Mathematical Tools for Physicists

Edited by
George L. Trigg

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Mathematical Tools for Physicists

Edited by

George L. Trigg

WILEY-VCH

WILEY-VCH Verlag GmbH & Co. KGaA

Editor:

Dr. George L. Trigg
New Paltz, New York, USA

Editorial Advisor:

Professor Stuart P. Smith
California State University, Hayward (CSUH), USA

This edition is based on Review Articles originally written for the "Encyclopedia of Applied Physics", edited by George L. Trigg, 25 Volumes, VCH Publishers/Wiley-VCH, New York/Weinheim, 1991–2000.

Cover Picture

Sound pressure plot of the acoustic waves in water around an aluminum cylinder. Arrow and deformation plot shows the deformation of the cylinder. Model computed by using the FEMLAB® Structural Mechanics Module. © COMSOL AB, Stockholm, Sweden. Printed with kind permission of COMSOL AB, www.comsol.com.

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.ddb.de.

©WILEY-VCH Verlag GmbH & Co. KGaA Weinheim, 2005

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such are not to be considered unprotected by law.

Printed in the Federal Republic of Germany.
Printed on acid-free paper.

Composition: Laserwords Private Ltd., Chennai, India
Printing: betz-druck gmbh, Darmstadt
Bookbinding: J. Schäffer GmbH, Grünstadt
ISBN-13: 978-3-527-40548-0
ISBN-10: 3-527-40548-8

Contents

Preface *vii*

List of Contributors *ix*

Algebraic Methods *1*
A. J. Coleman

Analytic Methods *33*
Charlie Harper

Fourier and Other Mathematical Transforms *83*
Ronald N. Bracewell

Fractal Geometry *109*
Paul Meakin

Geometrical Methods *127*
V. Alan Kostelecký

Green's Functions *159*
Kazuo Ohtaka

Group Theory *189*
M. Hamermesh

Mathematical Modeling *213*
Kenneth Hartt

Monte-Carlo Methods *249*
K. Binder

Numerical Methods *281*
Christina C. Christara and Kenneth R. Jackson

Perturbation Methods *385*
James Murdock

Quantum Computation *417*
Samuel L. Braunstein

Quantum Logic *439*
David J. Foulis, Richard J. Greechie, Maria Louisa Dalla Chiara, and Roberto Giuntini

Special Functions *475*
Charlie Harper

Stochastic Processes *513*
Melvin Lax

Symmetries and Conservation Laws *565*
Gino Segrè

Topology *587*
S. P. Smith

Variational Methods *619*
G. W. F. Drake

Index *657*

Preface

Mathematics is a central structure in our knowledge. The rigor of mathematical proof places the subject in a very special position with enormous prestige. For the potential user of mathematics this has both advantages as well as disadvantages. On the one hand, one can use mathematics with confidence that in general the concepts, definitions, procedures, and theorems have been thoroughly examined and tested, but the sheer amount of mathematics is often very intimidating to the non-expert. Since the results of mathematics once proved stay in the structure forever, the subject just gets larger and larger, and we do not have the luxury of discarding older theories as obsolete. So the quadratic formula and the Pythagorean theorem are still useful and valid even though they are thousands of years old. Euclid's Elements is still used as a text in some classrooms today, and it continues to inspire readers as it did in the past although it treats the mathematics from the time of Plato over 2300 years ago.

Despite the prestige of mathematical proof, most mathematics that we use today arose without proof. The history of the development of calculus is a good example. Neither Newton nor Leibniz gave definitions of limits, derivatives, or integrals that would meet current standards. Even the real number system was not rigorously treated until the second half of the nineteenth century. In the past, as in modern times, large parts of mathematics were initiated and developed by scientists and engineers. The distinction between mathematicians and scientists was often rather vague. Consider for example, Newton, Euler, Lagrange, Gauss, Fourier, and Riemann. Although these men did important work in mathematics, they were also deeply involved in the sciences of their times. Toward the end of the nineteenth century a splitting occurred between mathematics and the sciences. Some see it in the development of non-Euclidean geometry and especially axiomatic methods reminiscent of Euclid.

At this time mathematics appeared to be taking its own path independent of the sciences. Here are two cases that participated in this division. In the late nineteenth century Oliver Heaviside developed the Operational Calculus to solve problems in electrical engineering. Although this calculus gave solutions in agreement with experiment, the mathematicians of Heaviside's time could not justify or condone his procedures. Physicists also found the Dirac delta function and Green's functions extremely useful and developed an appropriate calculus for their use, but the underlying mathematical theory was not available. It was not until the early 1950's that Laurent Schwartz was able to give a rigorous mathematical foundation for these methods with

his Theory of Distributions. Also, early in the twentieth century the relatively new subject of Group Theory was seen as being of use in applications to chemistry and physics, but the few texts available at the time were written in a rather abstract and rigorous mathematical style that was not easily accessible to most non-mathematicians. The subject was quickly labeled the "Gruppenpest" and ignored by many researchers. Needless to say, today group theory with its applications to symmetry is a fundamental tool in science.

With the complexity of each field in science and engineering growing so rapidly, a researcher in these fields has little time to study mathematics for its own sake. Each field has more material than can possibly be covered in a typical undergraduate program, and even graduate students must quickly pick a sub-area of specialization. Often, however, there is a sense that if we just knew more mathematics of the right sort, we could get a better grasp of the subject at hand. So, if we are still in school, we may take a mathematics course, or if not in school, we may look at some mathematical texts. Here some questions arise: which course should we take, do we have the correct prerequisites, what if our mathematics instructor has no knowledge of our field or any applications that we are interested in, are we really in the right course? Furthermore, most texts in mathematics are intended for classroom use. They are generally very proof oriented, and although many now include some historical remarks and have a more user friendly tone, they may not get to the point fast enough for the reader outside of a classroom.

This book is intended to help students and researchers with this problem. The eighteen articles included here cover a very wide range of topics in mathematics in a compact, user oriented way. These articles originally appeared in the Encyclopedia of Applied Physics, a magnificent twenty-three volume set edited by George L. Trigg, with associate editors Eduardo S. Vera and Walter Greulich and managing editor Edmund H. Immergut. The full Encyclopedia was published in the 1990's by VCH, a subsidiary of John Wiley & Sons, New York. Each article in this volume covers a part of mathematics especially relevant to applications in science and engineering. The articles are designed to give a good overview of the subject in a relatively short space with indications of applications in applied physics. Suggestions for further reading are provided with extensive bibliographies and glossaries. Most importantly, these articles are accessible. Each article seeks to give a quick review of a large area within mathematics without lapsing into vagueness or overspecialization.

Of course not all of mathematics can be covered in this volume: choices must be made in order to keep the size of the work within bounds. We can only proceed based on those areas that have been most useful in the past. It is certainly possible that your favorite question is not discussed here, and certainly the future will bring new mathematics and applications to prominence, but we sincerely expect that the articles in this volume will be valuable to most readers.

