomial	$P(x; N, p) = \dfrac{N!}{x!(N-x)!} p^x (1-p)^{N-x}$ $\bar{x} = m = Np$ $\sigma^2 = Np(1-p) = m(1-p)$
Gaussian	$P(x; m, \sigma) = \dfrac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-m)^2/2\sigma^2}$
Poisson	$P(x; m) = \dfrac{m^x}{x!} e^{-m}$ $m = Np$ $\sigma^2 = m$

But the sum on the right is the series for e^m and $e^{-m}e^m = 1$. The same trick can be used to verify that the mean is m:

$$\sum_{x=0}^{\infty} x P(x) = \sum_{x=0}^{\infty} x \frac{m^x}{x!} e^{-m} = \sum_{x=1}^{\infty} x \frac{m^x}{x!} e^{-m}.$$

The index of summation can be changed from x to $y = x-1$:

$$\sum_{x=0}^{\infty} x P(x) = \sum_{y=0}^{\infty} \frac{(y+1)}{(y+1)!} m^y m e^{-m} = m \sum_{y=0}^{\infty} \frac{m^y}{y!} e^{-m} = m.$$

One can show that the variance for the Poisson distribution is $\sigma^2 = \overline{(x-m)^2} = m$.

Table J.1 compares the binomial, Gaussian, and Poisson distributions. The principal difference between the binomial and Gaussian distributions is that the latter is valid for large N and is expressed in terms of the mean and standard deviation instead of N and p. Since the Poisson distribution is valid for very small p, there is only one parameter left, and $\sigma^2 = m$ rather than $m(1-p)$.

The Poisson distribution can be used to answer questions like the following:

1. How many red cells are there in a small square in a hemocytometer? The number of cells N is large; the probability p of each cell falling in a particular square is small. The variable x is the number of cells per square.
2. How many gas molecules are found in a small volume of gas in a large container? The number of tries is the total number of molecules. The probability that an individual molecule is in the smaller volume is $p = V/V_0$, where V is the small volume and V_0 is the volume of the entire box.
3. How many radioactive nuclei (or excited atoms) decay (or emit light) during a time dt? The probability of decay during time dt is proportional to how long dt is: $p = \lambda dt$. The number of tries is the N nuclei that might decay during that time.

The last example is worth considering in greater detail. The probability p that each nucleus decays in time dt is proportional to the length of the time interval: $p = \lambda dt$. The average number of decays if many time intervals are examined is

$$m = Np = N\lambda\, dt.$$

The probability of x decays in time dt is

$$P(x) = \frac{(N\lambda dt)^x}{x!} e^{-N\lambda dt}.$$

As $dt \to 0$, the exponential approaches one, and

$$P(x) \to \frac{(N\lambda dt)^x}{x!}.$$

The overwhelming probability for $dt \to 0$ is for there to be no decays: $P(0) \approx (N\lambda dt)^0/0! = 1$. The probability of a single decay is $P(1) = Ndt$; the probability of two decays during dt is $(N\lambda dt)^2/2$, and so forth.

If time interval t is finite, it is still possible for the Poisson criterion to be satisfied, as long as $p = \lambda t$ is small. Then the probability of no decays is

$$P(0) = e^{-m} = e^{-N\lambda t}.$$

The probability of one decay is

$$P(1) = (N\lambda t) e^{-N\lambda t}.$$

This probability increases linearly with t at first and then decreases as the exponential term begins to decay. The reason for the lowered probability of one decay is that it is now more probable for two or more decays to take place in this longer time interval. As t increases, it is more probable that there are two decays than one or none; for still longer times, even more decays become more probable. The probability that n decays occur in time t is $P(n)$. Figure J.1 shows plots of $P(0)$, $P(1)$, $P(2)$, and $P(3)$, vs $m = N\lambda t$.

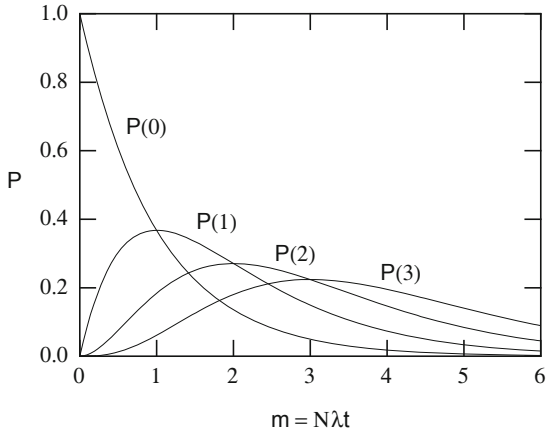


Fig. J.1 Plot of $P(0)$ through $P(3)$ vs. $N\lambda t$

Problems

Problem 1. In the USA 400,000 people were killed or injured one year in automobile accidents. The total population was 200,000,000. If the probability of being killed or injured is independent of time, what is the probability that you will escape unharmed from 70 years of driving?

Problem 2. Large proteins consist of a number of smaller subunits that are stuck together. Suppose that an error is made in inserting an amino acid once in every 10^5 tries; $p = 10^{-5}$. If a chain has length 1000, what is the probability of making a chain with no mistakes? If the chain length is 10^5?

Problem 3. The muscle end plate has an electrical response whenever the nerve connected to it is stimulated. I. A. Boyd and A. R. Martin (The end plate potential in mammalian muscle. *J Physiol* **132**: 74–91 (1956)) found that the electrical response could be interpreted as resulting from the release of packets of acetylcholine by the nerve. In terms of this model, they obtained the following data:

Number of packets reaching the end plate	Number of times observed
0	18
1	44
2	55
3	36
4	25
5	12
6	5
7	2
8	1
9	0

Analyze these data in terms of a Poisson distribution.

Appendix K
Integrals Involving e^{-ax^2}

Integrals involving e^{-ax^2} appear in the Gaussian distribution. The integral

$$I = \int_{-\infty}^{\infty} e^{-ax^2} dx$$

can also be written with y as the dummy variable:

$$I = \int_{-\infty}^{\infty} e^{-ay^2} dy.$$

These can be multiplied together to get

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx dy\, e^{-ax^2} e^{-ay^2}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx dy\, e^{-a(x^2+y^2)}.$$

A point in the xy plane can also be specified by the polar coordinates r and θ (Fig. K.1). The element of area $dxdy$ is replaced by the element $rdrd\theta$:

$$I^2 = \int_0^{2\pi} d\theta \int_0^{\infty} r\, dr\, e^{-ar^2} = 2\pi \int_0^{\infty} r\, dr\, e^{-ar^2}.$$

To continue, make the substitution $u = ar^2$, so that $du = 2ardr$. Then

$$I^2 = 2\pi \int_0^{\infty} \frac{1}{2a} e^{-u}\, du = \frac{\pi}{a} \left[-e^{-u}\right]_0^{\infty} = \frac{\pi}{a}.$$

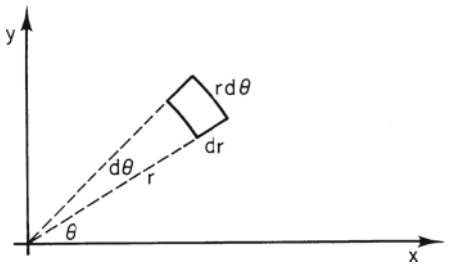


Fig. K.1 An element of area in polar coordinates

The desired integral is, therefore,

$$I = \int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}. \tag{K.1}$$

This integral is one of a sequence of integrals of the general form

$$I_n = \int_0^{\infty} x^n e^{-ax^2} dx.$$

From Eq. K.1, we see that

$$I_0 = \frac{I}{2} = \frac{1}{2}\sqrt{\frac{\pi}{a}}. \tag{K.2}$$

The next integral in the sequence can be integrated directly with the substitution $u = ax^2$:

$$I_1 = \int_0^{\infty} xe^{-ax^2} dx = \frac{1}{2a} \int_0^{\infty} e^{-u} du = \frac{1}{2a}. \tag{K.3}$$

A value for I_2 can be obtained by integrating by parts:

$$I_2 = \int_0^{\infty} x^2 e^{-ax^2} dx.$$

Let $u = x$ and $dv = xe^{-ax^2} dx = -(1/2a)d(e^{-ax^2})$. Since $\int udv = uv - \int vdu$,

$$\int_0^{\infty} x^3 e^{-ax^2} dx = -\frac{xe^{-ax^2}}{2a} + \frac{1}{2a} \int e^{-ax^2} dx.$$

This expression is evaluated at the limits 0 and ∞. The term xe^{-ax^2} vanishes at both limits. The second term is $I_0/2a$. Therefore,

$$I_2 = \frac{1}{2 \times 2a}\sqrt{\frac{\pi}{a}}.$$

This process can be repeated to get other integrals in the sequence. The even members build on I_0; the odd members

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

build on I_1. General expressions can be written. Note that $2n$ and $2n+1$ are used below to assure even and odd exponents:

$$\int_0^\infty x^{2n} e^{-ax^2} dx = \frac{1 \times 3 \times 5 \times (2n-1)}{2^{n+1} a^n} \sqrt{\frac{\pi}{a}}, \qquad \text{(K.4)}$$

$$\int_0^\infty x^{2n+1} e^{-ax^2} dx = \frac{n!}{2a^{n+1}}, \qquad (a > 0). \qquad \text{(K.5)}$$

The integrals in Appendix I are of the form

$$\int_{-\infty}^\infty e^{-x^2/2\sigma^2} dx.$$

This integral is $2I_0$ with $a = 1/(2\sigma^2)$. Therefore, the integral is $\sqrt{2\pi\sigma^2}$.

Integrals of the form

$$J = \int_0^\infty x^n e^{-ax} dx,$$

can be transformed to the forms above with the substitution $y = x^{1/2}$, $x = y^2$, $dx = 2y\,dy$. Then

$$J = \int_0^\infty y^{2n} e^{-ay^2} 2y\,dy = 2\int_0^\infty y^{2n+1} e^{-ay^2} dy.$$

Therefore,

$$\int_0^\infty x^n e^{-ax} dx = \frac{n!}{a^{n+1}} = \frac{\Gamma(n+1)}{a^{n+1}}. \qquad \text{(K.6)}$$

The gamma function $\Gamma(n) = (n-1)!$ if n is an integer. Unlike $n!$, it is also defined for noninteger values. Although we have not shown it, Eq. K.6 is correct for noninteger values of n as well, as long as $a > 0$ and $n > -1$.

Problems

Problem 1. Use integration by parts to evaluate

$$I_3 = \int_0^\infty x^3 e^{-ax^2} dx.$$

Compare this result with Eq. K.5.

Problem 2. Show that $\int_{-\infty}^\infty x e^{-ax^2} dx = 0$. Note the lower limit is $-\infty$, not 0. There is a hard way and an easy way to show this. Try to find the easy way.

Appendix L
Spherical and Cylindrical Coordinates

It is possible to use coordinate systems other than the rectangular (or Cartesian) (x, y, z): In spherical coordinates (Fig. L.1), the coordinates are radius r and angles θ and ϕ:

$$\begin{aligned} x &= r \sin\theta \cos\phi, \\ y &= r \sin\theta \sin\phi, \\ z &= r \cos\theta. \end{aligned} \tag{L.1}$$

In Cartesian coordinates a volume element is defined by surfaces on which x is constant (at x and $x + dx$), y is constant, and z is constant. The volume element is a cube with edges dx, dy, and dz. In spherical coordinates, the cube has faces defined by surfaces of constant r, constant θ, and constant ϕ (Fig. L.2). A volume element is then

$$dV = (dr)(r\, d\theta)(r \sin\theta\, d\phi) = r^2 \sin\theta\, d\theta\, d\phi\, dr. \tag{L.2}$$

To calculate the divergence of vector $\mathbf{J}$, resolve it into components $\mathbf{J}_r$, $\mathbf{J}_\theta$, and $\mathbf{J}_\phi$, as shown in Fig. L.2. These components are parallel to the vectors defined by small displacements in the r, θ, and ϕ directions. A detailed calculation[1] shows that the divergence is

$$\operatorname{div} \mathbf{J} = \nabla \cdot \mathbf{J} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 J_r) + \frac{1}{r \sin\theta}\frac{\partial}{\partial \theta}(\sin\theta\, J_\theta) + \frac{1}{r \sin\theta}\frac{\partial}{\partial \phi}(J_\phi). \tag{L.3}$$

The gradient, which appears in the three-dimensional diffusion equation (Fick's first law), can also be written in spherical coordinates. The components are

$$\begin{aligned} (\nabla C)_r &= \frac{\partial C}{\partial r}, \\ (\nabla C)_\theta &= \frac{1}{r}\frac{\partial C}{\partial \theta}, \\ (\nabla C)_\phi &= \frac{1}{r \sin\theta}\frac{\partial C}{\partial \phi}. \end{aligned} \tag{L.4}$$

Figure L.2 also shows that the element of area on the surface of the sphere is $(r\, d\theta)(r \sin\theta\, d\phi) = r^2 \sin\theta\, d\theta\, d\phi$. The element of solid angle is therefore

$$d\Omega = \sin\theta\, d\theta\, d\phi.$$

This is easily integrated to show that the surface area of a sphere is $4\pi r^2$ or that the solid angle is 4π sr.

$$\begin{aligned} S &= r^2 \int_0^\pi \sin\theta\, d\theta \int_0^{2\pi} d\phi = 2\pi r^2 \int_0^\pi \sin\theta\, d\theta \\ &= 2\pi r^2 \left[-\cos\theta\right]_0^\pi = 4\pi r^2. \end{aligned}$$

Similar results can be written down in cylindrical coordinates (r, ϕ, z), shown in Fig. L.3.

Table L.1 shows the divergence, gradient, and curl in rectangular, cylindrical, and spherical coordinates, along with the Laplacian operator ∇^2.

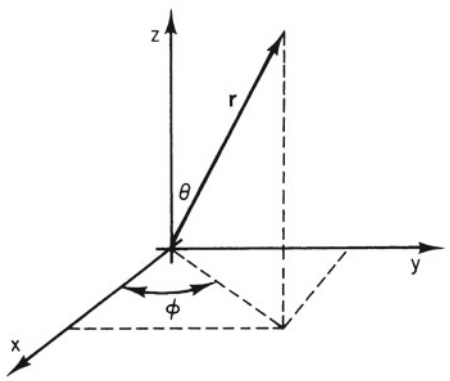


Fig. L.1 Spherical coordinates.

[1] H. M. Schey (2005). *Div, Grad, Curl, and All That*. 4th. ed. New York, Norton.

DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

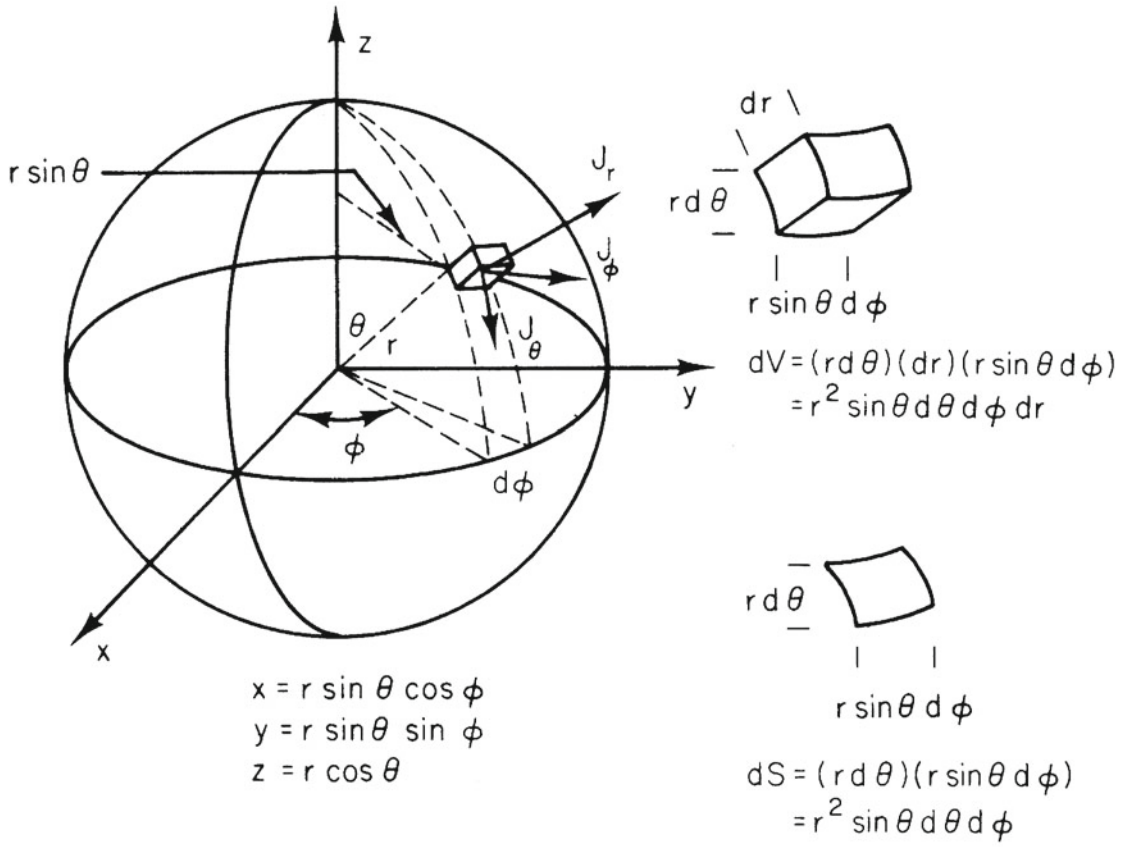


Fig. L.2 The volume element and element of surface area in spherical coordinates

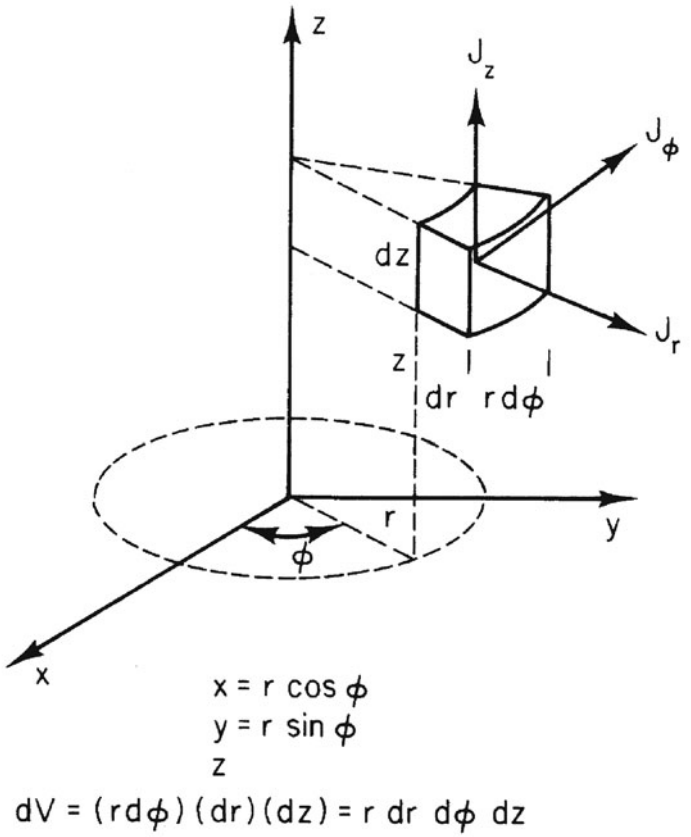


Fig. L.3 A cylindrical coordinate system

Table L.1 The vector operators in rectangular, cylindrical, and spherical coordinates