Stuart P. Smith
CSUH – January 2005

List of Contributors

K. Binder
Institut für Physik,
Johannes-Gutenberg-Universität Mainz,
Mainz,
Germany

Ronald N. Bracewell
Electrical Engineering Department,
Stanford University,
Stanford, California,
USA

Samuel L. Braunstein
SEECS,
University of Wales,
Bangor,
UK

Christina C. Christara
Computer Science Department,
University of Toronto,
Toronto, Ontario,
Canada

A. J. Coleman
Department of Math and Statistics,
Queens University,
Kingston, Ontario,
Canada

G. W. F. Drake
Department of Physics,
University of Windsor,
Windsor, Ontario,
Canada

David J. Foulis
University of Massachusetts,
Amherst, Massachusetts,
USA

Roberto Giuntini
Università di Firenze,
Florence,
Italy

Richard J. Greechie
Louisiana Tech University,
Ruston, Louisiana,
USA

M. Hamermesh
School of Physics and Astronomy,
University of Minnesota,
Minneapolis, Minnesota,
USA

Charlie Harper
Department of Physics,
California State University,
Hayward, California,
USA

Kenneth Hartt
Physics Department,
University of Rhode Island,
Kingston, Rhode Island,
USA

Kenneth R. Jackson
Computer Science Department,
University of Toronto,
Toronto, Ontario,
Canada

V. Alan Kostelecký
Physics Department,
Indiana University,
Bloomington, Indiana,
USA

Melvin Lax
Physics Department,
City College of the City
University of New York,
New York,
USA

and

Bell Laboratories,
Lucent Technologies,
Murray Hill, New Jersey,
USA

Maria Louisa Dalla Chiara
Università di Firenze,
Florence,
Italy

Paul Meakin
Department of Physics,
University of Oslo,
Oslo,
Norway

James Murdock
Iowa State University,
Ames,
USA

Kazuo Ohtaka
Laboratory of Applied Physics,
Faculty of Engineering,
Chiba University, Chiba-shi,
Japan

Gino Segrè
Department of Physics,
University of Pennsylvania,
Philadelphia, Pennsylvania,
USA

S. P. Smith
Department of Mathematics and
Computer Science,
California State University,
Hayward, California,
USA

Algebraic Methods

A. J. Coleman
Department of Math and Statistics, Queens University, Kingston, Ontario, Canada

Introduction *2*
1 Groups *2*
2 Fields *3*
2.1 The Characteristic of $\mathbb{F}$ *4*
2.2 Algebraically Closed Fields *4*
2.3 Rational Functions *5*
3 Linear Spaces *5*
3.1 Independence of Vectors *6*
3.2 Change of Basis *6*
3.3 Linear Maps and Their Associated Matrices *7*
3.4 Determinants *8*
3.5 Eigenvectors and Eigenvalues *8*
3.6 Canonical Form of Matrices *10*
3.7 Dual Space *12*
3.8 Tensors *12*
4 Creating Algebraic Structures *13*
5 Rings *15*
5.1 Examples of Rings *15*
5.2 Polynomial Rings *18*
5.2.1 Binomial and Multinomial Theorem *19*
5.2.2 Fundamental Theorem of Algebra *19*
6 Algebras *21*
6.1 Examples of Algebras *23*
7 Modules *28*
7.1 Examples of Modules *28*
7.2 Morphisms of Modules *28*
Glossary *30*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

List of Works Cited *31*
Further Reading *31*

Introduction

The use of mathematics by physicists, and in particular of algebra, has increased in a remarkable degree during the last 50 years, both in the amount of space occupied in journal articles and in the type and abstractness of the methods employed.

Following N. Bourbaki, it is now conventional to characterize as algebraic structures those parts of mathematics that employ operations, such as addition, which act on a finite set of objects to produce a unique corresponding object. Such operations are contrasted with ideas like limit in calculus or closure in topology, which associate a number or other mathematical object to an infinite set or sequence. Thus, whereas the passage from (2, 3) to $2 + 3 = 5$ is an algebraic operation, to go from the infinite sequence $n \to 1/n$ (where n is any positive integer) to the limit 0 is a topological operation. The present section is concerned chiefly with algebra.

In this brief article it is impossible to describe all the many algebraic structures which occur in the literature of applied physics. Therefore we have selected those which are absolutely essential for understanding the contemporary literature under the following rubrics: Groups; Fields; Linear Algebra; Rings; Algebras and Modules. As to style, we have attempted to steer a course between that which physicists would have liked 20 years ago and the austerity of contemporary pure mathematicians with which all physicists will be happy 20 years from now. This should leave all readers equally unhappy! Our definitions are seldom painstakingly detailed but rather highlight the essential ideas leaving the reader to use common sense to fill them out. We shall assume that the reader is familiar with elementary properties of vectors and matrices. Recall that a square matrix A is invertible or nonsingular if there is a matrix B such that $AB = BA = I$, where I is the identity matrix. In this case A and B are inverses of each other and we denote B by A^{-1}. Although, logically, rings should be discussed before fields, teaching experience suggests that the reverse order is pedagogically sounder.

NOTATION: We shall adopt the following widely used symbolism: $\mathbb{N}:=$ the natural numbers, $\{1, 2, 3, \ldots\}$; $\mathbb{Z}:=$ the positive and negative integers and zero; $\mathbb{R}:=$ the real numbers; $\mathbb{C}:=$ the complex numbers; $i:=\sqrt{-1}$; $\mathbb{Q}:=$ the rational numbers. We shall employ Einstein's summation convention in the restricted form that in any monomial an index which is repeated as a subscript and as a superscript will be interpreted as summed over its range unless the contrary is explicitly stated.

1
Groups

A group is a set, G, say, together with a binary operation which we temporarily denote by "$*$", which satisfies certain definite rules. A binary operation is one which combines two elements of G to

obtain an element of G. Perhaps our first encounter with a group occurs when as babies we push our blocks around on the floor using the translation group in two dimensions! Later in grade school we learn the properties of the integers. Under addition the integers $\mathbb{Z}$ exemplify the axioms of a group:

(i) A group $(G,*)$ is a set, G, together with an operation, $*$, which to any two elements x and y of G associates an element $z = x * y$ of G. For example, in $(\mathbb{Z}, +)$, $2 + 3 = 5$, $5 + (-3) = 2$. This property is described by saying that G is closed under the binary operation $*$.
However, for the structure $(G,*)$ to be dignified with the title "group," it must satisfy the additional properties:

(ii) The operation is associative, that is for any x, y, z in G, $(x * y) * z = x * (y * z)$.

(iii) There is a unique neutral or identity element, n, such that $x * n = n * x = x$ for all x in G.

(iv) For any element x in G there is a unique element y in G such that $x * y = n$. In this case, x and y are said to be inverses of each other.

Thus while $(\mathbb{N}, +)$ satisfies (i) and (ii) it is not a group because (iii) and (iv) fail. However, $(\mathbb{Z}, +)$ is a group when we take $n := 0$.

If G has a finite number of elements, the group is a finite group and the number of elements is called the order of the group. If $x * y = y * x$ for all $x, y \in G$, the group is Abelian or commutative.

The set of symmetries of any mathematical or physical structure constitutes a group under composition of symmetries. Such groups play a major role in physics for analyzing the properties of space-time, understanding crystal structure, and classifying the energy levels of atoms, molecules, and nuclei. Indeed, the role of groups is so important in physics that an article of the *Encyclopedia* is devoted to them. We therefore shall not explicitly pursue the detailed properties of groups further, even though they will occur as substructures in rings and fields.

2 Fields

Whereas a group consists of a set together with a single binary operation, a field consists of a set together with two binary operations linked together by a distributive law. The two operations are usually called addition and multiplication. The familiar fields are the real numbers, $\mathbb{R}$; the complex numbers, $\mathbb{C}$; and the rational numbers, $\mathbb{Q}$. We shall use the symbol $\mathbb{F}$ for an arbitrary field. Strictly speaking, we should employ a notation such as $(\mathbb{F}, +, \times)$ to denote a field; however, the relevant operations are generally obvious from context in which case it is sufficient to use $\mathbb{F}$ alone.

$(\mathbb{F}, +, \times)$ is a field if:

(i) $(\mathbb{F}, +)$ is a commutative or Abelian group. That is, $x + y = y + x$ for any x and y in $\mathbb{F}$.

(ii) The elements of $\mathbb{F}$ other than zero form a group under multiplication.

(iii) Multiplication distributes over addition. That is, if a, b, c, belong to $\mathbb{F}$ then $a \times (b + c) = a \times b + a \times c$, and $(b + c) \times a = b \times a + c \times a$.

These properties are, of course, familiar for the reals, complexes, and rationals, but there are fields, such as the quaternions, for which multiplication is not commutative. There are also fields with only a finite number of elements.

A field always has at least two elements, 0 and 1.

2.1 The Characteristic of $\mathbb{F}$

Since a field is closed under addition, $\mathbb{F}$ contains $1+1$, which cannot equal 1 since this would imply that $1=0$, which we excluded. But $1+1$ might equal 0 in which case $(1+1)+1=1$ and one can easily check that $\mathbb{F}=\{0,1\}$ can serve as the set of a field of two elements. This is the smallest possible field and is both famous and useful since it plays a key role in the design of electric circuits, such as those which occur in computers.