Rectangular x, y, z	Cylindrical r, ϕ, z	Spherical r, θ, ϕ
Gradient		
$(\nabla C)_x = \dfrac{\partial C}{\partial x}$	$(\nabla C)_r = \dfrac{\partial C}{\partial r}$	$(\nabla C)_r = \dfrac{\partial C}{\partial r}$
$(\nabla C)_y = \dfrac{\partial C}{\partial y}$	$(\nabla C)_\phi = \dfrac{1}{r}\dfrac{\partial C}{\partial \phi}$	$(\nabla C)_\theta = \dfrac{1}{r}\dfrac{\partial C}{\partial \theta}$
$(\nabla C)_z = \dfrac{\partial C}{\partial z}$	$(\nabla C)_z = \dfrac{\partial C}{\partial z}$	$(\nabla C)_\phi = \dfrac{1}{r\sin\theta}\dfrac{\partial C}{\partial \phi}$
Laplacian		
$\nabla^2 C = \dfrac{\partial^2 C}{\partial x^2} + \dfrac{\partial^2 C}{\partial y^2} + \dfrac{\partial^2 C}{\partial z^2}$	$\nabla^2 C = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial C}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2 C}{\partial \phi^2} + \dfrac{\partial^2 C}{\partial z^2}$	$\nabla^2 C = \dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial C}{\partial r}\right) + \dfrac{1}{r^2\sin\theta}\dfrac{\partial}{\partial \theta}\left(\sin\theta\dfrac{\partial C}{\partial \theta}\right) + \dfrac{1}{r^2\sin^2\theta}\dfrac{\partial^2 C}{\partial \phi^2}$
Divergence		
$\nabla\cdot\mathbf{j} = \dfrac{\partial j_x}{\partial x} + \dfrac{\partial j_y}{\partial y} + \dfrac{\partial j_z}{\partial z}$	$\nabla\cdot\mathbf{j} = \dfrac{1}{r}\dfrac{\partial(r j_r)}{\partial r} + \dfrac{1}{r}\dfrac{\partial j_\phi}{\partial \phi} + \dfrac{\partial j_z}{\partial z}$	$\nabla\cdot\mathbf{j} = \dfrac{1}{r^2}\dfrac{\partial(r^2 j_r)}{\partial r} + \dfrac{1}{r\sin\theta}\dfrac{\partial(\sin\theta\, j_\theta)}{\partial \theta} + \dfrac{1}{r\sin\theta}\dfrac{\partial j_\phi}{\partial \phi}$
Curl		
$(\nabla\times\mathbf{j})_x = \dfrac{\partial j_z}{\partial y} - \dfrac{\partial j_y}{\partial z}$	$(\nabla\times\mathbf{j})_r = \dfrac{1}{r}\dfrac{\partial j_z}{\partial \phi} - \dfrac{\partial j_\phi}{\partial z}$	$(\nabla\times\mathbf{j})_r = \dfrac{1}{r\sin\theta}\times\left[\dfrac{\partial(\sin\theta\, j_\phi)}{\partial \theta} - \dfrac{\partial(j_\theta)}{\partial \phi}\right]$
$(\nabla\times\mathbf{j})_y = \dfrac{\partial j_x}{\partial z} - \dfrac{\partial j_z}{\partial x}$	$(\nabla\times\mathbf{j})_\phi = \dfrac{\partial j_r}{\partial z} - \dfrac{\partial j_z}{\partial r}$	$(\nabla\times\mathbf{j})_\theta = \dfrac{1}{r\sin\theta}\times\left[\dfrac{\partial j_r}{\partial \phi} - \dfrac{\sin\theta\,\partial(r j_\phi)}{\partial r}\right]$
$(\nabla\times\mathbf{j})_z = \dfrac{\partial j_y}{\partial x} - \dfrac{\partial j_x}{\partial y}$	$(\nabla\times\mathbf{j})_z = \dfrac{1}{r}\dfrac{\partial(r j_\phi)}{\partial r} - \dfrac{1}{r}\dfrac{\partial j_r}{\partial \phi}$	$(\nabla\times\mathbf{j})_{\partial\phi} = \dfrac{1}{r}\left[\dfrac{\partial(r j_\theta)}{\partial r} - \dfrac{\partial j_r}{\partial \theta}\right]$

Appendix M
Joint Probability Distributions

In both physics and medicine, the question often arises of what is the probability that x has a certain value x_i while y has the value y_j. This is called a *joint probability*. Joint probability can be extended to several variables. This appendix derives some properties of joint probabilities for discrete and continuous variables.

M.1 Discrete Variables

Consider two variables. For simplicity assume that each can have only two values. The first might be the patient's health with values *healthy* and *sick*; the other might be the results of some laboratory test, with results *normal* and *abnormal*. Table M.1 shows the values of the two variables for a sample of 100 patients. The joint probability that a patient is healthy *and* has a normal test result is $P(x = 0, y = 0) = 0.6$; the probability that a patient is sick and has an abnormal test is $P(1, 1) = 0.15$. The probability of a false positive test is $P(0, 1) = 0.20$; the probability of a false negative is $P(1, 0) = 0.05$.

The probability that a patient is healthy regardless of the test result is obtained by a summing over all possible test outcomes: $P(x = 0) = P(0, 0) + P(0, 1) = 0.6 + 0.2 = 0.8$.

In a more general case, we can call the joint probability $P(x, y)$, the probability that x has a certain value

Table M.1 The results of measurements on 100 patients showing whether they are healthy or sick and whether a laboratory test was normal or abnormal

	Healthy ($x = 0$)	Sick ($x = 1$)
Normal test ($y = 0$)	60	5
Abnormal test ($y = 1$)	20	15

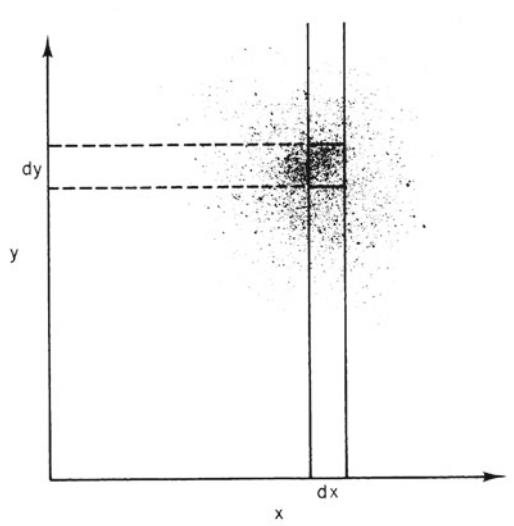


Fig. M.1 The results of measuring two continuous variables simultaneously. Each experimental result is shown as a point

independent of y, $P_x(x)$, and so forth. Then

$$\begin{aligned} P_x(x) &= \textstyle\sum_y P(x, y) \\ P_y(y) &= \textstyle\sum_x P(x, y). \end{aligned} \tag{M.1}$$

Since any measurement must give some value for x and y, we can write

$$\begin{aligned} 1 &= \textstyle\sum_x P_x(x) = \sum_x \sum_y P(x, y), \\ 1 &= \textstyle\sum_y P_y(y) = \sum_y \sum_x P(x, y). \end{aligned} \tag{M.2}$$

M.2 Continuous Variables

When a variable can take on a continuous range of values, it is quite unlikely that the variable will have *precisely* the

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

value x. Instead, there is a probability that it is in the interval (x, dx), meaning that it is between x and $x + dx$. For small values of dx, the probability that the value is in the interval is proportional to the width of the interval. We will call it $p_x(x)dx$. The extension to joint probability in two dimensions is $p(x, y)dxdy$. This is the probability that x is in the interval (x, dx) and y is in the interval (y, dy). Figure M.1 shows each outcome of a joint measurement as a dot in the xy plane. The probability that x is in (x, dx) regardless of the value of y is

$$p_x(x)dx = \left(\int p(x, y)dy \right) dx. \tag{M.3}$$

It is proportional to the total number of dots in the vertical strip in Fig. M.1. Normalization requires that

$$1 = \int p_x(x)dx = \int dx \int dy\, p(x, y). \tag{M.4}$$

The first strip could be taken horizontally:

$$1 = \int p_y(y)dy = \int dy \int dx\, p(x, y).$$

Figure M.2 shows a perspective drawing of $p(x, y)$. The volume of the shaded column is $p(x, y)dxdy$. The volume of the slice is $p_x(x)dx$. The entire volume under the surface is equal to 1.

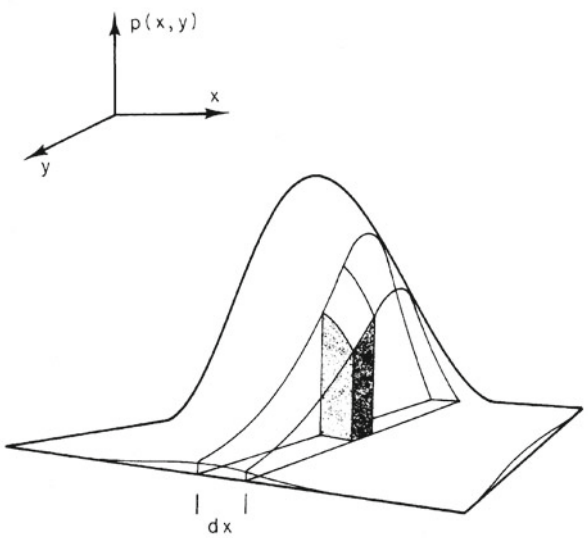


Fig. M.2 Perspective drawing of $p(x, y)$

Appendix N
Partial Derivatives

When a function depends on several variables, we may want to know how the value of the function changes when one or more of the variables is changed. For example, the volume of a cylinder is

$$V = \pi r^2 h.$$

How does V change when r is changed while the height of the cylinder is kept fixed?

$$V(r + \Delta r) = \pi(r + \Delta r)^2 h = \pi(r^2 + 2r\Delta r + \Delta r^2)h.$$

Subtracting the original volume, we have

$$\Delta V = \pi(2r\Delta r + \Delta r^2)h.$$

In the limit of small Δr, this is

$$dV = 2\pi h r dr.$$

This is the same answer we would have gotten if h had been regarded as a constant. The *partial derivative* of V with respect to r is defined to be

$$\left(\frac{\partial V}{\partial r}\right)_h = \lim_{\Delta r \to 0}\left(\frac{V(r + \Delta r, h) - V(r, h)}{\Delta r}\right) = 2\pi r h.$$

The subscript h in the partial derivative symbol means that h is held fixed during the differentiation. Sometimes it is omitted; when it is not there, it is understood that all variables except the one following the ∂ are held fixed.

If the cylinder radius is held fixed while the height is varied, we can write

$$\Delta V = V(r, h + \Delta h) - V(r, h) = \pi r^2 \Delta h.$$

The partial derivative is

$$\left(\frac{\partial V}{\partial h}\right)_r = \lim_{\Delta h \to 0}\left(\frac{V(r, h + \Delta h) - V(r, h)}{\Delta h}\right) = \pi r^2.$$

Suppose now that we allow small changes in both r and h. The difference in volume is

$$\Delta V = V(r + \Delta r, h + \Delta h) - V(r, h).$$

We can add and subtract the term $V(r, h + \Delta h)$:

$$\begin{aligned}\Delta V &= V(r + \Delta r, h + \Delta h) - V(r, h + \Delta h) \\ &\quad + V(r, h + \Delta h) - V(r, h) \\ &= \frac{V(r + \Delta r, h + \Delta h) - V(r, h + \Delta h)}{\Delta r}\Delta r \\ &\quad + \frac{V(r, h + \Delta h) - V(r, h)}{\Delta h}\Delta h.\end{aligned}$$

In the limit as Δr and $\Delta h \to 0$, the first term is

$$\left(\frac{\partial V}{\partial r}\right)_h \Delta r,$$

evaluated at $h + \Delta h$. If the derivatives are continuous at (r, h), the derivative evaluated at $(r, h + \Delta h)$ is negligibly different from the derivative evaluated at (r, h). Therefore, we can write

$$dV = \left(\frac{\partial V}{\partial r}\right)_h dr + \left(\frac{\partial V}{\partial h}\right)_r dh.$$

This result is true for several variables. For a function $w(x, y, z)$,

$$dw = \left(\frac{\partial w}{\partial x}\right)_{y,z} dx + \left(\frac{\partial w}{\partial y}\right)_{x,z} dy + \left(\frac{\partial w}{\partial z}\right)_{x,y} dz. \quad \text{(N.1)}$$

The derivatives are evaluated as though the variables being held fixed were ordinary constants. If $w = 3x^2 y z^4$,

$$\left(\frac{\partial w}{\partial x}\right)_{y,z} = 6xyz^4,$$

$$\left(\frac{\partial w}{\partial y}\right)_{x,z} = 3x^2 z^4,$$

$$\left(\frac{\partial w}{\partial z}\right)_{x,y} = 12x^2 y z^3.$$

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

It is also possible to take higher derivatives, such as $\partial^2 w/\partial x^2$ or $\partial^2 w/\partial x \partial y$. One important result is that the order of differentiation is unimportant, if the function, its first derivatives, and the derivatives in question are continuous at the point where they are evaluated. Without filling in all the details of a rigorous proof, we will simply note that

$$f = \frac{\partial w}{\partial x} = \lim_{\Delta x \to 0} \left(\frac{w(x + \Delta x, y) - w(x, y)}{\Delta x} \right)$$

$$g = \frac{\partial w}{\partial y} = \lim_{\Delta y \to 0} \left(\frac{w(x, y + \Delta y) - w(x, y)}{\Delta y} \right).$$

The mixed partials are

$$\begin{aligned} \frac{\partial^2 w}{\partial y \partial x} &= \frac{\partial f}{\partial y} = \lim_{\Delta y \to 0} \left(\frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \right) \\ &= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \left(\frac{w(x + \Delta x, y + \Delta y) - w(x, y + \Delta y) - w(x + \Delta x, y) + w(x, y)}{\Delta x \, \Delta y} \right) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 w}{\partial x \partial y} &= \frac{\partial g}{\partial x} \\ &= \lim_{\substack{\Delta y \to 0 \\ \Delta x \to 0}} \left(\frac{w(x + \Delta x, y + \Delta y) - w(x + \Delta x, y) - w(x, y + \Delta y) + w(x, y)}{\Delta x \, \Delta y} \right). \end{aligned}$$

The right side of each of these equations is the same, except for the order of the terms. Thus,

$$\frac{\partial}{\partial x} \frac{\partial w}{\partial y} = \frac{\partial}{\partial y} \frac{\partial w}{\partial x}.$$

Problems

Problem 1. If $w = 12x^3 y + z$, find the three partial derivatives $\partial w/\partial x$, $\partial w/\partial y$, and $\partial w/\partial z$.

Problem 2. If $V = xyz$ and $x = 5$, $y = 6$, $z = 2$, find dV when $dx = 0.01$, $dy = 0.02$, and $dz = 0.03$. Make a geometrical interpretation of each term.

Appendix O
Some Fundamental Constants and Conversion Factors

The values of the fundamental constants are from the 2010 least-squares adjustment, available at http://www.nist.gov/pml/data/index.cfm

Symbol	Constant	Value	SI units
c	Velocity of light in vacuum	2.997925×10^{8}	m s^{-1}
e	Elementary charge	1.602177×10^{-19}	C
F	Faraday constant	9.64853×10^{4}	C mol^{-1}
g	Standard acceleration of free fall	9.80665	m s^{-2}
h	Planck's constant	6.626070×10^{-34}	J s
$\hbar$	Planck's constant (reduced)	1.054572×10^{-34}	J s
		6.582119×10^{-16}	eV s
k_B	Boltzmann's constant	1.380649×10^{-23}	J K^{-1}
		8.617343×10^{-5}	eV K^{-1}
m_e	Electron rest mass	9.109383×10^{-31}	kg
$m_e c^2$	Electron rest energy	8.187105×10^{-14}	J
		5.10999×10^{5}	eV
m_p	Proton rest mass	1.672622×10^{-27}	kg
N_A	Avogadro's number	6.022141×10^{23}	mol^{-1}
r_e	Classical electron radius	2.817940×10^{-15}	m
R	Gas constant	8.31446	J mol^{-1} K^{-1}
u	Mass unit (^{12}C standard)	1.660539×10^{-27}	kg
uc^2	Mass unit (energy units)	9.31494×10^{8}	eV
ϵ_0	Electrical permittivity of free space	8.85419×10^{-12}	C^2 N^{-1} m^{-2}
$1/4\pi\epsilon_0$		8.98755×10^{9}	N m^2 C^{-2}
σ_{SB}	Stefan Boltzmann constant	5.67037×10^{-8}	W m^{-2} K^{-4}
λ_C	Compton wavelength of electron	2.42631×10^{-12}	m
μ_B	Bohr magneton	9.274010×10^{-24}	J T^{-1}
μ_0	Magnetic permeability	$4\pi \times 10^{-7}$	T m A^{-1}
	of space	$\approx 12.566 \times 10^{-7}$	
μ_N	Nuclear magneton	5.050784×10^{-27}	J T^{-1}

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

Some of the more useful conversion factors for converting from older units to SI units are listed. (Taken from *Standard for Metric Practice,* ASTM E 380-76, Copyright 1976 by the American Society for Testing and Materials, Philadelphia)

To convert from	To	Multiply by
Angstrom	Meter	1.000000×10^{-10}
Atmosphere (standard)	Pascal	1.013250×10^{5}
Bar	Pascal	1.000000×10^{5}
Barn	$Meter^2$	1.000000×10^{-28}
Calorie (thermochemical)	Joule	4.184000
Centimeter of mercury (0 °C)	Pascal	1.33322×10^{3}
Centimeter of water (4 °C)	Pascal	9.80638×10^{1}
Centipoise	Pascal second	1.000000×10^{-3}
Curie	Becquerel	3.700000×10^{10}
Dyne	Newton	1.000000×10^{-5}
Electron volt	Joule	1.60218×10^{-19}
Erg	Joule	1.000000×10^{-7}
Fermi (femtometer)	Meter	1.000000×10^{-15}
Gauss	Tesla	1.000000×10^{-4}
Liter	$Meter^3$	1.000000×10^{-3}
Mho	Siemens	1.000000
Millimeter of mercury	Pascal	1.33322×10^{2}
Poise	Pascal second	1.000000×10^{-1}
Roentgen	Coulomb per kilogram	2.58×10^{-4}
Torr	Pascal	1.33322×10^{2}

Index

A scan, 373
AAPM, 405, 406, 421, 469, 507
AAPM Report 96, 492
Abbey, H., 588
Abduct, 7
Abductor muscle, 7
Aberration
  chromatic, 412
  spherical, 412
Ablation, 404
Able, K. P., 229
Able, M. A., 229
Abraham, R., 277
Abramowitz, M., 209, 263, 359, 361, 362
Absolute temperature, 61, 62
Absorbed dose, 452
Absorption coefficient, 387
Absorption edge, 428
Acceleration, 571
Accommodation, 412
Acetabulum, 8
Acetylcholine, 110, 143, 194, 597
ACHD, 399
Achilles tendon, 6
Acoustic impedance, 366
Acoustic shadow, 378
Actin, 85
Actinometry, 405
Action potential, 141
  foot, 181
  Gaussian approximation, 190
  propagating, 166
  space-clamped, 165
Activating function, 209
Active transport, 519
Activity, 81, 506, 511
  cumulated, 511, 512
Activity vector, 188
Acton, F. S., 340
Adair, R. K., 256, 257, 260, 337
Adenosine triphosphate (ATP), 3
Adiabatic approximation, 440
Adiabatic process, 58
ADP, 72, 555
Adrenal gland, 294
Afterloading, 523
Agre, P., 118
Ahlen, S. P., 442, 444, 448
Ahrens, E. T., 557
Air
  acoustic impedance, 367
  attenuation, 371
  density and specific heat, 65
  speed of sound, 366
Air bladder, 27
Alberts, B., 481
Albumin, 136
Aldosterone, 294
Algae, 229
Aliasing, 314, 329
  in an image, 350
Allen, A. P., 51
Allen, R. D., 108
Almond, P. R., 488
α particle, 448, 506, 525
Alternans, 298
Altes, T. A., 559
Alveoli, 1, 20, 82, 110, 270
Alzheimer's disease, 226
American Association of Physicists in Medicine, *see* AAPM
Ampere (unit), 151
Ampere's law, 216, 225, 227
  and Biot-Savart law, 232
Amplitude attenuation coefficient, 370
Ampullae of Lorenzini, 256, 266
Anaplasia, 401
Anderka, M., 313
Anderson K. E., 422
Anderson, H. L., 379
Anderson, J. R., 533
Anesthetic, 169
Aneurysm, 27
Angiography, 399, 476
Angioplasty, 524
Angioscopy, 399
Angstrom, 3
Angular Frequency, 308
Angular momentum, 536
Angular wave number, 346
Anisotropy, 28, 200, 201
Annihilation radiation, 438, 510
Anode, 202
Anode-break excitation, 181, 210
Anomalous rectification, 264
Antidiuretic hormone, 294
Antineutrino, 509
Antiscatter grid, 474
Antonini, E., 81
Anumonwo, J. B., 172
Aorta, 21, 479
Apoptosis, 480
Appa Y., 422
Aquaporins, 118
Aquaspirillum magnetotacticum, 228
Aqueous, 411
Arakane K., 423
Armato, S. G., 470
Armstrong, B. K., 402
Armstrong, C. M., 394, 421
Arqueros, F., 450, 515, 532
Arterial spin labeling, 556
Arteriole, 21
Artificial insemination, 77
Artificial kidney, 126
Asano H., 423
Ascites, 122
Ashcroft, F., 250
Astigmatism, 412
Astumian, R. D., 256, 259, 260, 265, 337, 338
Ataxia-tangliectasia, 483
Atherosclerosis, 399
Atkins, P. W., 75, 83
Atmosphere, 46
  pressure variation, 64, 80
Atmosphere (pressure unit), 14
Atomic deexcitation, 434
Atomic energy levels, 425
Atomic number, 503
Atoms per unit volume, 388
ATP, 3, 71, 555
Atrioventricular node, *see* see AV node
Atrium, 193
Attenuation
  of sound wave, 370
  water, 371
Attenuation coefficient, 387
  amplitude, 370

R. K. Hobbie, B. J. Roth, *Intermediate Physics for Medicine and Biology*,
DOI 10.1007/978-3-319-12682-1, © Springer International Publishing Switzerland 2015

effective, 391
intensity, 370
linear, 433
mass, 433
Attenunator
ladder, 179
Attix, F. H., 430, 431, 438, 444, 446, 450, 453, 455, 462, 464, 469, 489
Attractor, 276, 284
Auditory evoked response, 328
Auger electron, 435, 449, 507, 524
cascade, 436, 461, 507
Augmented limb leads, 197
Autocorrelation function, 318, 543
of exponential pulse, 325
and energy spectrum, 325
and power spectrum, 320
of noise, 334
of sine wave, 319
of square wave, 319
AV node, 194, 203
Average, 585
ensemble, 56
time, 57
Average reference recording, 205, 211
Avogadro's number, 64, 89, 433
definition in SI units, 388
Axel, L., 546, 555
Axelrod, D., 101, 114, 115
Axial vector, 6
Axon, 2, 141
cable model, 156
electric field, 154
membrane capacitance, 155, 169
membrane capacitance and conductance, 157
membrane equivalent circuit, 162
membrane time constant, 156
myelinated, 142, 167
potassium gate, 164
potassium Nernst potential, 162
potential outside, 185
sodium conductance, 164
sodium gate, 165
sodium Nernst potential, 162
space-clamped, 161, 165
surface charge density, 155
unmyelinated, 142, 167
voltage-clamped, 161
Axoplasm, 142
Ayotte, P., 495

B scan, 373, 378
Background
natural, 492
Backscatter
factor, 455
Backx, P. H., 194
Bacteria, 2
magnetotactic, 228, 235
orientation in a magnetic field, 235
Bacteriophage, 2, 527
Badeer, H. S., 19, 31
Bagavathiappan, S., 422
Bainton, C. R., 276
Balloon angioplasty, 524
Bambynek, W., 435
Banavar, J. R., 45, 51
Bandettini, P. A., 557
Bar (unit), 13
Barach, J. P., 223, 229, 232
Barium, 472
Barium fluorobromide, 470
Barker, A. T., 226
Barlow, H. B., 415, 421
Barn (unit), 428
Barnes, F. S., 257, 266
Barold, S. S., 203
Barr M. L., 422
Barr, G., 91, 114
Barr, R. C., 161, 167
Barrett, H. H., 345, 353, 362
Barrett, J. N., 251
Bart K., 115
Bartels, L. W., 379
Barth, R. F., 489
Bartlett, A. A., 42, 51, 289
Barysch M. J., 421
Barysch, M. J., 403
Basal cell, 401, 403
Basal cell carcinoma (BCC), 402, 403, 486
Basal metabolic rate, 65
Base, 573
Basford, J., 27, 30
Basilar membrane, 370
Basin of attraction, 281
Bass, M., 422, 423
Basser, P. J., 110, 115, 136, 558
Bastian, J., 256
Battocletti, J. H., 555
Baylor, D. A., 415, 422
Bazylinski, D. A., 228
Beam hardening, 471
Bean, C. P., 91, 115, 129, 133, 134, 177
Beaurepaire, E., 423
Beck, R. E., 134
Becklund, O. A., 348, 362
Becquerel (unit), 495, 506, 511
Beer's law, 388
Bees, 229, 235
Begon, M., 281
BEIR, 494, 495, 525, 526, 532
Belousov–Zhabotinsky reaction, 290
Bénard, H., 296
Bender, M. A., 531
Benedek, G. B., 11, 15, 30, 95, 105, 115
Bennison, L. J., 48
Bequerel second, 506
Berg, H. C., 85, 101, 115
Berg, M. J., 48, 51
Berg, W. A., 379
Berger, M. J., 515, 532
Bergmanson, J. P. G., 403, 421
Berliner, L. J., 545
Berlinger, W. G., 51
Berne, B. J., 393, 422
Bernoulli equation, 19, 29
Bernoulli process, 587
Bernstein, M. A., 553
Berry, E., 422, 423
Berwick M., 422
Bessel function, 263, 359, 361, 372
modified, 208
Beta decay, 506, 508
spectrum, 509
Bethe, H., 442
Bethe–Bloch formula, 444
Bevington, P. R., 307
Beyer, R. T., 371, 379
Bianchi, A. M., 328, 329
Bidomain model, 200, 201, 205, 208, 233
Bieber, M. T., 113, 115
Biersack, J. P., 439, 442, 445, 446, 448
Bifocal lenses, 412
Bifurcation, 282
diagram, 285
Bilderback D., 115
Bilinear interpolation, 360
Bilirubin, 400
Binding energy
electronic, 507
nuclear, 505
per nucleon, 505
Binkert, C. A., 379
Binomial probability distribution, 55, 106, 587
Bioheat equation, 404
Biomagnetism, 213
Biot-Savart law, 216
and Ampere's law, 232
and magnetic field around an axon, 219
current in, 218
Bipolar electrode, 203
Birch, R., 463
Bird, R. B., 94
Birds, 229
Birefringence, 393
Births, 341
spontaneous, 313, 314
Bistable systems, 298
Blackbody, 395
Blackbody radiation
and heat loss, 398
vs. frequency, 397
vs. wavelength, 396
Blackman, R. B., 314
Blackman-Tukey method, 329
Blagoev, K. B., 3
Blair, D. F., 101, 115
Blakemore, N., 228
Blakemore, R. P., 228
Bland, E. F., 39, 51
Bloch equations, 539
Bloch, F., 442, 539
Blood flow
pulmonary artery, 316
Blood pressure, 269
Blood-brain barrier, 123, 519
Blount, W. P., 11
Blue (color), 415
Blume, S., 399, 422
Boas, D. A., 422
Boccara, A. C., 423
Bockris, J. O'M., 247
Bodurka, J., 557
Bohr, N., 383, 442
Boice, J. D., 490

Bolch, W. E., 511, 532
BOLD (Blood Oxygen Level Dependent), 543, 557
Boltzmann factor, 58, 62–65, 67, 89, 239, 395, 537
Boltzmann's constant, 61
Bone scan, 521
Boone, J. M., 438
Born charging energy, 151, 247
Born, C. G., 488
Boron neutron capture therapy (BNCT), 489
Bouma, B. E., 422, 423
Boundary layer, 23
Boundary-element method, 200
Bourland, J. D., 202, 555
Boyd, I. A., 597
Bracewell, R. N., 315, 321
Brachytherapy, 484, 489, 523
  high-dose-rate, 523
Bradshaw, P., 24, 30
Bradycardia, 203
Bragg peak, 489, 499
Bragg rule, 446
Bragg-Gray relationship, 490
Bramson, M. A., 422
Braun, T. J., 289
Breathing
  energy loss due to, 65
Bremsstrahlung, 437, 441, 461, 462
  energy fluence, 462
Bren, S. P. A., 258
Brenner, D. J., 494
Březina, V., 314
Brezinski, M. E., 392, 422
Brightness contrast, 474
Brill, A. B., 532
Brink, S., 39, 51
Broad, W. J., 478
Broad-beam geometry, 432
Bronzino, J. D., 328, 329
Brooks, A. L., 494, 500
Brooks, R. A., 478, 480
Brown, J. H., 45, 51
Brown, J. H. U., 423
Brown, R., 89, 108
Brown, R. W., 548
Brownian motion, 65, 89, 108, 332
Buchanan, J. W., 533
Buchsbaum, D., 524
Bucky, G., 474
Budd, T., 449
Budinger, T. F., 511, 512, 514, 517, 533
buffer, 110
Bui, T-A., 51
Buildup factor, 454, 515
Buka, R.L., 403, 422
Bulk modulus, 15
Bundle branch block, 198
Bundle of His, 194
Buonocore, M. H., 350, 362
Buoyancy, 15
Burch, W. M., 109
Burnes, J. E., 200
Bystander effect, 483, 507
Cable equation, 156, 159
  and ladder attenuator, 179
Calcaneus, 6, 7
Calcium, 295
  -induced calcium release, 111
  buffer, 110
  diffusion, 110
  waves, 111
Callaghan, P., 554
Calland, C. H., 126
Calorie, 58, 80
  dietary, *see* Kilocalorie
Calorimetry, 498
Cameron, J. R., 369
Cancer
  and power-line-frequency fields, 257
  prostate, 557
Candela (unit), 410
Capacitance, 149
  concentric cylinders, 176
  cylindrical membrane, 182
  membrane, 169
  resistance and diffusion, 172
Capillary, 2, 21, 110, 121
Capillary blockade, 519
Carbohydrate, 3
^{11}C, 523
^{14}C dating, 532
Carbon dioxide
  production, 270
  regulation, 271, 273, 275
Carbon monoxide, 79
Carbonyl group, 252
Carcinoma
  basal cell, 402, 486
  squamous cell, 402
Cardiac arrest, 76
Cardiac cell, 1
Cardiac output, 22
Carlsson, G. A., 431
Caro, C. G., 24, 30
Carr–Purcell (CP) Sequence, 547
Carr–Purcell–Meiboom–Gill (CPMG) Sequence, 547
Carslaw, H. S., 96, 99, 105, 112, 115
Carson, P. L., 371, 379
Carstensen, E. L., 258
Cartilage
  articular, 136
Castelli, W. P., 51
Catalyst, 79
Cataracts, 118, 480
Catfish, 256
Catheter, 399
Cathode, 202
  virtual, 203
    "dog bone", 205
Cathode-break excitation, 210
Cathode-ray tube, 213
Cavitation, 375
Cavity radiation, *see* Blackbody radiation
Cebeci, T., 24, 30
Cell
  eukaryotic, 2
  membrane, 2
  producing or absorbing a substance, 101
  prokaryotic, 2
  size, 2
  sorting, 236
Cell culture, 480
Cell survival
  fractionation curve, 482
Cellular automata, 291, 298
Center of gravity, 4
Center of mass, 385
Central slice theorem, 351
Centrifuge, 27, 28, 79
Centripetal acceleration, 27, 215
Cerebral cortex, 223
Cerutti, S., 328, 329
Cesium iodide, 470
Chamberlain, J. M., 422, 423
Chance, B., 389, 392, 422, 423
Chandler, W. K., 263
Chang, W., 422
Channel
  selectivity, 252
Channelopathies, 250
Channels
  calcium, 250
  chloride, 251
  delayed rectifier, 250
  ion, 143
  potassium, 250
  sodium, 250
  two-state model, 252
Chaos, 335
  deterministic, 279, 285
  in heart cells, 291
Chaotic behavior, 284
Characteristic x rays, 461
Charcoal, activated, 48
Charge
  free and bound, 150
Charge distribution
  cylindrically symmetric, 145
  line, 145
  on cell membrane, 147
  plane sheet, 145
  point, 145
  spherically symmetric, 145
Charge inversion, 245
Charge screening, 426
Charged-particle equilibrium, 453, 464
Charman, W. N., 413, 422
Chase, M., 527
Chavez, A. E., 438
Cheeseman, J., 403, 422
Chemical dosimeter, 469
Chemical potential, 66, 92, 120, 121
  ideal gas, 67
  solute, 67, 73
  water, 73
Chemical shift, 555, 562
  images, 555
Chemostat, 40
Chemotaxis, 98, 101
Chemotherapy, 46
Chen, J., 422
Chen, W., 557
Cherry, S. R., 517, 520, 522, 523, 531
Cheyne–Stokes respiration, 289