More generally, if p is a prime number, we can obtain a field containing p elements in which the sums of j 1's are numbers which are distinct if $0 \le j < p$ and equal to 0 if $j=p$. When this occurs in any field $\mathbb{F}$ we say that p is the characteristic of $\mathbb{F}$ and that $\mathbb{F}$ has finite characteristic. When there is no such p we say that $\mathbb{F}$ has characteristic zero. A field of characteristic zero has an infinite number of elements. If $\mathbb{F}$ has only a finite number of elements it will contain p^n elements, where p is a prime and n is a positive integer. If $n>1$, $\mathbb{F}$ will contain a subfield of the above type with p elements. The fields with p^n elements are called Galois fields. They are important in coding and communication theory. A finite field is necessarily commutative.

2.2 Algebraically Closed Fields

We know that the square of a real number is positive, so there is no real number x such that $x^2=-1$. In other words, in $\mathbb{R}$ there is no element x satisfying the equation $x^2+1=0$. If $\mathbb{F}$ has the property that for every equation of the form $a_j x^j = 0$, $0 \le j \le n$, where the a_j belong to $\mathbb{F}$, there is an element of $\mathbb{F}$ which satisfies the equation, we say that $\mathbb{F}$ is algebraically closed. Otherwise, it is not algebraically closed. Clearly $\mathbb{R}$ is not algebraically closed. If we assume that there is a "number" i such that $i^2+1=0$, then, as we know, the field containing $\mathbb{R}$ and i is the complex numbers, $\mathbb{C}$. It was proved by Gauss that $\mathbb{C}$ is algebraically closed.

Notice that if σ is a $1:1$ map of $\mathbb{C}$ onto itself, such that $\sigma(x+iy)=x-iy$ for all $x,y \in \mathbb{R}$, then σ preserves all the properties of a field and is therefore an automorphism of $\mathbb{C}$. Recall that an isomorphism of two algebraic structures is a bijective (or one-to-one) correspondence between their sets, which preserves all the relations among their elements, and that an automorphism is an isomorphism of an algebraic structure onto itself. Note that $\sigma(x)=x$ if $x \in \mathbb{R}$ and that $\sigma(i)=-i$, which is the root other than i of the equation $x^2+1=0$. An automorphism of $\mathbb{C}$ must send 0 into 0 and thus must either leave i fixed (and so everything in $\mathbb{C}$ is fixed) or, like σ, send i to $-i$. The set consisting of σ and the identity map is a group of order two under composition of mappings. It is the Galois group of $\mathbb{C}$ over $\mathbb{R}$. Alternatively, it is also called the Galois group of the equation $x^2+1=0$ with respect to the reals. For more detail about fields, their algebraic extensions, and their Galois groups, the reader is referred to Jacobson (1964) or any of the multitude of algebra texts at the same level.

Are there fields containing $\mathbb{R}$ other than $\mathbb{C}$ which are at most finite dimensional over $\mathbb{R}$? The answer was given by Frobenius. There is one and only one, the quaternions, but in this field multiplication is not commutative. We shall see below that the quaternions can be realized

as linear combinations with real coefficients of the Pauli matrices and the 2×2 identity matrix. The significance of the field of quaternions is dramatized by the observation that if it did not exist there would be no spin in physics, therefore no sigma and pi bonds in chemistry, and therefore no life on planet earth if, indeed, there were any stars or planets!

2.3 Rational Functions

If we adjoin a symbol x to any field $\mathbb{F}$ and form all possible sums, differences, products, and quotients involving x and the elements of $\mathbb{F}$, the result is a set which is closed under any finite sequence of these operations and forms a field, denoted by $\mathbb{F}(x)$, which we might describe as the field of rational functions in x over $\mathbb{F}$. There is a subset, denoted by $\mathbb{F}[x]$, of polynomials of the form $a_j x^j$ where j is summed from 0 to some $n \in \mathbb{N}$, where n is arbitrary and the $a_j \in \mathbb{F}$. If a_n is not zero we say that the polynomial has degree n. As we shall remark below, the polynomials constitute a ring. As usual $x^0 := 1$, by definition, so when $n = 0$ the preceding polynomial reduces to a_0. Thus $\mathbb{F}$ is contained in $\mathbb{F}[x]$. The field $\mathbb{F}(x)$ consists of all possible quotients of elements of $\mathbb{F}[x]$.

Suppose that the rational function $R(x) = P(x)/Q(x)$, where P and Q are polynomials. Suppose further that $Q(x) = Q_1(x)Q_2(x)$, where Q_1 and Q_2 are polynomials with no common factor. Since we could have used long division to ensure that R is the sum of a polynomial and a rational function, the numerator of which has degree strictly less than the degree of Q, we may assume that $\deg(P)$ – the degree of P – is less than $\deg(Q)$. It is relatively easy to show that it is then possible to find polynomials P_1 and P_2 with $\deg(P_i) < \deg(Q_i)$ such that

$$\frac{P}{Q} = \frac{P_1}{Q_1} + \frac{P_2}{Q_2}.$$

This is the fundamental theorem of the so-called method of partial fractions, by repeated application of which it follows that any rational function can be expressed as a sum of a polynomial and rational functions whose denominators have no nontrivial factors.

In particular, if $\mathbb{F}$ is algebraically closed (e.g., $\mathbb{F} = \mathbb{C}$), then Q is a product of factors such as $(x - a)^m$, where $a \in \mathbb{F}$. A summand in $R(x)$ of the form $g(x)/(x - a)^m$ with $\deg(g) < m$ can, by Taylor's theorem applied to $g(x)$, be expressed as the sum $c_j(x - a)^{-j}$, where $1 \leq j < m$ and $c_j = g^{(m-j)}(a)/(m - j)!$. Here $g^{(k)}$ is the kth order derivative of g.

The method of partial fractions is quite useful for finding integrals of rational functions. Books on calculus explain helpful tricks for obtaining the partial fraction decomposition of a given rational function.

3 Linear Spaces

The theory of linear space with its related concepts of linear transformation, eigenvector, matrix, determinant, and Jordan canonical form is certainly one of the most important and most useful part of mathematics. The abstract concept of linear space is frequently approached by a long discussion of the problem of solving systems of linear equations. We take a direct approach defining a linear space as consisting of a field $\mathbb{F}$ whose elements are called scalars, a set V, called vectors, and two operations called vector addition and scalar multiplication together

with a set of rules governing the relation among these various entities. The vectors under addition form an additive Abelian group $(V, +)$. Under multiplication by scalars the set V is closed. Thus, $v \in V$ and $a \in \mathbb{F}$ imply that $av \in V$. Another important property is that multiplication by scalars distributes over addition of vectors. That is $a(v_1 + v_2) = av_1 + av_2$ for all $a \in \mathbb{F}$ and $v_i \in V$.

3.1 Independence of Vectors

This seems to be the most difficult idea in teaching elementary courses in linear algebra – possibly, the only difficult idea! Two nonzero vectors v_1 and v_2 are linearly dependent if there are scalars a^1 and a^2, not both zero, such that $a^1 v_1 + a^2 v_2 = 0$, where, of course, by 0 we mean the zero vector. It is clear that neither a^1 nor a^2 is zero, and thus each vector is a scalar multiple of the other. More generally, if, given n vectors v_i, $1 \le i \le n$, there exist scalars a^i such that $a^i v_i = 0$, where all $v_i \neq 0$ and not all $a^i = 0$; then we say that the n vectors are linearly dependent. If no such relation holds, the vectors are linearly independent. For example, for $n \in \mathbb{N}$ there are no numbers a_n other than 0 such that $\Sigma_n a_n \times \cos(n\vartheta) = 0$ for all ϑ. Thus the functions $\vartheta \to \cos(n\vartheta)$ are linearly independent.