Chick, W. L., 295
Chittka, L., 422
Cho, Z.-H., 351, 357, 362, 548
Cholesterol
  Raman spectrum, 393
Christian, P., 523, 532
Chromatic aberration, 412
Chromosome, 481, 482
Chronaxie, 179, 202
Chronic granulocytic leukemia, 290
CIE, 401, 409
Cilia, 370
Circadian rhythm, 341
Circulation, 20, 294
Circulatory system, 20
Clark, J., 190–192, 200, 208, 220
Clark, V. A., 39, 51
Clarke, J., 230
Classical electron radius, 430, 443
Clausius-Clapeyron equation, 83
Clearance, 40, 293, 404
Clement, G. T., 375, 379
Clostridium, 153
Cloud chamber, 449
Cochlea, 369
Cochlear duct, 370
Cochlear implant, 202, 370
Cochran, W. W., 229
Coffey, J. L., 516
Cohen, A., 329
Cohen, B. L., 494, 495, 526
Cohen, D., 228
Cohen, L. G., 225
Coherence, 393
Coherent scattering, 427, 431
Cole C., 422
Collagen, 136
Collective dose, 493
Collett B., 115
Collimator, 471
  gamma camera, 520, 530
  multi-leaf, 487
Collision kerma, 464
Collision time, 90
Color blindness, 76, 415
Color flow imaging, 375
Color vision, 415
Colyvan, M., 297
Commission Internationale de l'Eclairage, *see* CIE
Common bundle, 194
Compass
  in birds, 229
Competitive binding assay, 503
Complex exponential, 311
Complex notation, 322
Complex numbers, 311
Compound interest, 33
Compound microscope, 420
Compounds and mixtures, 434
Compressibility, 15, 364
Compressive strength, 13
Compton scattering, 427, 428, 466
  cross section, 430
  differential cross section, 430
Compton wavelength, 429
Computed Radiography (CR), 470, 474
Computed tomography
  spiral, 478
Concentration
  and potential difference, 240
Concentration work, 67
Conductance, 151
Conduction system, 194, 203
Conduction velocity
  myelinated, 167, 169
  unmyelinated, 167, 168
Conductivity
  anisotropic, 200, 201
  interior and exterior, 189
  non-uniform, 200
  tensor, 200
Conductor, 148
Cones (retinal), 409, 413
Conformal radiation therapy
  three-dimensional, 487
Congestive heart failure, 38
Conjunctivitis, 403
Constant-field model, 248
Contact lens, 113, 403
Continuity equation, 86
  differential form, 88
  integral form, 87
  with creation or destruction, 89
Continuous slowing down approximation, 447
Contrast
  brightness, 474
  exposure, 474
  film, 465
  noise brightness, 475
  noise exposure, 475
Contrast agent, 472
Control system, 278
Convection coefficient, 80
Conversion factors, 610
Convolution, 346
  theorem, 346, 348
Cook, G., 82, 83
Cook, G. J. R., 557
Cooley, J. W., 315
Coordinate system
  rotating, 539
^{64}Cu, 48
Cormack, A. M., 478
Cornea, 113, 403, 411
Cornsweet, T. N., 413, 422
Correlation function, 317
Cortisol, 340
Cosgrove, D. O., 379
Coster-Kronig transition, 435, 507
Coulomb, 143
Coulomb's Law, 143
Coulter counter, 177
Coulter, W. H., 177
Countercurrent exchange
  and heat loss, 138
Countercurrent transport, 127
Coupling medium, 377
Couriel D. R., 422
Coursey, B. M., 524, 532
Covino, B. G., 169
Cowen, A. R., 470
Cox, J. D., 486
Crank, J., 96, 105, 115
Crank-Nicolson method, 111
Creatinine clearance test, 47
Cristy, M., 532
Critical damping, 582
Cross correlation
  and signal averaging, 328
Cross product, 6
Cross section, 90, 388
  differential, 388
  energy transfer, 431
  scattering, 388
  total, 388
Cross-correlation function, 318
Crouzy, S. C., 255
Crowder, S. W., 142
Crowe, E., 511, 515, 533
Crowell, J. W., 289
CSDA, 447
CT, 478
Cuevas, J. M., 521
Cumulated activity, 511, 512
Cumulated mean activity per unit mass, 514
Curie (unit), 512
Curie temperature, 228
Curl, 224
Curl in different coordinate systems, 601
Curran, P. F., 126
Current
  bound, 227
  electric, 151
  free, 227
  total, 17
  volume, 17, 85
Current density, 17
  electric, 151
Current dipole, 187
  electric and magnetic measurements compared, 222
  in spherical conductor, 233
Current dipole moment, 188
  of the heart, 195
Current source, 186
  potential due to, 186
Cyan, 415
Cyclic GMP, 415
Cyclotron, 215
Cyclotron frequency, 215
Cylindrical coordinates, 601
Cystic fibrosis, 589
Cytokinesis, 481

da Luz, L. C. Q. P., 511, 533
da Silva, F. C., 515, 533
Damage, *see* Radiation damage
Dasari, R. R., 422
Data
  surrogate, 336, 337
Davis, L., 160
de Araujo, F. F., 229
de Boer J. F., 422
de Boer, J. F., 393
de Broglie, L., 383
de Jong, N., 379

Dead-time correction, 530
Death rate, 38, 41, 46, 47
DeBlois, R. W., 177
Debye (unit), 175
Debye length, 242
Debye–Hückel model, 244
Decay
  constant, 35
  exponential, 35
  Multiple paths, 41
  rate, 35
    variable, 38
  with constant input, 41
Decibel, 331, 368
Deckers, R., 379
Deexcitation
  atomic, 434
DeFelice, L. J., 254, 327
Deffeyes, K. S., 49, 51, 229
Defibrillation, 76
Defibrillator, 202, 204
Deformation, 12
Degeneracy, 63
Degrees of freedom, 57, 59, 280
Delaney, C., 352, 362
Delaney, T. F., 489
Delannoy, J., 558
Delay-differential equation, 288, 290
Delayed rectifier channels, 250
Delmar, M., 193
Delta function, 323
Delta rays, 449
Deltoid muscle, 25
Demand pacemaker, 209
Demir, S. S., 165, 172
den Boer, J. A., 548, 553
Dendrites, 141
  in cerebral cortex, 223
Denier van der Gon, J. J., 203
Denk, W., 256
Denny, M. W., 15, 30, 65, 80, 82, 83, 113–115, 366, 367, 371, 379, 418, 421, 422
Density
  current, 17
  optical, 465
Density effect, 444, 446
Density gradient separation, 27
Density of states factor, 63
Density-weighted image, 554
Deoxyhemoglobin, 543
Deoxyribonucleic acid (DNA), 3, 25
Department of Energy, 494
Dephasing, 539, 544
Depolarization, 142
Depression, 226
Depth of field, 412
Derivative control, 279
Derivative, partial, 61
Dermis, 402
Detecting weak magnetic fields, 229
Detective quantum efficiency (DQE), 476
Detector
  thin-film transistor, 474
Deterministic effects, 480
Detriment, 490, 491
Deuteron, 506
DeVita, V. T., 485
Dewaraja, Y. K., 515, 532
DeWerd, L. A., 471
Dextrose, 136
Diabetes, 295
Diabetes insipidus, 118
Dialysis, 123
  renal, 126
Diamagnetism, 227
Diamantopoulos, L., 423
Diastole, 20
Diastolic interval, 298
Diatomic molecule, 385
DiChiro, G., 478, 480
Dichromate vision, 415
Dickerson, R. H., 82, 83
Dielectric, 149
  saturation, 245
Dielectric constant, 150
  lipid, 151
  lipid bilayer, 169
  water, 151
Diem, M., 393, 422
Differential equation, 35
  characteristic equation, 581
  homogeneous and inhomogeneous, 581
  linear, 581
  nonlinear, 279
  second order, 582
Diffey, B. L., 401, 403, 422
Diffusion, 89, 125, 563
  and capacitance, 172
  and chemical reaction, 111
  and drift, 102
  and electrotonus, 179
  anisotropic, 110
  as random walk, 106
  between concentric spheres, 173
  between two spheres, 174
  circular disk, 174
  constant, 92
  in one dimension, 98
  in three dimensions, 98, 107
  in two dimensions, 98, 106
  of photons, 390
  of spins in MRI, 558
  self, 94
  sphere to infinity, 173
  tensor, 110, 558
  time-dependent, 391
  to a cell producing or absorbing a substance, 101
  to or from a disk, 99, 112
  to or from a spherical cell, 98
  trace (of tensor), 558, 564
  with a buffer, 110
Diffusion constant
  and viscosity, 94
  determination of, 105
  effective, 111
  photon, 390
  vs. molecular weight, 94
Diffusion equation, 95
  general solution, 104
DiFrancesco, D., 172
Digital detector, 470
  storage phosphor, 470
  thin-film transistor, 470
Digital subtraction angiography, 476
Dimensional analysis, 30
Dimethylchlortetracycline, 402
Diopter (unit), 412
Dipole
  energy in a magnetic field, 536
  magnetic, 535
Dipole moment
  electric, 175, 245
Dirac, P. A. M., 323
Direct Radiography (DR), 470, 474
Discrete Fourier transform, 311
Disease, incidence and prevalence, 47
Dispersion, 586
Displacement, 3, 569
Displacement current, 217
Ditto, W. L., 291
Diuresis, 123
Divergence, 88
Divergence in different coordinate systems, 601
Divergence theorem, 108
Diving, 15
DNA, 78, 481, 483, 507
Dobson unit, 401
Dog bone, 205
Dogfish, 256
Doi, K., 470, 474
Domains, 227
Donnan equilibrium, 239
Doppler effect, 374, 393
Dore, C. J., 379
dos Santos, D. S., 533
Dosdall, D. J., 202
Dose
  absorbed, 452
  effective, 491
  entrance, 465
  equivalent, 491
  exit, 465
Dose equivalent, 490
Dose factor, 511
Dose fractions, 484
Dose rate
  in photochemistry, 419
Doss, M., 494
Dot product, 12
Doubling time, 36
Douglas, J. G., 489
Doyle, D. A., 252
Drexler, W., 422
Drift, 89
  and diffusion, 102
Driving force, 92, 94
Driving pressure, 121, 124
Droplet keratopathy, 403
Drosophilia melanogaster, 252
Dubois, A., 423
Ducros M. G., 422
Duderstadt, J. J., 390, 422
Dummer R., 421
Duncan, W., 489
Durbin, R. P., 133
Duty cycle, 561

Dye
voltage-sensitive, 205
Dyer, A. G., 422
Dynamic pressure, 19
Dynamical system, 280
Dysplasia, 401

e, definition of, 34
Eames, C. and R., 31
Eames, M., 379
Ear, 369
canal, 369
drum, 369
external, middle, inner, 369
inner, 256
mechanoreceptors, 256
ossicles, 369
sensitivity, 368
Early-responding tissue, 483, 484
Echo planar imaging (EPI), 553
Eckerman, K. F., 511, 532
Eddy currents, 225
Edema, 122
and osmotic pressure, 122
cerebral, 123, 136
pulmonary, 125
Eder, 237
Eder, S. H. K., 229
Edge spread function, 349
EEG, *see* Electroencephalogram
Effective dose, 491
Efimov, I. R., 205
Ehman, R. L., 556
Einstein (unit), 417
Einthoven's triangle, 197
Eisberg, R., 227, 230, 383, 386, 387, 422, 505, 508, 509
Eisenman, G., 115
Ejection fraction, 521
El-Samad, H., 279
Elastic recoil pressure, 20
Elastography, 375, 399, 556
Electric dipole moment, 245
Electric displacement, 217
Electric field
definition, 144
Electrical potential, 147
Electrical stimulation, 202
Electrocardiogram, 185
and solid angle theorem, 207
leads, 196
left ventricular hypertrophy, 198
normal, 198
R-R interval, 208
right bundle branch block, 199
right ventricular hypertrophy, 198
Electrode
anodal, 203, 209
cathodic, 203, 209
pacing, 203, 209
spherical, 177
Electrode heating
during magnetic stimulation, 234
in a changing magnetic field, 234
Electroencephalogram, 185, 205, 210, 211, 223, 329
average reference recording, 205
Electromagnetic spectrum, 382
Electromagnetic wave, 382
Electromotive force, 225
Electromyogram, 185, 202, 329
Electron
Auger, 435
classical radius, 430
for radiation therapy, 488
rest energy, 432
rest mass, 432
Electron capture, 509
Electron microscope, 383
Electron volt (unit), 62, 151, 382
Electroporation, 257
Electrosurgery, 177
Electrotonus, 159, 167
Elementary charge, 63
Elias, W. J., 379
Elliott, D. M., 28, 30
Ellis, S., 375, 379
Elliston, C. D., 494
Elson, H. R., 488
Embedding, 335
dimension, 335
time lag, 335
Emission
spontaneous and stimulated, 542
Emissivity, 395, 398
of skin, 398
Emmer, M., 379
Emmetropia, 412
End-plate potential, 597
Endo, A., 532
Endocardial and epicardial membrane currents, 193
Endoplasmic reticulum, 111
Energy
equipartition theorem, 65
fluence, 404, 406, 407
fluence rate, 406, 408
imparted, 450, 452
ionization, 384, 426
kinetic, 11
lost and imparted, 447
radiant, 407
transferred, 450
net, 452
Energy exchange, 60
Energy levels, 57, 60, 383
atomic electron, 425
band, 395
hydrogen atom, 383
rotational, 385
splitting, 395
vibrational, 386
Energy spectrum
of a pulse, 324
of exponential pulse, 325
Energy-momentum relationship
relativistic, 429
Engdahl, J. C., 521, 533
Enquist, B. J., 51
Ensemble, 54
Ensemble and time averages, 60
Ensemble average, 56
Enthalpy, 73
Entrainment, 282, 296
Entropy, 62, 68
ideal gas, 81
of mixing, 72
Entry region, 24
Enzyme, 79
Eosinophilic granuloma, 521
Epicardium, 200
Epidemiological studies, 257
Epidermis, 401
Epiphyses, 521
Epiphysis, 8–10, 521
Epstein, A. E., 291, 292
Equal anisotropy ratios, 201
Equilibrium
approach to, 273
charged-particle, 453
transient, 453
diffusive, 66
in volume exchange, 68
particle, 66
radiation, 452
rotational, 4
thermal, 60, 66
translational, 4
Equilibrium absorbed dose constant, 511, 514
Equilibrium and steady state
distinction, 59
Equilibrium constant, 71
Equilibrium state, 57
Equipartition of energy, 65
Equivalent dose, 491
Erector spinae muscle, 26
Ergodic hypothesis, 60
Erhardt, J. C., 521
Error
mean square, 304
Error function, 105, 160, 179, 202, 209
Erythema, 401, 419
Erythrocyte, 2, 123
Escape peak, 466
Escherichia coli, 527
Escherichia coli, 2, 85, 101
Esenaliev, R. O., 393, 422
Essenpreis, M., 422
Ethanol, 48
Eukaryotes, 2
Eustachian tube, 369
Evans, R. D., 508, 532
Even function, 308
Evoked response, 205, 224, 328
Excess absolute risk, 493
Excess relative risk, 493
Exchange term, 446
Excitable media, 298
waves in, 290
Exit dose, 465
Exitance, 406, 408
Expectation value, 390
Exponent, 573
Exponential
complex, 311
decay, 35

function, 34
growth, 33
Exponentials, fitting, 42
Exposure, 464
Exposure contrast, 474
Extensive variables, 69
Exterior potential
and electrocardiogram, 188
and solid angle, 189
arbitrary pulse, 190
cardiac depolarization, 187
from action potential, 189
from depolarization, 188
general case far from axon, 190
nerve impulse, 187
ratio to interior, 188
Extinction coefficient, 417
Extracellular fluid, 142
Extracorporeal photopheresis, 404
Eye
emmetropic (normal), 412
human, 410
hypermetropic (farsighted), 412
insect, 410
myopic (nearsighted), 412
resolution, 413
under water, 421