If n vectors v_i are such that any vector v can be written as $v = a^i v_i$ for some choice of scalars a^i, we say that the set $\{v_i\}$ spans V. If the v_i are also linearly independent then the coefficients a^i are unique. We then say that $B = \{v_i\}$ is a basis of V, that the linear space V has dimension n, and that a^i are the components of v with respect to B. A basic theorem assures us that the dimension depends only on the space V and not on the choice of basis. If a linear space does not have a finite basis it is infinite dimensional.

3.2 Change of Basis

How do the components of a given vector change if the basis is changed? This was a key question which led to the theory of invariants in the mid-19th century and opened up the development of much of contemporary algebra. It also led to the emergence of the tensor calculus which was essential for Einstein's exposition of General Relativity Theory.

Suppose V is a linear space and that $B = \{v_i\}$ and $B' = \{v'_i\}$ are two different bases for V. Then there is a matrix, P^j_i called the transition matrix from B to B' such that $v'_i = P^j_i v_j$. Thus if the vector $x = x^j v_j = x'^i v'_i = x'^i P^j_i v_j$, it follows that $x^j = P^j_i x'^{\,i}$. If in the usual matrix notation we regard x^j as the jth component of a column vector $\mathbf{x}$, and P^j_i as the element in the jth row and ith column of the transition matrix, $\mathbf{P}$, from the old base B to the new base B', the preceding equation takes the form

$$\mathbf{x} = \mathbf{P}\mathbf{x}' \quad \text{or} \quad \mathbf{x}' = \mathbf{P}^{-1}\mathbf{x}.$$

There is an even more convenient notation. Define $P(B', B) := \mathbf{P}$; then the preceding equations imply that $\mathbf{P}^{-1} = P(B, B')$. Subsequently we shall need the formulas

$$\mathbf{x} = P(B', B)\mathbf{x}' \quad \text{and} \quad \mathbf{x}' = P(B, B')\mathbf{x}.$$

To understand tensor notation it will prove important to note that, whereas $\mathbf{P}$ sends the old to the new basis, it sends the new coordinates to the old ones. This observation underlies duality in homological algebra and the distinction between

covariant and contravariant tensors, which we define below.

3.3 Linear Maps and Their Associated Matrices

Suppose that V and U are linear spaces of dimension n and m with bases $B_v = \{v_i\}$ and $B_u = \{u_j\}$, respectively. A transformation, function, or map from V to U sends each vector $x \in V$ to a vector, say, y of U, which we denote by $Ax := y$. If A has the property that for any two vectors x and $x' \in V$ and arbitrary scalars a and $a' \in \mathbb{F}$, $A(ax + a'x') = aAx + a'Ax'$, we say that A is linear. The condition that a map be linear is very restrictive. Nonetheless, linear maps play a big role in the application of mathematics to physics (as well as statistics, economics, biology, etc.) for the same reason that the derivative is important in analysis. For example, if $f(x)$ is a real-valued function of the real variable x, such that $f(0) = 0$, then $f'(0)x$ is the linear function of x which is the best possible linear approximation to $f(x)$ near 0.

A linear map or transformation, $A: V \to U$, can be completely described by an $m \times n$ matrix A_i^j, such that $Av_i = A_i^j u_j$, which we describe as the matrix associated to the linear transformation or map A with respect to the bases B_u and B_v. It has m rows and n columns. If we denote this matrix by $A(u,v)$ then if the bases in V and U are changed to B'_v and B'_u, respectively,

$$A(u', v') = P(B'_u, B_u)A(u, v)P(B_{v'}, B'_v),$$

where we use the notation of Sec. 3.2. In terms of coordinates, if $\mathbf{y} = \mathbf{Ax}$, then $y^j = A_i^j x^i$, where, as follows from the context, $1 \le j \le m$ and $1 \le i \le n$.

In the particular case that $V = U$ of dimension n, with $B_u = B_v$, $A(u, u)$ is an $n \times n$ matrix which we denote by $A(u)$. We deduce that for a change of basis

$$A(u') = P(B'_u, B_u)A(u)P(B_u, B'_u).$$

We thus associate to any linear transformation a matrix which is unique, *once bases are chosen* for the domain and codomain of the transformation. But conversely, if the bases are given, then there is a unique linear transformation associated with a given matrix of the appropriate shape. Thus there is a bijection (i.e., a one-to-one correspondence) between $m \times n$ matrices with entries in $\mathbb{F}$ and linear maps from a linear space of dimension n into one of dimension m. We have found how the bijection changes when the bases are altered. It is this bijection which gives meaning to the familiar addition and multiplication of matrices.

A linear map between two spaces over the same field $\mathbb{F}$ has the property of preserving the linear structure and is said to be a homomorphism (i.e., a structure-preserving map), so it is common to denote by $\mathrm{Hom}(V_1, V_2)$ the set of all linear maps between linear spaces V_1 and V_2 where both have the same field. If $A \in \mathrm{Hom}(V_1, V_2)$ then V_1 is the domain of A and V_2 is the codomain of A. The kernel of A, frequently denoted by $\ker(A)$, is the set of all elements in the domain which are mapped onto 0 by A. The range of A consists of all elements of the codomain of the form Ax for some x in the domain. Of course these last four definitions are valid for any function, not merely linear maps. However, when A is linear it can be easily proved that both the kernel and the range are linear subspaces of their ambient spaces. This is probably the secret of the power and relative simplicity of the theory of linear spaces. When $V_1 = V_2 = V$, we denote $\mathrm{Hom}(V,V)$ by $\mathrm{Hom}(V)$.

If G is a map from V_1 to V_2 and F one from V_2 to V_3, we denote the composition

of these two maps by FG and, having fixed bases in the spaces, we define the matrix corresponding to FG as the product of the matrices corresponding to F and G. This "explains" the usual rule for matrices that $(FG)_i^j = F_k^j G_i^k$, where i, k, and j range from 1 to the dimensions of V_1, V_2, and V_3, respectively.

The composition (or product) of two maps can be well-defined if the range of the first is in the domain of the second. The sum of two maps is only meaningful if the codomain is an additive group in order for the sum of Fx and Gx to be meaningful. In this case it is possible to let $F + G$ denote the map such that $(F + G)x = Fx + Gx$ for all x in the intersection of the domains of F and G. When the domain and codomain are fixed linear spaces over the same field $\mathbb{F}$ we can do even better and give $\mathrm{Hom}(V_1, V_2)$ the structure of a linear space over $\mathbb{F}$. This implies that the set of all $m \times n$ matrices with entries from $\mathbb{F}$ is a linear space of dimension mn over $\mathbb{F}$.

The dimension of the range of a linear operator is called the rank of the operator and also the rank of any matrix associated with the operator by a particular choice of bases. The dimension of the kernel of a linear transformation is called the nullity of the transformation and of its associated matrices. It follows from this definition that the various matrices obtained from one another by a change of basis all have the same rank and nullity. The rank of a product of operators or matrices is not greater than the minimum rank of its factors.

3.4 Determinants

If $\mathbb{F}$ is a commutative field, to any square matrix, it is possible to assign a number in $\mathbb{F}$ which is expressible as a polynomial in the elements of the matrix and which vanishes only if the matrix is not invertible. To two square matrices which are related as in Sec. 3.3 by a change of basis, we assign the same number, and therefore it is meaningful to also assign this number to the associated linear transformation belonging to $\mathrm{Hom}(V)$. The function, det, from $\mathrm{Hom}(V)$ into $\mathbb{F}$, has the following properties: (i) $\det(AB) = \det(A)\det(B)$; (ii) $\det(fI) = f^n$, where n is the dimension of V, I is the identity map, and f is any element of $\mathbb{F}$. The usual definition of the determinant follows from these properties (MacDuffee, 1943). In particular since, for a fixed basis, the equation $Ax = y$ is equivalent to the system of equations $A_i^j x^i = y^j$, Cramèr's rule implies

$$\det(A)x^i = \det(Y^i),$$

where Y^i is the matrix obtained from (A_i^j) by replacing its ith column by the column vector (y^k) where i, j, k run from 1 to n. Thus if $\det(A) \neq 0$ there is a unique x for every y so A is invertible; whereas if $\det(A) = 0$, there is an x only for particular y satisfying the n conditions $\det(Y^k) = 0$. Thus for a finite dimensional linear space V, $A \in \mathrm{Hom}(V)$ is invertible if and only if $\det(A) \neq 0$.