Factorial, 587
Faez, T., 371, 379
Fainting, 203
Falen, S. W., 517, 520, 533
False negative, 587, 605
False positive, 587, 605
Far field, 372
Farad (unit), 149
Faraday constant, 64
Faraday induction law, 224, 544
Faraday, M., 224
Farmer, J., 392, 422
Farrell, T. J., 422
Fast Fourier Transform (FFT), 315, 341
Fatt, I., 113, 115
Fédération CECOS, 76
Feedback
carbon dioxide regulation, 270
loop, 269
negative, 269
one time constant, 273
positive, 269
steady-state conditions, 270
summary, 288
time constant and fixed delay, 287
two time constants, 276
Feinstein, A. R., 588
Feld, M. S., 422
Femur, 8
Feng, T.-C., 422
Fenster, A., 371, 379
Fercher, A. F., 392, 422
Ferrara J. L. M., 422
Ferrimagnetism, 227
Ferromagnetism, 227
Ferrous sulfate, 469
Feynman's ratchet, 337
Feynman, R., 337
Fibrillation
ventricular, 204
Fibula, 6, 7
Fick tracer method, 108
Fick's first law, 92
Fick's second law, 95, 104, 242, 390, 404
Field of view, 349
Film
as x-ray detector, 465
gamma, 465
speed, 465
Film-screen combination, 474
Filtration
of x-rays, 471
Filtration coefficient, 123
Finegold, L., 228
Finlay, J. C., 399, 423
First law of thermodynamics, 57
Fish
breathing, 80
gills, 128
Fisher, H. L., 516, 533
Fishman, H. M., 256, 266
Fishman, R. A., 123
Fission
nuclear, 506, 519
Fit
linear, 304
polynomial, 305
Fitzgerald, A. J., 394, 422, 423
FitzHugh-Nagumo model, 296
Fitzmaurice, M., 422
Fixed point
stable, 276
unstable, 276
Flannery, B. P., 84, 115, 362
Fletcher D. A., 115
Fletcher, D. A., 85
Flotte, T., 422
Flounder
ocean, 256
Flow
laminar, 15
Poiseuille, 17
total, 17
Flow (mathematical), one dimensional, 276
Flow effects (in MRI), 555
Flow rate, 85
Fluence
energy, 404, 436
particle, 85, 390, 436
photon, 390
volume, 85
Fluence rate, 85, 390
volume, 17, 123
Fluid
Newtonian, 16, 91
Fluorescein, 402
Fluorescence, 205, 386, 427, 435
Fluorescence yield, 435, 437
^{18}F, 215, 510, 523
^{19}F, 557
Fluorodeoxyglucose, 523
Fluoroscopy
for fitting shoes, 497
Flux, 85
total, 17
volume, 85
Flux density, 85
volume, 17
Flux transporter (magnetic), 230
Focal length, 411
Follicle-stimulating hormone, 341, 342
Food consumption, 44
Force, 4
generalized, 69
nuclear, 505
surface, 12
Ford, M. R., 516, 533
Foster, K. R., 258, 261
Fourier integral, 320, 322
complex, 322
Fourier series, 308
complex, 311
discrete, equally-spaced data, 310
for piecewise continuous function, 315
for square wave, 316
sample duration not a period, 313
square wave, 311
Fourier theorem, 316
Fourier transform, 320, 322, 543, 551
complex, 322
discrete, 311
discrete (2-d), 351
fast, 315
of exponential pulse, 322
Fourtanier, A. M., 402, 422
Fovea, 413
Fowler, P. H., 448, 449
Fox, J. J., 292
Fox, R. A., 524
Fractional distribution function, 528
Fractional rate of growth or decay, 38
Fractionation curve, 482
Fraden, J., 399, 422
Framingham study, 39
Francis, J. E., 467
Frankel, R. B., 228, 229, 231
Fraunhoffer zone, 372
Free body diagram, 8
Free currents, 227
Free expansion, 78
Free induction decay (FID), 544, 546
Free radical, 469, 481
Freeston, I. L., 226
Freitas, C., 226
French L. E., 422
Frequency, 308
angular, 308
half-power, 331
spatial, 346, 349
Frequency response, 295, 330
Fresnel zone, 372
Frey, E. C., 532
Fricke dosimeter, 469
Friedman, R. N., 223
Fritzberg, A. R., 524, 532
Fruit fly, 252
shaker mutant, 252
Fuchs, A. F., 31
Fujimoto, J. G., 422

Function
piecewise continuous, 315
Functional MRI, 557
Fundamental constants, 609
Fundamental equation of thermodynamics, 69
Fung, Y. C., 12, 13, 30
Furocoumarins, 402
Fusion
nuclear, 506
Fuster, V., 422

Gadolinium, 543, 557
Gadolinium oxysulfide, 466
Gain, 272
open-loop, 272
^{67}Ga (gallium), 520, 522
Gamma (film), 465
Gamma camera, 520
Gamma decay, 506, 507
Gamma function, 600
Gammaitoni, L., 337
Gangrene, 153
Gans, D. S., 423
Gao, L., 350, 362
Gap junctions, 193
Gap phase, 481
Garfinkel, A., 291, 298
Gas
ideal, 64, 65
Gas constant, 64
Gas multiplication, 468
Gasiorowicz, S., 396, 422
Gaskill, J. D., 349, 362
Gastrocnemius, 6, 7
Gating current, 253
Gauss, 213
Gauss (unit), 213
Gauss's law, 144
for magnetic field, 215
free and bound charge, 150
Gaussian distribution, 96, 104, 592
Gawande, A., 589
Geddes, L. A., 142, 202
Geiger counter, 468
Geiger's rule, 499
Geise, R., 475, 477
Gel dosimeter, 469
General thermodynamic relationship, 69
Generalized force, 69
Generalized functions, 323
Gennari, F. J., 123
George, A. L., 193
Germanium, 469
Gersten, K., 23, 31
Gevenois, P. A., 480
Gevins, A., 224
Giaccia, A. J., 480, 481, 483, 485
Giasson, 422
Giasson, C. J., 403
Gibbs factor, 81
Gibbs free energy, 70
Gibbs paradox, 72
Gibbs phenomenon, 316
Gielen, F. L. H., 223, 236
Giercksky, K. E., 422
Gilhuijs, K. G., 379
Gillooly, J. F., 51
Gimm, H. A., 428
Gingl, Z., 337, 338
Ginzburg, L., 297
Glaser, K. J., 556
Glass, L., 280–282, 284, 286, 288–291, 300, 335–337
Glazier, D. S., 45, 51
Glide-Hurst, C. K., 371, 379
Globulin, 136
Glomerulus, 17, 28, 123
Glucose, 3, 72, 123
Glucose regulation, 295
Glutamate, 143
Gluteus medius, 8
Gluteus minimus, 8
Glycine, 143
Glycoaminoglycans, 263
GMP, cyclic, 415
Goff, J. P., 279
Goitein, M., 486–489
Goiter, 292
Goldberg, M. J., 51
Goldman–Hodgkin–Katz equations, 249
Gompertz mortality function, 47
Goodsell, D. S., 3, 30
Gorby, Y., 228
Gou, S-y., 101, 115
Gould, J. L., 229
Gouy-Chapman model, 241
Gradient, 92, 148
diffusion, 558
magnetic, 549
phase-encoding, 550
readout, 550
slice-selection, 548
Gradient in different coordinate systems, 601
Gradiometer, 231
Graininess, 475
Granger, H. J., 121
Grant, R. M., 369
Gras, J. L., 109
Grass, Inc., 205
Gray (unit), 452, 491
Gray body, 396
Gray, D. E., 379
Gray, R. A., 290, 291
Green (color), 415
Green's function, 104
Gregory, K., 422
Greigite, 228
Greinix H., 422
Greivenkamp, J. E., 413, 422
Grenier, R. P., 531
Grid
antiscatter, 474
Griffiths, D. J., 213
Grosberg, A. Y., 245
Grossweiner, L., 386, 389, 391, 404, 405, 422
Growth rate
variable, 38
Grundfest, H., 190, 191, 220
Guevara, M. R., 291
Gulrajani, R. M., 185, 200
Guttman, R., 46
Guyton, A. C., 21, 121, 143, 276, 289, 340
Gyromagnetic ratio, 536, 537

h gate, 164, 181
Haas, R. A., 589
Hafemeister, D., 257
Hagen, S. J., 40, 43, 51
Hair cells, 256, 370
Halberg, F., 313, 341
Haldane, J. B. S., 48, 51
Half-life, 35, 507
effective, 513
Half-power frequency, 331
Half-value layer, 498
Hall, E. J., 480, 481, 483, 485, 494
Hall, J. E., 121, 123, 128, 143
Hall, M. A., 286
Hallett, M., 225
Halliday, D., 398
Hämäläinen, M., 229, 230
Hämäläinen, R., 334
Hamblen, D. P., 467
Hamill, O. P., 251, 255
Hamilton, L. J., 390, 422
Hanlon, E. B., 394, 422
Hao O-Y., 422
Hao, D., 557
Hao, O-Y., 403
Hari, R., 224
Harisinghani, M. G., 557
Harmonic images, 374
Harmonic oscillator, 277, 386, 583
Harmonics, 308
Harri, R., 334
Harris, C. C., 467
Harris, J. W., 189
Hart, H., 51
Hartmann, W. M., 370, 379
Hartree–Fock approximation, 444
Haskell, R. C., 422
Hasselquist, B., 521, 522
Hastings, H. M., 298
Hawkes, D. J., 431
Hayes, N. D., 288
Haynie, D. T., 54, 83
HCl, 417
He, B., 235
Heald, M. A., 410, 422
Health Physics Society, 494
Heart block, 203
second degree, 209
sinus exit, 209
Heart failure, 122, 203
Heart rate, 50
Heat capacity, 65, 293
air, 65, 80
at constant volume, 65
ideal gas, 65
specific, 65
water, 80
Heat conduction, 109, 112, 404
Heat engine, 75
Heat flow, 58, 65
Heat loss
and countercurrent exchange, 138

Heat stroke, 289
Heaton, C. L., 401
Hecht, S., 413, 414, 422
Hee, M. R., 422
Helical CT, 478
Hellman, L., 51
Helmholtz, 142
Hematocrit, 24
Hemmingsen, A. M., 45, 51
Hemochromatosis, 228
Hemoglobin, 2, 25, 392, 543, 557
Hemolysis, 400
Hemosiderosis, 228
Hempelmann, L. H., 497
Hendee, W. R., 363, 373, 376, 379, 471, 474, 477
Henriquez, C. S., 192, 200, 201
Hereditary spherocytosis, 123
Hermann, M., 422
Herrick, J. F., 17, 31
Hershey, A., 527
Hertz (unit of frequency), 308
Hielscher, A., 391, 422
High-temperature superconductors, 230
Higson, D. J., 494
Hilborn, R. C., 277, 280, 286
Hildebrand, J. H., 72, 83, 91, 115, 121
Hill, W., 152
Hille, B., 31, 162, 247, 250, 256
Hiller J. E., 423
Hine, H. G., 531
Hip, 7, 8
Hitzenberger, C. K., 422
Hobbie, R. K., 194, 197, 198, 450
Hodgkin's disease, 485
Hodgkin, A., 142, 155, 160, 161, 179
Hodgkin–Huxley model, 3, 161, 163, 165, 166, 283, 290
Hofbauer G. F., 421
Hoffmann, P. M., 85, 115
Hogstrom, K. R., 488
Hole, 461
Homeotherms, 45
Honig, B., 243
Hormesis, 494
Horowitz, P., 152
Hosaka, H., 222
Hot tub, 289
Hounsfield (unit), 478
Hounsfield, G. N., 478
Howell, R. W., 507, 533
Hu, X-p., 350, 351
Huang, D., 392, 422
Huang, S., 53, 84
Hubbell, J. H., 428, 429, 431, 432, 435, 438, 472
Hubbert's peak, 49
Human Immunodeficiency Virus (HIV), 2
Humm, J. L., 507
Hunt, J. G., 515
Husband, J. E., 557
Huxley, A. F., 142, 161
HVL, 498
Hydraulic permeability, 123
  pore model, 129
Hydrogen peroxide, 481
Hydrogen spectrum, 383
Hydrostatics, 13
Hydroxyl free radical, 481
Hynynen, K., 375, 379
Hyperbolic functions, 174
Hyperopia, 412
Hyperplasia, 401
Hyperpolarization, 181, 559
Hyperthermia, 404
Hyperthyroidism, 524
Hypertrophy, 401
Hypoproteinemia, 122
Hypothalamus, 270, 294
Hysteresis, 13, 20
  magnetic, 227

Ice ages, 337
ICRP, 490, 491
ICRU, 370, 374, 379, 390, 438, 439, 441, 444–448, 450, 462–464, 471, 472, 474, 492, 501, 511
Ideal gas, 78
Ideal gas law, 64, 67
Ideal solution, 121
Ideker, R. E., 292
Iliakis, G., 51
Illuminance, 406
Ilmoniemi, R. J., 226, 334
Image
  density-weighted, 554
  T_1-weighted, 554
  T_2-weighted, 554
Image contrast and MR pulse parameters, 554
Image quality, 474
Imaginary number, 311
Impact parameter, 442
Impedance
  acoustic, 366
  characteristic, 367
  reactive component, 367
  transformation in middle ear, 369
Impulse approximation, 440, 443
Impulse function, 323
Impulse response, 331, 345
Incidence, 402
Incidence of disease, 47
Incoherent scattering, 427, 431
Incremental-signal transfer factor, 466
Incubator
  neonatal, 399
Incus, 369
Index of refraction, 381, 387
  of aqueous, 411
  of the lens, 411
  of vitreous, 411
Inelastic scattering, 427
Inflammatory response, 122, 401
Influence function, 104
Infrared
  photography, 399
  radiation, 382, 398
  spectroscopy, 391
Ingard, K. U., 363, 379
Inman, V. T., 8, 9, 31
Inner ear, 369
Insulin, 295
Integral control, 279
Integrals
  trigonometric, 579
Intensifying screen, 465
Intensity
  of sound, 367
  radiant, 407
  reference, 368
Intensity attenuation coefficient, 370
Intensity-modulated proton therapy (IMPT), 489
Intensity-modulated radiation therapy, 487
Interferometer, 393
Internal conversion, 507
Internal radiotherapy, 524
International Commission on Radiation Units and Measurements, *see* ICRU
International Commission on Radiological Protection, *see* ICRP
International Committee on Non-ionizing Radiation Protection (ICNIRP), 261
Interpolation, 352
Interstitial fluid, 121
Intestinal parasites, 589
Intestinal villi, 128
Intracellular fluid, 142
Intravascular ultrasound, 399
Inulin, 417
Inverse problem, 207
Inversion recovery (IR), 546
Iodine, 472
^{123}I, 520
^{131}I, 522, 524
Iodine deficiency, 292
Ion
  Born charging energy, 151
Ionization chamber, 468, 490
Ionization energy, 384, 426
  average, 444
Ionizing radiation, 425
^{192}Ir, 524
Iris, 411
Iron oxide, 543
Iron stores in the body, 228
Irradiance, 406, 408
Isodose contours, 486
Isosbestic point, 392
Isotonic saline, 136
Isotope, 504
Isotropic radiation, 409
Itoh, M., 423
Itzkan, I., 422
Izatt, J. A., 394

Jackson, D. F., 431
Jackson, J. D., 431
Jacques, S. L., 422
Jaeger, J. C., 96, 99, 105, 112, 115
Jaksch P., 422
Jalife, J., 172, 205, 223, 283
Jalinous, R., 226
Jang, I-K., 422
Janks, D. L., 205, 210
Jaramillo, F., 337

Jaundice, neonatal, 400
Jayakumar, T., 422
Jeanmonod, D., 375, 379
Jeffrey, K., 203, 209
Jeffreys, B. S., 159
Jeffreys, H., 159
Jejunum, 178
Jewett, J. W., 7, 148, 423
Jiang, D., 209
Jiang, Y., 252
Johnsen, S., 381, 422
Johnson noise, 255, 332, 333
Johnson, G. F., 51
Johnson, J. B., 333
Joint probability distribution, 605
Jones, J. P., 362, 548
Jones, R. H., 531
Jordan, A., 229
Joseph, P. M., 546
Josephson junction, 230
Joule (unit), 11
Juziene, A., 422