The theory of $\mathrm{Hom}(V_1, V_2)$ is really equivalent to the theory of systems of linear equations in several variables. This topic occurs in articles of this book devoted to NUMERICAL METHODS and to MATHEMATICAL MODELING and in at least one hundred elementary textbooks; so we shall not pursue it here.

3.5 Eigenvectors and Eigenvalues

If $A \in \mathrm{Hom}(V)$ then for any $x \in V$, $Ax \in V$. In general we shall not expect Ax to

equal x or indeed, even, that Ax be parallel to x. However, in the latter case Ax would be a multiple of x, say, λx. The equation $Ax = \lambda x$ is equivalent to $(\lambda I - A)x = 0$. By the preceding section, if the determinant of $\lambda I - A$ is different from zero, the only possible solution of this equation is $x = 0$, which is of no great interest. When there is a nontrivial solution of this equation it will be somewhat unusual and is called an eigenvector of A and can occur only for special values of λ. Such a value of λ is the eigenvalue of A corresponding to the particular eigenvector x. The eigenvalue, λ, will satisfy the nth degree algebraic equation

$$f(z; A) := \det(zI - A) = 0.$$

The nth degree polynomial $f(z;A)$ is called the characteristic function of A, and the preceding equation is the characteristic equation of A. Any eigenvalue of A satisfies its characteristic equation. For each zero of the characteristic equation there is at least one nontrivial eigenvector.

There is a one-to-one correspondence between the operators in $\mathrm{Hom}(V)$ and the set of $n \times n$ matrices over $\mathbb{F}$, and this set spans a linear space over $\mathbb{F}$ of dimension n^2. If we interpret A^0 as the identity operator, I, it follows that the operators A^k for $0 \le k \le n^2$ are linearly dependent. That is, there are $c_j \in \mathbb{F}$ such that $c_j A^j = 0$, where not all c_j are zero. Thus there exists at least one polynomial, $p(z)$, such that $p(A) = 0$. From the algorithm for long division it easily follows that there is a unique monic polynomial (i.e., a polynomial with highest coefficient 1) of minimal degree with this property. We shall denote this so-called minimal polynomial of A by $m(z;A)$. A famous theorem of Hamilton asserts that A satisfies its characteristic equation. That is, $f(A;A) = 0$. Since $\deg(f) = n$, $\deg[m(z;A)] \le n$. Since $m(z;A)$ divides any polynomial $p(z)$ such that $p(A) = 0$, it follows that $m(z;A)$ divides $f(z;A)$.

The form of $m(z;A)$ provides information about A.

(i) $m(z : A) = z^p$ implies that $A^p = 0$ but that $A^{p-1} \neq 0$. Such an operator is called nilpotent, with nilpotency index p.

(ii) $m(z;A) = (z-1)^p$ implies that $A - I$ is nilpotent with index p. Thus in this case $A = I + N$, where N is nilpotent. An operator of this form is called unipotent.

(iii) Suppose the minimal polynomial of A has no multiple zeros, which is equivalent to saying that m and its derivative have no common factors. Then there is a basis of V consisting of eigenvectors of A. Equivalently, among the matrices associated with A there is one which is diagonal. In this case we say that A and its associated matrices are diagonalizable or semisimple.

(iv) If $m(z;A) = (z-\lambda)^p$, then, of course, $p \le n$. A basis can be chosen so that the matrix corresponding to A has zero entries except along the diagonal where there are so-called Jordan blocks, which in case $n = 4$, for example, would be

$$\begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix}.$$

That is $n_i \times n_i$ matrices with λ on the diagonal and 1's on the first superdiagonal, $\Sigma n_i = n$, $1 \le n_i \le p$, and for at least one value of i, $n_i = p$.

In the preceding we have assumed that the entries of the matrix A could

be arbitrary. However, if they are real and nonnegative the remarkable Perron-Frobenius theorem (Senata, 1973) about the eigenvalues and eigenvectors of A gives information which is useful in many contexts; we thus state it here. A matrix $M = (m_{ij})$ is connected or indecomposable if for any two indices i and j there is a sequence $r_k, 1 \leq k \leq s$, such that the continued product $m_{ir_1} m_{r_1 r_2} m_{r_2 r_3} \ldots m_{r_s j} \neq 0$. We write $M > 0$ if all $m_{ij} > 0$, and $M \geq 0$ if all $m_{ij} \geq 0$. Then, if $M \geq 0$ is a real connected matrix, it has a largest simple positive eigenvalue, $r(M) = r$, and an associated column vector $x > 0$, such that $Mx = rx$ where $r > 0$; any other eigenvalue λ of M has absolute value less than or equal to r. Further, if $N \geq 0$ is another real matrix of the same dimension, such that $M - N \geq 0$, then $r(N) \leq r(M)$ with equality only if $N = M$. This theorem can be used to quickly give the basic classification of Kac-Moody algebras.

3.6 Canonical Form of Matrices

In Sec. 3.3 we noticed that distinct matrices were associated with the same linear operator, so there is a sense in which such matrices are "equivalent." Recall that by an equivalence relation a set is partitioned into distinct mutually exclusive subsets which exhaust the given set. One method of partitioning a set is into the orbits of a group which acts on the set. Thus if g belongs to a group G which is acting on a set S and we denote by gs the element of S into which g sends s, the orbit of s is the set $M_s = \{gs | g \in G\}$. It follows that $x \in M_s$ implies that $M_x = M_s$. Given an equivalence relation on a set of matrices, the problem considered in this section is that of choosing a canonical or "simplest" matrix in each equivalence class. There are different canonical forms depending on the types of matrices we consider and the different group actions contemplated.

Possibly the basic and most general situation is that considered by H. J. S. Smith in 1861. It is that of Sec. 3.3 where the equation $A(u', v') = PA(u, v)Q$ occurs in slightly different notation. There P and Q are arbitrary invertible $m \times m$ and $n \times n$ matrices, respectively. By choosing B'_u so that the last elements of the basis span the kernel of A and the first ones span a subspace which is complementary to the kernel, while the first elements of B_u span the range of A, one can arrive at Smith's canonical matrix which has 1's in the (i, i) positions for $1 \leq i \leq r$ where r is the rank of A, and zero everywhere else. It would be difficult to demand anything "simpler." It follows that with this meaning of equivalence there are $p + 1$ equivalence classes of $m \times n$ matrices where p is the minimum of $\{m, n\}$.

At first one is surprised that there are so few classes. However, on second thought, one notices that we have been acting on a space of matrices of dimension mn by the group $Gl(n, \mathbb{F}) \times Gl(m, \mathbb{F})$ ($= G$, say), which has $n^2 + m^2 \geq 2mn$ parameters; there is plenty of redundancy unless one of m and n is 1 and the other is 1 or 2.

If we consider an action on the set of $n \times n$ matrices by a smaller group we shall expect more equivalence classes. For $(P, Q) \in G$, subgroups of G can be defined by imposing restrictions on P and Q.

Recall the following definitions. If A is a square matrix the transpose of A, denoted by A^t, is obtained from A by interchanging rows and columns or by reflecting across the main diagonal. The operation of taking the transpose is an involution, that is $(A^t)^t = A$. If $A^t = A$, then we say A is symmetric. If $A^t = -A$, then A is antisymmetric or skew-symmetric.

An important property of transposition is $(AB)^t = B^t A^t$. It is worth noting that once the basis of V has been fixed, the mapping defined by transposition of matrices can be transferred to the associated linear transformations, thus defining an involution on $\mathrm{Hom}(V)$.

If σ is an automorphism of $\mathbb{F}$, we can define an operation on the matrix A by replacing each of its elements by its conjugate under the automorphism, and denote the new matrix by A^σ. If the field is commutative $(AB)^\sigma = A^\sigma B^\sigma$. In particular, when $\mathbb{F} = \mathbb{C}$, complex conjugation is an automorphism of period two. We follow a common custom and denote the complex conjugate of A by $\overline{A}$, so $\overline{AB} = \overline{A}\,\overline{B}$.

The Hermitian conjugate of A is denoted by $A^* = \overline{A}^t$ and satisfies $(AB) = B^* A^*$. A matrix A is Hermitian if $A^* = A$ and anti-Hermitian if $A^* = -A$.

The approach of this section is based on that of Turnbull and Aitken (1932), a superb book which goes far beyond our brief summary. They distinguish five subgroups of G.

(i) The Collinearity Group is characterized by $PQ = I$. It arises in Sec. 3.3 when $v = u$ and $v' = u'$. Under the action of this group, a square matrix, A, can be reduced to Jordan canonical form, that is to a sum of diagonal blocks, each of which has the form $\lambda I + N$, where λ is an eigenvalue of A and N is a nilpotent matrix, all of whose entries are zero except for 1's along the first superdiagonal. A particular eigenvalue occurs on the diagonal of the canonical form as many times as its multiplicity in the characteristic equation. For any eigenvalue the dimension of the largest Jordan block is equal to the multiplicity of the eigenvalue in the minimal polynomial $m(z;A)$. Thus if the zeros of $m(z;A)$ are simple, A is diagonalizable.

(ii) The Congruent Subgroup is defined by the condition $P^t = Q$. Under this group, symmetry or antisymmetry of A is invariant. A symmetric matrix can be diagonalized. If $\mathbb{F}$ is closed under taking square-roots, we can choose as the canonical element of an equivalence class a diagonal matrix which has only 0's or 1's on the diagonal. If $\mathbb{F} = \mathbb{R}$, the diagonal could also contain -1. In the real case, Sylvester's Law of Inertia asserts that the number of 1's and the number of -1's are invariants. A nonsingular antisymmetric matrix has even rank r and there is a canonical form under the congruent group which contains zeros everywhere except for $r/2$ blocks of 2×2 antisymmetric matrices down the diagonal; each has 1 and -1 off the diagonal and 0 on the diagonal.

(iii) The Conjunctive Subgroup is defined by the condition $P = Q^*$. It changes Hermitian matrices into Hermitian matrices. For real matrices, the conjunctive and the congruent transformations are the same. For any $\mathbb{F}$, one may choose a diagonal matrix as canonical. If $\mathbb{F} = \mathbb{C}$, the diagonal can consist of 1's and 0's.

(iv) The Orthogonal Group is defined by $PQ = I$ and $P = Q^t$ and is thus a subgroup of the groups (i) and (ii). It will preserve symmetry or antisymmetry of a matrix. A symmetric matrix will be equivalent to a diagonal matrix whose diagonal elements are eigenvalues of the original matrix. An antisymmetric matrix will be equivalent to one with zeros everywhere except for 2×2 blocks on the diagonal, the determinants of these blocks

being equal to the negatives of the squares of eigenvalues of the original matrix.

(v) The Unitary Subgroup is defined by $PQ = I$ and $P = Q^*$, and is thus a subgroup of (i) and (iii). It preserves the property of a matrix being Hermitian or anti-Hermitian. If $\mathbb{F} = \mathbb{R}$, groups (v) and (iv) are the same. Under this group, a Hermitian matrix is equivalent to a diagonal matrix whose nonzero elements are eigenvalues of the original matrix. An anti-Hermitian matrix is equivalent to one with 2-dimensional blocks on the diagonal whose determinants are the negatives of the squares of eigenvalues of the original matrix.

3.7 Dual Space

We have already noted that $\mathrm{Hom}(V, U)$, where V and U are linear spaces of dimension n and m, respectively, over a common field $\mathbb{F}$, can be given a structure of a linear space of dimension nm over $\mathbb{F}$. We can, of course consider $\mathbb{F}$ as a linear space of dimension 1 over F. Thus, $\mathrm{Hom}(V, \mathbb{F})$ is a linear space of dimension n over $\mathbb{F}$ and therefore isomorphic to $\mathbb{F}^n$ and hence also to V. It is called the dual space of V and usually denoted by V^*. This use of the asterisk can be distinguished from its use to indicate Hermitian conjugation by the context. The elements of V^* are linear functions on V with values in $\mathbb{F}$. We shall denote them by lower case Greek letters. Recall that the Kronecker symbol δ^i_j takes the value 1 if $i = j$ and 0 otherwise.

If $\alpha \in V^*$ and $x = x^j v_j$ is an arbitrary vector in V expressed in terms of the basis B_v, then $\alpha(x) = x^j \alpha(v_j) = a_j x^j$, where $a_j = \alpha(v_j)$. It is possible to define various bases for V^*. The basis which is said to be dual to B_v, and may be denoted by B^*_v, is defined as follows. Recall that a linear function on V is completely determined by the values it assumes for the elements of a basis of V.

Let α^i be a linear function such that $\alpha^i(v_j) = \delta^i_j$ for all j, $1 \le j \le n$. Then $\alpha^i(x) = x^i$. Thus α^i is the ith coordinate function. It easily follows that α^i are linearly independent and that $\alpha = a_j \alpha^j$, where $a_j = \alpha(v_j)$. Thus any element of V^* is a linear combination of the n elements α^j, $1 \le j \le n$, so that $B^*_v = \{\alpha^j\}$ is a basis for V. Just as the x^i are coordinates of an arbitrary element of V with respect to B_v, so a_i are coordinates of an arbitrary element of V^*. Since $a_i = \alpha(v_i)$, when the basis of V is changed, a_i changes by the same transformation as, or cogrediently with, the basis. As we noted at the end of Sec. 3.2, the x^i transform contragrediently to the basis. This distinction reflects the fact that the definition of the linear function $\alpha: x \to \alpha(x)$ is independent of the coordinate system used to describe it. A geometrical or physical entity which is described by a sequence of n numbers which transform like (a_i) is called a covariant vector. Similarly, an entity described by a sequence of n numbers which transform like (x^i) when the basis is changed is called a contravariant vector.

3.8 Tensors

Possibly it was algebraic geometers in the middle of the nineteenth century who first focused attention on the behavior of the coordinates of geometrical objects when the frame of reference is changed. But the first time this issue really impinged on physics was with the advent of Einstein's General Relativity Theory (GRT). The basic metric of GRT, $g_{ij} dx^i dx^j$, is clearly independent of the coordinate system but

since dx^i is a contravariant vector, g_{ij} will have to vary covariantly in both subscripts i and j. Then the n^2 symbols g_{ij} must be describing something (in fact, according to Einstein, the gravitational field!) which is a doubly covariant tensor.

The curvature of space-time, which allegedly explains black holes and how planets circle around the sun, is described by the Riemann-Christoffel tensor, R^i_{jkl}, which is contravariant in the index i and covariant in the other three.

The great advantage of the indicial notation, as it evolved in the writings of Eddington, Weyl, Synge, and other mathematical physicists between 1920 and 1940, is that it immediately indicates the behavior of the tensor when the underlying basis, or frame of reference, is changed. Thus if a_{ij} is a double covariant tensor and b^i is a contravariant vector (or first order tensor), then $a_{ij}b^k$ is a third order tensor covariant in two indices and contravariant in one. If we now contract on the indices j and k, we see immediately that $c_i = a_{ij}b^j$ is a covariant vector.

An algebraist would say that a_{ij} are the components of an element of $V^* \otimes V^*$, the tensor product of the dual space of V with itself. Similarly, $a^i b^j_k$ are the components of an element in the tensor product $V \otimes V \otimes V^*$. In general, the tensor product (see Sec. 4) of two linear spaces of dimension n and m is a linear space of dimension nm. In particular, $V^* \otimes U$ is isomorphic to $\mathrm{Hom}(V, U)$ and is spanned by a basis consisting of elements noted as $\alpha^i \otimes u_j$, where $1 \leq i \leq n$ and $1 \leq j \leq m$.