K edge
  in contrast agent, 472
K shell, 461
Kagan, A. R., 485
Kaiser, I. H., 313, 341
Kalender, W. A., 478–480
Kalmijn, Ad. J., 256
Kaluza, G. L., 524, 533
Kaminer, M., 375, 379
Kanal, E., 555
Kandel, S., 193
Kane, B. J., 142
Kannel, W. B., 51
Kaplan, D., 280, 286, 291, 335, 336
Karhu, J., 226
Karp, J. S., 523, 533
Kassell, N. F., 379
Kassirer, J. P., 123
Kassis, A. I., 480, 483, 507, 524, 533
Katchalsky, A., 126
Kathren, R. L., 494
Katz, B., 141, 143, 161
Keener, J. P., 270, 291
Keizer, J., 111, 115
Kelvin (unit), 61
Kempe, C. H., 44, 51
Keratin, 401
Keratitis, 403
Keratopathy, 403
Keriakes, J. G., 488
Kerma, 452
  collision, 452, 453
  radiative, 452
Kernicterus, 400
Ketcham, D., 473
Kevan, P. G., 410, 422
Kevles, B. H., 345, 362
Keynes, R. D., 252
Khammash, M., 279
Khan, F. M., 486–489, 533
Khurana, V. G., 261
Kidney, 40, 123, 293
  artificial, 126
  glomerulus, 17
Kilocalorie, 58, 80
Kim, B. H., 543
Kimlin, M. G., 422
Kinesin, 85
Kinetic energy, 11
King, K. F., 553
Kirchhoff's laws, 153
Kirschvink, J. L., 229, 260, 261
Kiss, L. B., 338
Kissel, L. H., 431
Kleiber's law, 45
Klein-Nishina
  factor, 456
  formula, 430
Kligman, L. H., 402, 422
Knobler R., 422
Knobler, R., 404
Knutson, J. S., 202
Knuutila, J., 334
Kobayashi, A. K., 260
Koh, D.-M., 557
Kohl, M., 379
Köhler, M. O., 379
Köhler, T., 26, 31
Koizumi, H., 423
Kondo, S., 494
Koo, T-W., 422
Kooiman, K., 379
Kooy, H. M., 489
Körner, M., 470
Korotkoff sounds, 21
Kovacs, G. T. A., 142
Kowalsky, R. J., 517, 520, 533
Krämer, U., 422
Kramer's law, 497
Kramer, E. M., 118
Kramer, J. R., 422
Krane, K. S., 398
Kremkau, F. W., 371, 379
Kricker, 421
Kricker, A., 402
Kripke, M. L., 402, 403, 422
^{81m}Kr, 522
Kupfer cells, 519
Kwong, K. K., 557

LaBarbera, M., 28, 31
Lahiri, B. B., 422
Lakshminarayanan, A. V., 357, 362
Lamb, H., 260
Lambert's law, 409
Laminar flow, 15, 16
Lanchester, F. W., 296
Langevin equation, 91
Langrill, D. M., 210
Lapicque, L., 202
Laplacian, 95
Large-signal transfer factor, 466
Larin, K. V., 422
Larina, I. V., 422
Larmor precession frequency, 538, 542
Laser surgery, 404
Lasser, T., 422
Late-responding tissue, 483, 484
Latent heat of vaporization, 79
Lateral geniculate nucleus, 557
Latitude, 465
Laüger, P., 155
Lauterbur, P. C., 548
Law of Laplace, 27, 82
Lazovich D., 422
Lazovich, D., 403
Le, T. H., 350, 351
Lead zirconate titanate, 371
Leads
  ECG, 196
  limb, 196
    augmented, 197
  precordial, 197
Least squares, 211, 303
  linear, 304
  nonlinear, 306
  polynomial, 306
Lebec, M., 423
LeBihan, D., 115
LeDuc, P. R., 279
Left and right bundles, 194
Left ventricular hypertrophy, 198
Legendre polynomial, 189, 192
Lei, H., 555
leibnitz, 53
Leigh, J., 423
Lens, 411
Lenz's law, 225, 227
LET, 447
Leuchtag, H. R., 249
Leukemia, 45, 290
Levenberg-Marquardt method, 307
Levitt, D., 91, 103, 115, 130, 133
Levitt, M. H., 543
Levy, D., 39, 51
Lewis number, 109
Lewis, C. A., 250
Li, T. K., 48
Liang, Z.-P., 548
Light
  photon properties, 382
  polarized, 393
  speed, 381
  wave nature, 382, 392
Lightfoot, E. N., 94
Lighthill, J., 16, 31
Lighthill, M. J., 316, 323
Likavec, M. J., 123
Limb leads, 196
Lime peel, 402
Limit cycle
  bifurcation of, 282
  stable, 281
  unstable, 281
Lin, C. P., 422
Lindemans, F. W., 203
Lindhard, J., 442, 444, 445
Lindner, A., 126
Lindsay, R. B., 371, 379
Line approximation, 192
Line integral, 148
Line spread function, 349
Linear energy transfer, 447, 481

Linear no-threshold model, 493, 494
Linear regression, 304
Linear system, 345
Linear-no-threshold model, 526
Linear-quadratic model, 482
Links, J. M., 521, 533
Liow, J. S., 533
Lipid bilayer, 169
Lissner, H. R., 7, 8, 31
Liter (unit), 21
Lithium iodide battery, 210
Lithotripsy, 375
Little, M. P., 494
Littmark, U., 439, 442, 445, 446, 448
Liu, H., 392, 422
Liver, 122, 228
Ljungberg, M., 532
Local control of tumor, 485
Loevinger, R., 511, 512, 514, 517, 533
Log-log graph paper, 43
Logarithm
  definition, 573
  natural, 573
Logistic equation, 42, 281
Logistic map, 284
Longitudinal wave, 363, 366
Lopes da Silva, F. H., 205
Lorente de Nó, R., 160
Lorentz force, 214, 234
Lorenz, E. N., 296
Lotka-Volterra equations, 49
Lounasmaa, O. V., 334
Lowe, H. C., 422
Lu, J., 256, 266
Lubin, J. H., 495
Lumen, 399
Lumen (unit), 409
Luminance, 406
Luminous
  efficacy, 410
  energy, 406
  exitance, 406
  exposure, 406
  flux, 406, 409
  intensity, 406, 410
Lung transplant, 522
Luo, C. H., 165, 193
Luther, S., 291
Lutz, G., 469
LVH, *see* Left ventricular hypertrophy
Lybanon, M., 306
Lynch, D. K., 410, 423
Lysaght, M. J., 126

m gate, 164
M mode, 374
Mackenzie, D., 51
Mackey, M. C., 280–282, 284, 288–290, 300
MacKinnon, R., 252
MacNeill, B. D., 399, 422
Macovski, A., 474
Macrophage, 228
Macrostate, 56, 59
  disordered, 56
  nonrandom, 56
  ordered, 56
  random, 56
Maddock, J. R., 101, 115
Madronich, S., 400, 401, 422
Magenta, 415
Magic angle, 560
Magleby, K. L., 251
Magnetic dipole, 226
  field far from, 542
Magnetic field, 213, 215
  crayfish axon, 219
  detection, 229
  earth, 229
  from a current dipole in a conducting sphere, 223, 233
  from axon, 219
  from depolarizing cell, 219
  from point source of current, 217
  lines, 216
  remanent, 227
Magnetic field intensity, 227
Magnetic flux quantum, 230
Magnetic moment, 214, 535
  and angular momentum, 536
  nuclear, 537
  torque, 214
Magnetic monopole
  absence of, 215
Magnetic permeability, 227
Magnetic resonance imaging, 535
  intravascular, 399
Magnetic shielding, 225
Magnetic stimulation, 235
Magnetic susceptibility, 213, 227, 228
Magnetically-shielded rooms, 230
Magnetism
  and special relativity, 213
Magnetite, 228, 229, 235
Magnetization, 227, 537
Magnetocardiogram, 219, 222
Magnetoencephalogram, 219, 223, 334
Magnetopneumography, 228
Magnetosome, 228, 229, 235, 260, 261
Magnetotactic bacteria, 228
Magnification, 347
Magnifying glass, 420
Maidment, A. D. A., 379
Mainardi, L. T., 328, 329
Maki, A., 423
Makita, S., 423
Mali, W. P., 379
Malleus, 369
Malmivuo, J., 185
Mammography, 477
Mannitol, 123, 136
Manoharan, R., 422
Manoyan, J. M., 446
Mansfield, P., 548
Mansfield, T. A., 101, 115
Maor, E., 35, 51
Mariappan, Y. K., 556
Maris, M., 423
Maritan, A., 51
Markov process, 587
Marshall, M., 449, 463
Martin, A. R., 597
Martin, E., 379
Mass attenuation coefficient, 433
  tissue, 471
Mass energy absorption coefficient, 437
  table (on the web), 465
Mass energy transfer coefficient, 437
Mass fraction, 434
Mass number, 503
Mass stopping power, 439
Mass unit
  atomic, 505
  unified, 505
Mattiello, J., 115, 558
Maughan, W. Z., 315
Mauricio, C. L., 515, 533
Mauro, A., 241, 242
Mavroidis, C., 126
Maxwell's equations, 234
May, R. M., 285
Mayaux, M. J., 76, 77
Mazumdar, J. N., 16, 21, 23, 31
McAdams,M. S., 422
McCollough, C. H., 490, 491, 494
McDonagh, A. F., 400, 422
McGrayne, S. B., 228
McKee, P. A., 39, 51
McMahon, T., 45, 51
McNamara, P. M., 51
Meadowcroft, P. M., 24, 31
Mean, 585
Mean absorbed dose, 511
Mean absorbed dose per unit cumulated activity, 511
Mean energy emitted per unit cumulated activity, 511
Mean energy per ion pair, 464
Mean energy per unit cumulated activity, 511
Mean free path, 90
Mean initial activity per unit mass, 514
Mean number per disintegration, 507
Mean number per transformation, 511
Mean square error, 304
Meandering waves, 292
Mechanoreceptors, 256
Medel, R., 379
Medical Internal Radiation Dose, *see* MIRD
MEG, 219
Meidner, H., 101, 115
Melanin, 401
Melanocyte, 401
Melanoma, 402, 403
Mellinger, M. D., 422
Membrane, 2, 117
  force on, 134
  permeability, 125
  permeable, 117
  semipermeable, 117
Membrane capacitance, 169
Merckel, L. G., 375, 379
Mermin, N. D., 593
Messner, W. C., 279
Metabolic rate, 45, 270
Metaplasia, 401
Metastable state, 507
Metastases, 485
Metric prefixes, 1

Metric system, 1
Mettler, F. A., 477, 492
Metz, C. E., 474
Michels, L., 379
Micron, 1
Microscope, 420, 421
Microstate, 56, 59
Microstates, 68
  energy dependence, 61
Microtubules, 85
Microwave radiation, 382
Mielczarek, E. V., 228, 291
Miller, J. M., 422
Milner T. E., 422
Milnor, W. R., 21, 24, 31, 316
Miners, 228
Mines, G. R., 290
Minimum auditory field (MAF), 368
Minimum auditory pressure (MAP), 368
Minimum erythemal dose, 401
Miralbell, R., 489
MIRD, 503, 511, 515, 516, 524
Mitochondria, 2
Mitosis, 85, 480, 481
Mitral valve, 193
Mixing
  entropy of, 72
  heat of, 73
Miyachi Y., 423
Mizuno M., 423
ml (unit), 21
Moan, J., 422
Mobile phones, 261
Mode, 586
Model (mathematical), 3
Models, 3
Modulation transfer function, 348, 413, 474
Modulus
  shear, 13
  Young's, 13
Moe, G. K., 283
Mole, 388
Mole fraction, 73
Molecular dynamics, 244
Moller, J. H., 198, 199
Moller, W., 228
^{99}Mo, 519
Molybdenum target x-ray tube, 477
Moment arm, 5, 10
Moment of inertia, 385
Monte Carlo calculation, 389, 391, 450, 515
Monteith, S., 375, 379
Montesinos, G. D., 450, 515, 532
Moonen, C. T., 379
Moore, J. W., 289
Moran, J., 126
Morel, A., 379
Morgan, K., 394
Morin, D. J., 213, 218
Morin, R., 555
Morrison, P., 3, 31
Morse, P. M., 363, 379
Mortality function, 47
Mortimer, M., 281
Moseley, H., 497
Moseley, M. E., 558
Moses, H. W., 202
Moskowitz, B. M., 228
Mosley, M. L., 523, 533
Moss, F., 337, 338
Moss, W. T., 486
Mossman, K. L., 494
Motamedi, M., 422
Mottle, 475
Motz, J. T., 422
Moulder, J. E., 257, 258, 261
Mouritsen, H., 229
Moy, G., 243
Moyal, D. D., 402, 422
MR Image and pulse parameters, 554
Muehllehner, G., 523, 533
Mugler, J. P., 559
Mullin, J. C., 202
Multipole expansion, 192
Mundy L., 423
Murata, K., 118
Murphy, E. A., 588
Murphy, M. R., 400, 422
Murray's law, 28
Murray, J. D., 43, 49, 51, 270
Muscle
  gastrocnemius, 6, 7
  gluteus medius, 8
  gluteus minimis, 8
  soleus, 6, 7
Mutual inductance, 236
Myelinated fiber
  time and space constants, 167
Myers, D. R., 118
Myers, K. J., 345, 353, 362
Myocardial cells
  and nerve cells compared, 193
  repolarization, 186
Myocardial infarct, 39, 48, 222, 399
Myoglobin, 81
Myopia, 412
Myosin, 85

n gate, 164
Nag, S., 523
Naito N., 423
Narmoneva, D. A., 30
Narrow-beam geometry, 432
Nath, R., 524, 532
National Institute of Standards and Technology, *see* NIST
National Nuclear Data Center, 511
National Research Council, 258
Natural background, 492
Natural logarithm, 573
Navier-Stokes equation, 28
NCRP, 477, 492, 494
Near field, 372
Near-infrared radiation, 392
Near-infrared spectroscopy, 399
Nedbal, L., 314
Negative resistance, 265, 266
Neher, E., 251
Nelson J. S., 422
Neonatal jaundice, 400
Neovascular membrane, 394
Neper, 370
Nephron, 123
Nephrotic syndrome, 122
Nernst equation, 64, 155, 239
  sodium, 161
Nernst potential
  calcium, 178
  potassium, 162
  sodium, 162
Nernst-Planck equation, 247
Nerve cell, 79
Nerves
  Sympathetic and parasympathetic, 194
Nesson, M. H., 260
Neuromagnetic current probe, 229
Neurotransmitter, 143
Neutrino, 509
Neutron therapy, 489
Newhauser, W. D., 488, 489
Newman D., 115
Newman, E. B., 369, 379
Newton (unit), 4
Newton's law of cooling, 80
Newton's second law, 11, 53
Newton's third law, 12
Newton, I., 381
Newtonian fluid, 16, 24, 91
Nicholls, A., 243
Nicholls, J. G., 141
Nichols, C. G., 264
Nickell, S., 389, 422
Nielsen, P., 228
Nioka, S., 423
NIST, 433, 438, 448, 465
^{13}N, 523
Nitrogen narcosis, 15
No-slip boundary condition, 15
Noble, D., 172
Nodal escape beats, 194
Node of Ranvier, 142, 167
Noise, 80, 254, 326, 327
  1/f, 335
  autocorrelation function, 334
  Johnson, 255, 332, 333
  pink, 335
  power spectrum, 333
  quantum, 475
  shot, 254, 332, 475
  source, 332
  white, 333
Noise brightness contrast, 475
Noise equivalent quanta, 476
Noise exposure contrast, 475
Nolte, J., 143
Noninvasive electrocardiographic imaging, 200
Nonlinear least squares, 306
Nonmelanoma skin cancer (NMSC), 402
Norepinephrine, 194
Normal distribution, 592
Nuclear force, 505
Nuclear scattering, 445
Nunez, P. L., 205
Nutation, 541, 549
Nyenhuis, J. A., 555
Nyquist frequency, 329

Nyquist sampling criterion, 309
Nyquist, H., 333

O'Brien, D., 44, 51
O'Reardon, J. P., 226
O'Sullivan N., 422
O'Sullivan, N., 403
OATZ, 228
Oberly, L. W., 521
Ocean flounder, 256
Odd function, 308
Oellrich, R. G., 400, 422
Oersted, H. C., 213
Ogawa, S., 557
Ohm (unit), 151
Ohm's law, 151, 190, 199
  and current density, 152
  anisotropic medium, 200
Oldendorf, W., 478
OLINDA/EXM, 511, 515
One-dimensional flow (mathematical), 276
Open-loop gain, 272
Operating point, 271
  approach to, 273
Optical coherence microscopy, 393
Optical coherence tomography (OCT), 392
Optical density, 465
  and exposure, 497
Optical mapping, 205
Optical transfer function, 348
Orbital magnetic moment, 227
Orbital quantum number, 425
Orear, J., 306
Organelle, 2
Orthogonality relations
  for trigonometric functions, 315
Orton, C., 379, 485, 494
Osmolality, 121
Osmolarity, 121
Osmoreceptors, 294
Osmosis
  in ideal gas, 118
  in liquid, 121, 123
  reverse, 125
Osmotic diuresis, 123
Osmotic flow
  difference from diffusion, 117
Osmotic pressure
  and capillaries, 121
  and chemical potential, 120
  and driving pressure, 121
  and interstitial fluid, 121
  in ideal gas, 120
  in liquid, 121
Ossicles, 369
Osteoarthritis, 136, 263
Othmer, J. G., 111, 115
Oudit, G. Y., 194
Oval window, 369
Øverbø, I., 428
Oxic-anoxic transition zone, 228
^{15}O, 523
^{18}O, 215
Oxygen consumption, 270
Oxyhemoglobin, 543
Ozone, 400

P wave, 195, 203
^{31}P, 555
Pacemaker, 202, 203, 282
Pacing electrode, 203
Packard, G. C., 308
Page, C. H., 225
Paine, P. L., 91, 115
Pair production, 428, 432
  free electron, 432
Pallotta, B. S., 251
Pancreas, 295
Panfilov, A. V., 291
Pankhurst, Q. A., 229
Paramagnetism, 227
Paramecium, 1
Parasympathetic nerves, 194
Parathyroid hormone, 295
Parisi, A. V., 422
Parisi, M., 118
Parkinson, J. S., 101, 115
Parseval's theorem, 324
Partial derivative, 61, 607
  order of differentiation, 608
Particle exchange
  equilibrium, 66
Particular solution, 584
Pascal (unit), 13
Patch-clamp recording, 251
Patterson, J. C., 523, 533
Patterson, M. S., 391, 399, 422, 423
Patton, H. D., 21, 31, 123, 128, 141, 143, 155, 271
Pauli exclusion principle, 384, 446
Pauli promotion, 446
Payandeh, J., 252
Payne, J. T., 357
Pearle P., 115
Pearle, P., 108
Peckham, P. H., 202
Péclet number, 113
Pecora, R., 393, 422
Pedley, T. J., 30
Peng, Q., 417, 422
Perfumes, 402
Perfusion, 404
Period doubling, 285
  in heart cells, 291
Periodogram, 329
Perkins, D. H., 448, 449
Permeability
  hydraulic, 123
  magnetic, 227
  membrane, 125
Perrin, J., 89, 108, 115
PET, 557
Peters, R. H., 45, 51
Petridou, N., 557
Phagocytosis, 519
Phantom, 515
Phase, 308
  resetting, 280–282
  singularity, 299
  space, 277
Phase encoding, 548, 551
Phase transfer function, 348
Phased array, 373
Phelps, M. E., 517, 520, 522, 523, 531, 532
Phenobarbital, 48
Phenothiazines, 402
Philip, J., 422
Phlebotomy, 125
Phosphate, 293
Phosphorescence, 387
Photochemotherapy, 404
Photodynamic therapy, 399
Photoelectric effect, 426, 428
Photographic emulsion, 448
Photometry, 405
Photomultiplier tube, 520
Photon, 382
  attenuation coefficient, 433
  characteristic, 435
  diffusion constant, 390
  exitance, 406
  fluence, 406
  fluence rate, 406
  fluorescence, 435
  flux, 406
  flux exposure, 406
  flux intensity, 406
  flux irradiance, 406
  flux radiance, 406
  secondary, 436, 437, 454
Photon absorption, 387
Photon scattering, 387
Photons, 257
Photopheresis, 404
Photopic vision, 409
Photosynthesis, 72, 314
  in aquatic plants, 418
Physiological recruitment, 209
π, 541
$\pi/2$, 544
Pickard, W. F., 256
Piezoelectric transducer, 371
Pigeon, 229, 235
Pillsbury, D. M., 401
Pinna, 369
Pirenne, M. H., 413, 414, 422
Pisano, E. D., 477
Pituitary, 270, 292
Planck's constant, 230, 257, 382
Planck, M., 396
Plane wave, 363, 409
Plank, G., 205
Plaque, 399
  vulnerable, 399
Plasma, 40
Plasma membrane, 239
Plasmon, 442
Plasticity, 205
Platzman, R. L., 464
Plonsey, R., 21, 160, 161, 167, 185, 190–192, 200, 202, 208, 220
Plug flow, 556
Pneumothorax, 472
Pogue, B. W., 391, 422
Poikilotherms, 45
Poincaré, J. H., 279

Point charge, 145
Point-spread function, 347, 413
  for an ideal imaging system, 347
Poise (unit of viscosity), 16
Poiseuille flow, 17, 21, 128
  departures from, 22
Poiseuille, J. L. M., 17
Poisson distribution, 389, 414, 595
  and shot noise, 255
Poisson equation, 242
Poisson statistics
  in cancer incidence, 493
  in cell survival, 485
  in image, 475
  in photon fluence, 480
  in radiation damage, 482
Poisson's ratio, 28
Poisson-Boltzmann equation, 242
  analytic solution, 263
  linearized, 242
Polarization, 394
  electric, 246
  of dielectric, 149
Polarized light, 393
Polk, C., 257, 259–261
Polynomial
  regression, 306
Population inversion, 417
Positron, 428, 432, 438
  decay, 508
Positron Emission Tomography (PET), 215, 523
Postulates of statistical mechanics, 59
Potassium channels
  shaker, 254
  slow, 193
Potassium conductance, 163
Potassium gate, 164
Potassium Nernst potential, 162
Potential
  electric, 147
  outside axon, 185
  transmembrane, 205
Potential Alpha Energy Concentration (PAEC) (unit), 526
Potential difference, 148
  and ion concentration, 240
Potential energy, 147
Powell, C. F., 448, 449
Power, 11, 153
Power law, 43
Power spectral density (PSD), 328
  Blackman-Tukey method, 329
  magnetic noise, 334
  periodogram, 329
Power spectrum, 317
Power, average, 324
Prausnitz, J. M., 72, 83, 91, 115
Precession, 538
Precordial leads, 197
Predator-prey problem, 49
Prephasing lobe, 562
Press, W. H., 76, 84, 111, 115, 191, 211, 219, 220, 303, 307, 314, 315, 328, 329, 360, 362
Pressure, 14
  diastolic, 20
  dynamic, 19
  systolic, 20
Preston, G. M., 118
Prevalence of disease, 47
Prickle layer, 401
Primary colors, 415
Principal quantum number, 425
Probability, 54
  binomial, 55
Product
  cross, 6
  dot, 12
  scalar, 12
  vector, 6
Projection, 522
Projection reconstruction, 548, 551
Projection theorem, 351
Projections, 351
Prokaryotes, 2
Proportional control, 279
Proportional counter, 468
Protein, 2
Proton therapy, 488
Protoplast, 2
Protozoans, 1
PRU (peripheral resistance unit), 21
Pryde, J. A., 91, 115
Pseudovector, 6
Psoralen, 404
Psoriasis, 403
Pterygium, 403
Puliafito, C. A., 422
Pulmonary embolus, 517
Pulse echo techniques, 373
Pulse oximeter, 392
Pupil, 411
Pupil size, 289
Purcell, E. M., 23, 31, 101, 114, 115, 213, 218, 539
Purkinje fibers, 194, 282
PUVA, 404
Pyramidal cells, 233
Pyroelectric crystal, 399

Q_{10}, 79, 164, 165
QRS wave, 195, 203
Quadrupole
  magnetic, 535
Quality factor, 490
Quantum mottle, 475
Quantum number, 57
Quantum numbers
  atomic electrons, 383
  rotational, 385
  vibrational, 386
Quasiperiodicity, 286
Quatrefoil pattern, 222
Québec, 495
Quinine water, 402

R project, 307
RADAR (Radiation Group Assessment Resource), 511, 515
Radial isochron clock, 3, 281, 296
Radian, 567
Radiance, 406–408
Radiant
  energy, 406, 407
  exposure, 406
  flux, 406
  intensity, 406, 407
  power, 406, 407
Radiation
  biological effects, 480
  electromagnetic, 257
  natural background, 492
  risk, 490
  weighting factor, 490
Radiation chemical yield, 463
Radiation damage
  type A, 482, 484
  type B, 482, 484
Radiation equilibrium, 452
Radiation risk model
  hormesis, 494
  linear no-threshold, 493
  threshold, 494
Radiation therapy, 486
  conformal, 487
  electron, 488
  intensity-modulated, 487
Radiation yield, 448, 461, 462
Radiative transition, 435
Radiobiology, 480
Radiochromic film, 469
Radiograph, 470
Radiographic signal, 475
Radioimmunoassay, 503
Radioimmunotherapy, 524
Radiometry, 405
Radiopharmaceuticals, 517
Radiotherapy
  internal, 524
^{226}Ra, 524
Radius
  atomic, 504
  nuclear, 504
Radon, 492, 495
^{222}Rn, 524, 525
Radon transformation, 353
Railroad tracks, 46
Raizner, A. E., 524, 533
Ramachandran, G. N., 357, 362
Raman scattering, 393
Raman spectroscopy, 399
Raman, C. V., 393
Random walk, 106
Range, 439, 447
Ranvier
  node of, 142
Rao, D. V., 533
Ratchet and pawl, 337
Ratliff, S. T., 485
Rattay, F., 209
Rayleigh, 296
Rayleigh scattering, 431
RBE, 490
Reactance, 367
Reaction-diffusion process, 111

Receptor
mechanical, 142
stretch, 141
temperature, 141
Receptor field, 413
Reconstruction from projections
filtered back projections, 352
Fourier transform, 351
Recruitment
physiological, 209
Rectifier channels, 248
Red (color), 415
Red blood cell, 2, 25, 123
Reddy, A. K. N., 247
Reduced mass, 385
Reduced scattering coefficient, 389
Reentrant circuit, 194, 204, 207
Reference action spectrum, 401
Reflection, 411
total internal, 421
Reflection at a boundary, 367
Reflection coefficient, 124, 125, 367
pore model, 133
Refraction, 411
Refractory period, 165, 181, 194
Regression
linear, 304
polynomial, 306
Rehm, K., 523, 524, 533
Reich, P., 50, 51
Reif, F., 54, 57, 59, 60, 68, 72, 84, 90, 115, 385, 587
Reinisch W., 422
Relative biological effectiveness, 490
Relative risk, 257
Relativity
special, *see* Special relativity
Relaxation oscillator, 194
Relaxation time
experimental, 544
longitudinal, 538, 543
non-recoverable, 544
spin-lattice, *see* Relaxation time, longitudinal
spin-spin, *see* Relaxation time, transverse
thermal, 405
transverse, 538, 544
Rem (unit), 490
Remanent magnetic field, 227
Renal arteries, 479
Renal dialysis, 123
Renal tubules, 128
Repeller, 276
Repolarization, 142, 194
Reservoir, 62, 70
Residence time, 512
Residuals, 304
Resistance, 151
parallel and series, 154
vascular, 21
Resistive pulse technique, 177
Resnick, R., 227, 230, 383, 386, 387, 398, 505, 508, 509, 532
Resolution
and spatial frequency, 349
Resonance, 584
Respiration of glucose, 72
Response
impulse, 345
Rest energy, 504
Rest mass, 429
Restenosis, 524
Resting potential, 142
atrial and ventricular cells, 193
Restitution, 298
Restricted linear collision stopping power, 447
Retina, 403, 411
Reverberation echo, 378
Reversal potential, 249, 250, 264
Reverse osmosis, 125
Reversible process, 69
Reynolds number, 22, 30, 49
Rheobase, 179, 202
Rieke, F., 415, 422
Rieke, V., 379
Rigel D., 422
Riggs, D. S., 42, 51, 270, 293, 295
Right ventricular hypertrophy, 198
Rinaldo, A., 51
Ripplinger, C. M., 205
Risk
excess, 493
models, 493
relative, 257
Ritenour, E. R., 363, 373, 376, 379, 471, 474, 477, 479
Roberson, P., 532
Robinson, D. K., 307
Robitaille, P.-M., 545
Rodieck, R. W., 410, 413, 423
Rodriguez, J., 352, 362
Rods (retinal), 409, 413
Roentgen (unit), 464, 497
Roentgen, W., 465
Rohs, R., 243
Roll-off, 331
Rollins, A. M., 394
Romberg integration, 191, 220
Romer, R. H., 46, 225
Rook A. H., 422
Rose, A., 415, 423
Rosenbaum, D. S., 205
Rossi, S., 225
Rossi-Fanelli, A., 81
Rotating coordinate system, 539
Rotation matrices, 540, 560
Roth, B. J., 201, 205, 210, 219, 223, 226, 229, 232–234, 236, 418, 423
Rottenberg, D. A., 533
Round window, 370
Rowlands, J. A., 470
Rudy, Y., 165, 193, 200
Ruohonen, J., 226
Rushton, W. A. H., 160, 167, 179
Ruska, E., 383
Ruth, T. J., 503, 533
Rutherford, E., 527
RVH, *see* Right ventricular hypertrophy
Rydberg constant, 417
Ryman, J. C., 532
SA node, 172, 194, 203
Sacks, O., 370, 379
Saedi, N., 375, 379
Safety in MRI, 555
Saint-Jalmes, H., 393, 423
Sakitt, B., 415, 423
Sakmann, B., 251
Salmelin, R., 224
Saltatory conduction, 167
Samuels S., 115
Santini, L., 555
Sap flow, 29
SAR, *see* Specific absorption rate
Sarcoplasmic reticulum, 111
Sarvas, J., 233
Sastry, K. S. R., 507
Sato, S., 236
Saturation, 227
Savage, V. M., 51
Savannah sparrow, 229
Scalar, 570
Scalar product, 12
Scaling, 43
Scattering
coherent, 427, 431
Compton, 427, 428
incoherent, 427, 431
inelastic, 427
nuclear, 445
Raman, 427
Rayleigh, 431
Thomson, 430
Scattering coefficient
reduced, 389
Scattering cross section, *see* Cross section
Schaefer, D. J., 555
Schaefer-Prokop, C., 470
Schaper, K. A., 533
Scher, A. M., 31, 294
Scherr, P., 91, 115
Schey, H. M., 88, 93, 102, 115, 144, 601
Schiff, G., 379
Schlichting, H., 23, 31
Schlomka, J. P., 472
Schmidt, C. W., 403, 423
Schmidt-Nielsen, K., 50, 51, 128, 138
Schmitt, F., 553
Schmitt, J. M., 392, 423
Schmitt, O. H., 258
Schrödinger equation, 53
Schroeder, D. V., 54, 84, 385, 396, 423
Schroter, S. T., 30
Schueler, B. A., 490, 491
Schultz, J. S., 134
Schulz, R. J., 485
Schuman, J. S., 422
Schwan, H. P., 258
Schwartz, D., 76, 77
Schwarz T., 422
Schwiegerling, J., 422
Scintillation camera, 520
Scintillation detector, 466, 520
Scotopic vision, 409, 413
Scott, A. C., 156, 161
Scott, G. C., 233
Scott, R. L., 72, 83, 91, 115, 121