4
Creating Algebraic Structures

What experimental apparatus is for the physicist, the Cartesian product and quotient structures are for the algebraist. These are the principal tools with which he makes new mathematical structures.

If A and B are two sets, the Cartesian product of A and B is denoted by $A \times B$ and defined as the set $\{(x, y) | x \in A, y \in B\}$. Thus it is a new set consisting of ordered pairs with the first element of the pair belonging to A and the second to B. If $A \neq B$, $A \times B \neq B \times A$, since by definition two ordered pairs (a, b) and (c, d) are equal only if $a = c$ and $b = d$.

Things become more interesting when A and B have some algebraic structure which can be used to impose structure on the Cartesian product. For example, suppose that $A = B = \mathbb{Z}$. We define the addition of pairs $\in \mathbb{Z} \times \mathbb{Z}$ by $(x, y) + (u, v) := (x + u, y + v)$. Notice that the plus signs on the right and left have quite different meanings. One acts on pairs of integers; the others on integers. If we think of $+3$ as a translation by 3 units along the number line, we can call $(\mathbb{Z}, +)$ a translation group in one dimension. We could then think of $(\mathbb{Z} \times \mathbb{Z}, +)$ as the translation group of a two-dimensional lattice. Another familiar example is the idea due to Gauss of imposing the structure of the complex numbers on $\mathbb{R} \times \mathbb{R}$.

The direct sum of two vector spaces provides us with another important example of this construction. Suppose X and V are two linear spaces over the same field $\mathbb{F}$ with bases $\{e_i\}$, $1 \leq i \leq n$, and $\{f_j\}$, $1 \leq j \leq m$ respectively. For $x, y \in X$, $u, v \in V$, and $\alpha \in \mathbb{F}$, define (i) $(x, u) + (y, v) := (x + y, u + v)$; (ii) $\alpha(x, u) := (\alpha x, \alpha u)$. By these definitions we have imposed on $X \times V$ the structure of a linear space for which the $n + m$ elements $\{(e_i, 0), (0, f_j)\}$ form a basis. This new linear space is called the direct sum of the linear spaces X and V, and has dimension $n + m$, and is denoted by $X \oplus V$.

An apparently minor variation on the preceding definition leads us to an important but quite different object – the tensor product of X and V which is denoted by $X \otimes V$. This is a linear space which has a basis of elements belonging to the Cartesian product $X \times V$, but addition and scalar multiplication are different. (i) $(x_1 + x_2, v_1 + v_2) = (x_1, v_1) + (x_1, v_2) + (x_2, v_1) + (x_2, v_2)$; (ii) $\alpha(x, v) = (\alpha x, v) = (x, \alpha v)$. These conditions imply that $X \otimes V$ is a vector space over $\mathbb{F}$ of dimension mn, with $\{(e_i, f_j)\}$ as a basis.

Recall that an equivalence relation ρ on a set S partitions S into mutually exhaustive subsets which we call equivalence classes. A binary relation ρ on a set is an equivalence relation if it has the following three properties: (i) reflexive, $x\rho x$ for all $x \in S$, (ii) symmetric, $x\rho y$ implies $y\rho x$, (iii) transitive, $x\rho y$ and $y\rho z$ imply $x\rho z$. A subset of the partition of S contains exactly all the elements of S which are related by ρ to any one member of the subset. For example, the nails in a hardware store can be partitioned by length. Thus $x\rho y$ means $\text{length}(x) = \text{length}(y)$.

Now consider the equivalence relation ρ on $\mathbb{Z} \times \mathbb{Z}$ such that $(a, b)\rho(u, v)$ if and only if $av = bu$. We have used only properties of the integers to partition $\mathbb{Z} \times \mathbb{Z}$ into equivalence classes. But the condition we used is identical with the equality of the rational numbers a/b and u/v. We have thus established a bijection, or one-to-one correspondence, between $\mathbb{Q}$ and the equivalence classes of $\mathbb{Z} \times \mathbb{Z}$ under the relation ρ.

For any set S with equivalence relation ρ, the new set whose elements are equivalence classes of S is denoted by S/ρ and called the quotient set of S by the relation ρ. Starting from $\mathbb{Z}$ we have just created the rationals $\mathbb{Q}$ as $(\mathbb{Z} \times \mathbb{Z})/\rho$. The notion of quotient structure frequently arises in physics when we have a group G acting on some set S.

Suppose that G acts transitively on S, that is if Q is a fixed point, and P is any point, there is at least one transformation in G which sends Q to P. The set of all transformations which leave Q fixed is a subgroup H of G – the so-called stabilizer of Q. For any two elements f and g of G we shall say that they are in the relation ρ, that is $f\rho g$, if $fg^{-1} \in H$. We easily prove that ρ is an equivalence relation and that the points of S are in one-to-one correspondence with the elements of G/ρ. Thus the physics of S can be transferred to G/ρ and the symmetries of the physical situation may become more transparent.

When the relation is defined by a subgroup H as above, G/ρ is usually denoted by G/H. Suppose we denote the equivalence class containing g by $\pi(g)$, if g is any element of G. That is π is a mapping from G to G/H, the so-called canonical map. We could ask whether it is possible to impose on G/H a structure of a group in such a way that for any $f, g \in G, \pi(fg) = \pi(f)\pi(g)$. The answer is yes – if and only if H is a normal subgroup of G. A normal subgroup is not only a group but has the additional property that for all $g \in G, gHg^{-1} = H$. Further $H = \{g \in G | \pi(g) = e\}$ where e is the neutral element of the new group G/H. When the subgroup H is not normal, G/H is not a group but is called a homogeneous space on which G acts transitively.

We shall meet below other examples of the use of quotienting as a method of creating new structures.

5
Rings

A ring like a field consists of a set, R, together with two binary operations which are usually called addition and multiplication. $(R, +, \times)$ is a ring if

(i) $(R, +)$ is a commutative additive group with zero;
(ii) $(R, \times)$ is closed under multiplication and may or may not have a unit;
(iii) multiplication distributes over addition, i.e., $a(x + y) = ax + ay$ for all a, x, and y in R.

We do not require that nonzero elements of R have reciprocals in R, nor that multiplication be commutative or associative, but we do not exclude these properties. Thus a field is a ring but not all rings are fields.

5.1
Examples of Rings

We now list five rings and one "almost ring" which occur frequently in the physics literature.

(a) *The Integers* $\mathbb{Z}$. Perhaps it was this example which led to the emergence of the concept of ring. The integers form a group under addition and are therefore closed under addition and subtraction. They are also closed under multiplication, which distributes over addition. However, the solution, x, of the equation $mx = n$, where $m, n \in \mathbb{Z}$, is not, in general, an element of $\mathbb{Z}$. In contrast with some other rings there are no divisors of zero in the integers. That is you cannot find two integers, neither of which is zero, whose product is zero.

(b) *Square Matrices.* Suppose $A = (a^i_j)$, $B = (b^i_j)$, and $C = (c^i_j)$ are $n \times n$ matrices with entries in a field $\mathbb{F}$; then we define $A + B$ and AB or $A \times B$ to be $n \times n$ matrices whose entries in the ith row and jth column are, respectively, $a^i_j + b^i_j$ and $a^i_k b^k_j$. (Recall the summation convention in the Introduction.) Here $1 \le i, j, k \le n$. Let $M_n(\mathbb{F}) = M_n$ denote the set of all $n \times n$ matrices with entries in the commutative field $\mathbb{F}$. Then one can verify that $(M_n, +, \times)$ is an associative ring which is noncommutative if $n \ge 2$. The zero element of the ring is the matrix all of whose entries are 0, whereas the unit or identity for multiplication is the matrix (δ^i_j) which has 1 on the diagonal and 0 elsewhere. Notice that if $n = 2$,

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 7 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix};$$

thus the ring of square matrices possesses zero divisors.