Screen, intensifying, 465
Screening
  atomic electron (charge), 426
SCUBA, 15
Second law of thermodynamics, 75
Sedimentation velocity, 27
Seed, W. A., 30
Seibert, J. A., 470
Selection rules, 384–386, 461
Self-diffusion, 94
Self-similarity, 286
Seltzer, S. M., 438, 472
Semicircular canals, 369
Semiconductor detector, 469
Semiempirical mass formula, 508
Semilog paper, 36
Semipermeable membrane, 117
Separation, 23
Sequestration, 519
Serruys, P. W., 423
Serway, R. A., 7, 148, 423
Setton, L. A., 30
Sevick, E. M., 391, 423
Seymour, R. S., 45, 51
Sgouros, G., 532
Shadowitz, A., 218
Shafer, K. E., 422
Shaker channels, 252
Shani, G., 469, 470
Shapiro, E. M., 263
Shapiro, L., 101, 115
Shark
  and earth's magnetic field, 231
Shaw, C. D., 277
Shaw, R., 349, 362
Shear modulus, 13
Shear rate, 16
Shear strain, 13
Shear stress, 13
Shear wave, 365
Sheehan, J., 379
Shells
  atomic, 426
Sheppard, A. R., 261
Sherwood number, 114
Shlaer, S., 413, 414, 422
Shot noise, 254, 332
Shrier, A., 291
SI system, 1, 226
Sick sinus syndrome, 203, 209
Sidtis, J. J., 533
Siegel, J. A., 524, 533
Siemens (unit), 151
Sievert (unit), 490, 491
Signal, 80, 326
  and noise, 327
  radiographic, 475
Signal averaging, 328
Signal-to-noise ratio, 476
Sigworth, F. J., 252–255
Silicon, 469
Silver halide, 465
Silver, H. K., 44, 51
Silverstein, M. E., 125
Sinc function, 549
Singh, M., 362, 548
Single-channel recording, 251
Single-photon emission computed tomography, 521
Sink, 276
Sinoatrial node, *see* see SA node
Sinogram, 353
Sinus exit block, 209
Skin depth, 234
Skofronick, J. G., 369
Slice selection, 548
Slichter, C. P., 538, 543
Slide projector, 420
Smith, M. A., 422
Smye, S. W., 394, 423
Smythe, W. R., 174
Snell's law, 411
Snell, J., 379
Sneyd, J., 270, 291
Snow blindness, 403
Snyder, W. S., 511, 515, 516, 533
Sobol, W. T., 471
Söderberg, P. G., 403, 421
Sodium
  -potassium pump, 155
  Nernst potential, 162
^{24}Na, 509, 510
Sodium gate, 164
Sodium spectral line, 394
Soffer, B. H., 410, 423
Solar constant, 400, 419
Soleus, 6, 7
Solid angle, 189, 389, 568
  defined, 567
  theorem, 207
Solute, 72, 73
  permeability, 125
Solute permeability
  pore model, 134
Solute transport, 125, 128
  pore model, 133
Solution
  ideal, 121
Solvent, 72
Solvent drag, 89, 125
  in an electric field, 247
Sonoda, E., 51
Sorenson, J. A., 517, 520, 522, 523, 531, 532
Sorgen, P. L., 193
Sound level measurement
  weighting, 369
Source, 276
Space clamp, 161
Space invariance, 347
Spano, M. L., 291
Sparks, R. B., 511, 515, 533
Spatial frequency, 346, 349
  and field of view, 349
  and resolution, 349
Special relativity, 213, 231, 234, 429, 504
Specific absorbed fraction, 511
Specific absorption rate (SAR), 555, 561
Specific heat
  water, 80
Specific heat capacity, 65, 404
Specific metabolic rate, 50
SPECT, 521
Spector, R., 51
Spectral efficiency function, 409
Spectroscopy
  infrared, 391
Spectrum
  beta decay, 509
  Wiener, 475
Spherical aberration, 412
Spherical coordinates, 601
Spherocytosis, 123
Spheroid degeneration of the eye, 403
Sphincter, pre-capillary, 21
Sphygmomanometer, 20
Spider's thread, 26
Spin
  electron, 227, 384
  nuclear, 536
Spin echo (SE), 546, 548
  fast or turbo, 553
Spin quantum number, 425
Spin warp encoding, 548
Spiral CT, 478
Spiral wave, 195, 291, 299
Spoiler gradient, 562
Spontaneous births, 314
Spontaneous emission, 542
Spring constant, 583
Squamous cell carcinoma (SCC), 402, 403
SQUID, 230, 236
Squid, 2
  giant axon, 161
SRIM, 439, 446, 448
Srinivas S. M., 422
Srinivasan, R., 205
Stabin, M. G., 511, 515, 524, 533
Stahlhofen, W., 228
Standard deviation, 586
Standard deviation of the mean, 586
Standing wave, 366
Stanfield J., 422
Stanley, P. C., 200
Stapes, 369
Stark, L., 270, 290, 413, 423
State space, 277
Stationarity, 347
Stationary random process, 327
Stationary system, 346
Statistical mechanics, 54
  postulates, 59
Staton, D. J., 222, 223
Steady state and equilibrium
  distinction, 59
Steel, G. G., 482
Stefan-Boltzmann law, 396
Stegun, I. A., 209, 263, 359, 361, 362
Stehling, M. K., 553
Steiner, R. F., 31
Steinmuller, D. R., 136
Steketee, J., 398, 423
Stenosis, 399
Stent, 524
  drug-eluting, 524
Steradian, 568
Stereocilia, 256
Steric factor, 125, 133
Stern, R. S., 404, 423

Steroid-eluting electrode, 204
Stewart, W. E., 94
Stimulated emission, 417, 542
Stimulus
  brain, 205
  current density, 172
  strength-duration curve, 179
  strength-interval curve, 181
Stinson, W. G., 422
Stirling's approximation, 591
Stochastic resonance, 337
Stocker, H., 189
Stokes' law, 23, 91, 94, 260
Stopping cross section, 439
Stopping interaction strength, 444
Stopping number
  per atomic electron, 444
Stopping power, 439
  electronic, 440
  for compounds, 446
  for electrons and positrons, 446
  mass, 439
  nuclear, 440
  radiative, 440
Storment, C. W., 142
Straggling, 447
Straight-line fit, 303, 304
Strain, 364
  normal, 12
  shear, 13
Strasburger, J. F., 222
Stratum corneum, 401
Stratum granulosum, 401
Streamline, 15, 18
Strength-duration curve, 179, 202
Strength-interval curve, 181
Stress, 12, 364
  normal, 12
  shear, 13
Strogatz, S. H., 276, 277, 280, 286
Stroink, G., 222, 225, 228
^{90}Sr, 524
Strother, S. C., 523, 533
Structure mottle, 475
Subchoroidal neovascular membrane, 394
Subexcitation, 464
Subshells
  atomic, 426
Suess, C., 499
Sulfides, 228
Sulfonamides, 402
Sulfonureas, 402
Sun protection factor (SPF), 403
Superconducting quantum interference device, *see* SQUID
Superconductor, 230
Superparamagnetic particles, 236
Superposition, 345
Surface tension, 69, 82
Surface-to-volume ratio, 200
Surfactant, 82
Surrogate data, 336, 337
Surroundings, 56
Survival curve, 480
Susceptibility
  electric, 149
Sutoh, Y., 423
Svassand, L. O., 422
Svedberg (unit), 28
Swanson, E. A., 422
Sweating, 293
Swihart, J. C., 249
Swinney, K. R., 219, 220
Sympathetic nerves, 194
Synapse, 110, 141, 143
Syncope, 203
Syncytium, 193
Synolakis, C. E., 19, 31
Synthesis, 481
System, 56
  linear, 345
  stationary, 346
System dynamics, 280
Systole, 19, 20
Szabo, A., 101, 115
Szabo, Z., 533

T rays, 394
T wave, 195
T-cell lymphoma, 404
T_1-weighted image, 554
T_2-weighted image, 554
Tachycardia
  ventricular, 195
Tack, D., 480
Tai, C., 209
Tait C., 422
Tait, C., 403
Takano, M., 422
Takata, M., 51
Takeda, S., 51
Talus, 6, 7
Tan, G. A., 222
Tanelian, D. L., 142
Tang, Y-h., 111, 115
Target entity, 388
Taylor's series, 575
  of exponential function, 576
Taylor, A. E., 121
Tearney, G. J., 422, 423
^{99m}Tc
  diphosphonate, 521
  pyrophosphate, 518
  sulfur colloid, 518
  tetrofosmin, 522
^{99m}Tc, 35, 507, 519, 522
  albumin, 519
  diphosphonate, 521
  generator, 519
  pertechnetate, 519
  red blood cells, 521
  sestamibi, 520
  sulfur colloid, 512, 529
  tetrofosmin, 520
Tectorial membrane, 370
Telegrapher's equation, 159
Temperature
  absolute, 61, 62
  centigrade or celsius, 62
  negative, 78
Temperature regulation, 293
"10-20" system for EEG leads, 205
Tendon
  Achilles, 6, 7
Tenforde, T. S., 257
Tensile strength, 13
Terahertz radiation, 382, 394
Teramura T., 423
Teramura, T., 403
Tesla (unit), 213
Tetley, I. J., 109
Tetrodotoxin, 251
Teukolsky, S. A., 84, 115, 362
TFT, 470
Thalamotomy, 379
^{201}Tl, 518, 520, 522
Theodoridis, G. C., 413, 423
Theriot J. A., 115
Theriot, J. A., 85
Thermal conductivity, 109, 404
Thermal equilibrium, 60, 61
Thermal noise
  retinal, 415
Thermal penetration depth, 405
Thermal relaxation time, 405
Thermodynamic identity, 69
Thermodynamic relationship, general, 69
Thermodynamics
  first law, 57
  second law, 75
Thermography, 399
Thermoluminescent dosimeter, 469
Thiazide diuretics, 402
Thin-film transistor, 470, 474
Thin-lens equation, 411
Thomas, D. L., 558
Thomas, L., 126
Thomas, S. R., 532
Thompson, D. J., 281
Thompson, J. H., 589
Thomson scattering, 430
Thomson, R. B., 589
^{232}Th, 532
Three-dimensional conformal radiation therapy, 487
Threshold detection, 337
Threshold stimulus strength, 181
Thyroid, 270
  stimulating hormone (TSH), 270, 292
Thyroid hormone, 293
  Thyroxine (T4), 270, 292
  Tri-iodothyronine (T3), 270
Tibia, 6, 7
Time gain compensation, 378
Tissue weighting factor, 491
Tittel, F. K., 422
TMS, 225
Tobacco, photosynthesis, 314
Tomography
  computed, 478
  optical coherence, 392
Tooby, P. F., 27, 31
Torque, 5, 535
  on rotating sphere, 260
Torr (unit), 14, 20
Total angular momentum quantum number, 425

Total internal reflection, 421
Total linear attenuation coefficient, *see* Attenuation coefficient
Toy model, 3
Trace of a matrix, 110
Transcranial magnetic stimulation, 225
Transducer, 141
  piezoelectric, 371
Transfer factor, 466
Transfer function, 330
  modulation, 348
  optical, 348
  phase, 348
Transformation, 506, 507
Transient charged-particle equilibrium, 453
Transition, 506, 507
  nonradiation, 435
  radiationless, 435
  radiative, 435
Transmission at a boundary, 367
Transmission coefficient, 367, 377
Transmittance, 465
Transport
  countercurrent, 127
  solute, 125, 128
  volume, 128
Transvenous pacing, 203
Transverse relaxation time
  recoverable and non-recoverable, 544
Transverse wave, 366
Traveling wave, 366
Trayanova, N., 192, 200, 205
Tree, 29
Trichromate vision, 415
Tricuspid valve, 193
Trigonometric integrals, 579
Triplet production, 432
Trochanter, greater, 8
Tromberg, B. J., 422
Trontelj, Z., 223
Trowbridge, E. A., 24, 31
Tsay, T.-T., 422
TTX, 251
Tubiana, M., 494
Tucker, R. D., 258
Tukey, J. W., 314, 315
Tumor eradication, 485
Tuna, 229, 235
Tung, C. J., 448
Turbidity, 389
Turbulence, 22
Turner, R., 553
Tympanic chamber, 370

Uehara, M., 239
Uffmann, M., 470
Ugurbil, K., 555–557
Ultracentrifuge, 79
Ultrafiltration, 124
Ultrasound
  diagnostic, 371
  intravascular, 399
Ultraviolet spectrum, 400
Ungar, I. J., 209
Unified mass unit, 505
Units of "power" and "energy", 329
Upton, A. C., 494
^{235}U, 519, 532
^{238}U, 524, 525
Urea, 123
UVA, 400, 402, 403
UVB, 400, 403, 404
UVC, 400, 403

van den Bongard, H. J., 379
van den Bosch, M. A., 379
van der Pol oscillator, 336
van der Steen, A. F. W., 379
van Eijk, C. W. E., 468
van Ginneken, B., 470
van Langenhove, G., 423
van't Hoff's law, 121, 124
Vapor pressure, 79, 83
Variable rate of growth or decay, 38
Variables
  extensive, 69
    electric charge, 69
    Length, 69
    Surface area, 69
    volume, 69
Variance, 586
Vascular resistance, 21
Vasodilation, 122
Vasopressin, 294
Vecchia, P., 261
Vector, 4, 569
  addition, 569
  components, 570
  unit, 570
Vector operators in different coordinate systems, 601
Vector product, 6
Velocity
  average, 570
  instantaneous, 570
  root-mean-square, 89
Velocity gradient, 16
Vena cava, 21
Ventilation rate, 270
Ventricle, 193
Ventricular fibrillation, 204, 291, 292, 298
Ventricular tachycardia, 195, 204, 291
Venules, 21
Vergence, 411
Verheye, S., 399, 423
Versluis, M., 379
Vestibular chamber, 370
Vetterling, W. T., 84, 115, 362
Villars, F. M. H., 11, 15, 30, 95, 105, 115
Virtual cathode, 203
  "dog-bone", 205
Virtual image, 420
Virus, 2
Viscoelasticity, 13
Viscosity, 13, 16, 91
  of water, 94
Viscous torque on a rotating sphere, 260
Visscher, P. B., 315
Visual cortex, 555, 557
Visual evoked response, 328
Vitreous, 411
Vlaardingerbroek, M. T., 548, 553
Vogel R. I., 422
Vogel, S. V., 16, 20, 23, 27, 30, 31
Vollrath, F., 26, 31
Volt (unit), 148
Voltage clamp, 161
Voltage difference, 147
Voltage divider, 154
Voltage source
  ideal, 332
Voltage-sensitive dye, 205
Volume transport, 128
  pore model, 133
  through a membrane, 123
Voorhees, C. R., 202, 203
Vreugdenburg T. D., 423
Vreugdenburg, T. D., 399

Wagner, H. N., 467, 533
Wagner, J., 111, 115
Wagner, R. F., 474, 476
Walcott, C. J., 229
Walker, G. C., 422
Walker, M. M., 229
Wang, D., 101, 115
Wang, H., 49, 51
Wang, L., 422
Wang, M. D., 114
Warburg equation, 113
Warloe, T., 422
Warner, G. G., 516, 533
Warshaw E. M., 422
Warwick, W., 589
Washout (in MRI), 555
Wassermann, E., 226
Watanabe, A., 190, 191, 220
Water, 3
  acoustic impedance, 367
  compressibility, 366
  density, 366
  dielectric constant, 151
  speed of sound, 366
  transmission vs. wavelength, 410
Water molecule
  dipole moment, 245
Waterstram-Rich, K. M., 523, 532
Watson, E. E., 511, 512, 514, 516, 517, 524, 532, 533
Watson, S. B., 533
Watt (unit), 11, 153
Wave
  longitudinal, 363
  plane, 363
  standing, 366
  traveling, 366
Wave equation, 181, 363, 364
Wave number, 346, 366
Weaver, J. C., 256, 257, 259, 260, 265, 337
Weaver, W., 54, 84
Weaver, W. D., 76
Webb, W. W., 256
Weber (unit), 224
Webster, J. G., 177, 392, 422, 423
Weighting factor

radiation, 490
Weinstock M. A., 422
Weiss, G., 202
Weiss, J. N., 291
Weiss-Fogh, T., 113, 115
Welder's flash, 403
Wells, P. N. T., 371, 379
Werner, B., 379
Wessels, B. W., 524, 532
West, G. B., 45, 51
Westfall, R. J., 532
Westphalen, A. C., 379
White noise, 333
White, C. R., 45, 51
White, P. D., 39, 51
White, R. J., 123
White-blood-cell count, 290
Wick, G. L., 27, 31
Wieben, O., 392, 423
Wiener spectrum, 475
Wiener theorem for random signals, 328
Wiesenfeld, K., 337
Wijesinghe, R., 229
Wikelski, M., 229
Wikswo, J. P., 53, 84, 205, 219, 220, 222–224, 228–230, 232, 236, 279
Wilders, R., 165, 172
Wiley, J. D., 177
Williams, C. R., 497
Williams, C. S., 348, 362
Williams, L. E., 503
Williams, M., 7, 8, 31
Williamson, S. J., 224
Willis C. D., 423
Wilson, B. C., 399, 422, 423
Winfree, A. T., 282–284, 291, 292, 298
Wintermark, M., 379
Witkowski, F. X., 292
Woods, M. C., 233
Woods, R. P., 533
Woosley, J. K., 219, 236
Work, 11, 58
concentration, 67
pressure-volume, 19, 21
Working level, 525
month, 526
Worthington, C. R., 171

X radiation, 382
X-ray absorption edge, 428
X-ray spectrum
bremsstrahlung, 461
characteristic, 461
thick target, 463
thin target, 462
X-ray tube, 470
Xenon, 499
^{129}Xe, 558, 559
^{133}Xe, 522
Xia, Y., 560
Xu, X. G., 489
Xu, Y., 235
Xylem, 29

Yaffe, M. J., 477
Yamashita, Y., 392, 423
Yasuno, Y., 393, 423
Yatagai, T., 423
Yazdanfar, S., 393, 394
Yehia, A. R., 298
Yi, M., 244
Ying, W., 201
Yodh, A., 389, 392, 422, 423
Young's modulus, 13, 364
Young, A. C., 294
Young, A. R., 401, 422
Young, T., 381
Yu, C. X., 487

Zadicario, E., 379
Zalewski, E. F., 405, 410, 423
Zanzonico, P. B., 520, 523, 532, 533
Zaret, B. L., 222
Zeng, Z-C., 51
Zeuthen, T., 118
Zhang, X-C., 394, 423
Zhang, Y., 422
Zhong, J., 557
Zhou, X.-H., 553
Zhu, T. C., 399, 423
Ziegler, J. F., 439, 442, 445, 446, 448
Zinovev, N. N., 422
Zipes, D. P., 223
Zumoff, B., 38, 51
Zwanzig, R., 101, 115