(c) *Quaternions* were invented by Sir William Rowan Hamilton in order to give a convenient description of rotations in 3-space.
In this section we shall use j, k, s, t as indices with their ranges restricted as follows: $1 \le j, k \le 3$, and $0 \le s, t \le 3$. The quaternions, H, form an associative ring with multiplicative identity, $I = e_0$, and contain three elements e_j satisfying the conditions $e_j e_k + e_k e_j = -2\delta_{jk} e_0$, so $e_j^2 = -e_0$. Further, $e_j e_k = e_m$ where (j, k, m) is an even permutation of $(1, 2, 3)$. As a ring, H will contain

$e_0 + e_0 := 2e_0$, etc., so that H contains $\mathbb{Z}e_0$. More generally if R denotes any commutative ring, we could assume that H contains Re_0 and note this explicitly by denoting the quaternions as $H(R)$. Hamilton considered only the possibility that $R = \mathbb{R}$, the real numbers, since his concern was rotations in the 3-dimensional space of Newtonian physics – not some esoteric space of super string theory! Over R we can define H by

$$H = \{x^s e_s | x^s \in R\}.$$

Then it follows that H is closed under addition and multiplication. If we demand that the associative and distributive properties hold, we obtain a noncommutative associative ring. That it is consistent to demand the preceding properties follows from the fact that they are satisfied by 2×2 matrices with entries in R if we represent e_0 by the identity matrix and e_j by $-i\sigma_j$, where σ_j are the three Pauli matrices:

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},$$

$$\sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus, if we set $E_0 = I$ and $E_j = -i\sigma_j$, we find that

$$X = x^s E_S = \begin{bmatrix} x^0 - ix^3 & -x^2 - ix^1 \\ x^2 - ix^1 & x^0 + ix^3 \end{bmatrix},$$

and that if $x^0 = 0$, then $\det(X) = \delta_{jk} x^j x^k$, which equals the square of the Euclidean length of the vector with components x^j.

If T is any invertible 2×2 matrix and $Y = TXT^{-1}$, then trace of $Y = \mathrm{tr}(Y) = \mathrm{tr}(X)$ and the determinant $\det(Y) = \det(X)$. Since $\mathrm{tr}(X) = x^0$, it follows that $x^0 = 0$ implies $y^0 = 0$. Further, $\delta_{jk} x^j x^k = \delta_{jk} y^j y^k$, that is, Euclidean distance is preserved so the transformation from (x^1, x^2, x^3) to (y^1, y^2, y^3) is orthogonal. In particular if $T = \exp(\vartheta \sigma_3) = \cos \vartheta\, I + \sin \vartheta\, \sigma_3$, this transformation is a rotation though an angle 2ϑ about the x^3 axis.

If R is a finite commutative ring with m elements, $H(R)$ would be a noncommutative ring with m^4 elements.

When is $H(R)$ a field? Define X' by $X = x^0 I + X'$ and $\overline{X}$ by $\overline{X} = x^0 I - X'$. It then follows that $X\overline{X} = \delta_{st} x^s x^t I$. If $R = \mathbb{R}$, this vanishes only if $X = 0$. Thus $\overline{X}$ divided by $\delta_{st} x^s x^t$ is the reciprocal of X. It is not difficult to verify that $H(\mathbb{R})$ satisfies the requirements of an anticommutative or skew field. This is the field discovered by Hamilton to which we alluded in Sec. 2.2.

(d) *Boolean "Ring".* In studying what he called "The Laws of Thought", George Boole was led to introduce an algebraic structure on the subsets of any fixed set in which union, $\cup$, and intersection, $\cap$, are analogs of addition and multiplication, respectively. The original set acts as the identity for multiplication, and the empty set serves as the zero for addition. The reader can verify that most of the properties of a commutative ring are satisfied by Boole's structure, but a given subset does not have an additive inverse so that $\mathscr{P}(S)$, the set of subsets of S, is not an additive group under the binary operation $\cup$.

(e) *Lie Rings.* Let $(L, +, \circ)$ be a set L together with two binary operators such that $(L, +)$ is an additive commutative group such that the operation $\circ$ distributes over addition, so that

$$x \circ (y + z) = x \circ y + x \circ z.$$

However, the Lie product is neither commutative nor associative but satisfies the properties:

$$x \circ y + y \circ x = 0$$

and

$$x \circ (y \circ z) + y \circ (z \circ x) + z \circ (x \circ y) = 0.$$

Because of the first of these conditions, we say that the Lie product is anticommutative. The second, which replaces the associativity property of the familiar rings, is referred to as the Jacobi identity. Lie groups are discussed in other articles of this work so we do not go into details here. We merely remark that the elements of a finite dimensional Lie group can be parametrized by continuous real variables and that Sophus Lie associated to such groups what he called an infinitesimal group which is a particular case of a Lie ring. Associativity of multiplication in the group implies the validity of the Jacobi identity in the corresponding Lie ring.
The Jacobi identity can be rewritten in the form

$$z \circ (x \circ y) = (z \circ x) \circ y + x \circ (z \circ y),$$

which is the same as

$$D(x \circ y) = (Dx) \circ y + x \circ (Dy),$$

if we set $Dw = z \circ w$, for fixed z and all $w \in L$. This last equation reminds us of the product rule in calculus, so we say that the linear map $D: w \to z \circ w$ is a derivation of L. The concept of Lie ring, which apparently (Witt, 1937) was first defined by Wilhelm Magnus in 1936, is a generalization of the concept of Lie algebra introduced under the name "infinitesimal group" by Lie and Killing independently before 1880.

(f) *Grassmann Ring.* As a final example of the concept of ring we briefly describe an algebraic structure invented by Hermann Grassmann about 1840 which is basic to the theory of fermions as well as the geometry of many dimensions.
Given a field, $\mathbb{F}$, and a finite vector space $(V, \mathbb{F}, +)$ of dimension n, it is possible to define a new vector space, $V^\wedge$, of dimension 2^n over $\mathbb{F}$ and a binary operation, denoted by $\wedge$, called the wedge or Grassmann product, which distributes over addition. $(V^\wedge, \mathbb{F}, +, \wedge)$ will be the Grassmann or exterior algebra. In order to define the product $\wedge$, which is the same as that for fermion creation operators in second quantization, we proceed by induction on the grade of homogeneous elements of the algebra. Recall that in the ring $\mathbb{F}[x, y]$ of all polynomials in x and y there are special subspaces such as $ax + by$, or $ax^2 + bxy + cy^2$, or $ax^3 + bx^2y + cxy^2 + dy^3$, of homogeneous elements of dimension 2, 3, 4, respectively. Any polynomial can be expressed as a sum of homogeneous polynomials, and the summands are unique. It turns out, analogously, that if $\dim(V) = n$, $V^\wedge$ contains $n + 1$ subspaces V^p, $0 \le p \le n$, such that any element x of $V^\wedge$ can be expressed in precisely one way as $x = \Sigma_0^n x^p$, where $x^p \in V^p$. An element of V^p is said to be homogeneous of grade p. If x and y are homogeneous of grades p and q, respectively, then $x \wedge y = (-1)^{pq} y \wedge x$ is of grade $p + q$. In particular, $V^0 = \mathbb{F}$ and $V^1 = V$ by definition. It follows that if x and y are of grade 1, that is belong to V, $x \wedge y = -y \wedge x$. So if $\mathbb{F}$ has characteristic other than 2 it follows that $x \in V$ implies that $x \wedge x = 0$.
If $\{v_i\}$ is a basis of V, the $n(n-1)/2$ elements $v_i \wedge v_j$ for $i < j$ are linearly

independent and span the subspace V^2 of $V^\wedge$. Similarly V^3 is spanned by $v_i \wedge v_j \wedge v_k := (v_i \wedge v_j) \wedge v_k = v_i \wedge (v_j \wedge v_k)$ with $i < j < k$ between 1 and n. The $\dim(V^3) = n(n-1) \quad (n-2)/6$. Proceeding in this way we define all the $n+1$ homogeneous subspaces. As is known, the sum of the coefficients of the nth power of a binomial is $(1+1)^n = 2^n$, so $V^\wedge$ has dimension 2^n. The preceding terse abstract definition does not immediately suggest that the Grassmann ring is significant for fermion physics. However, this becomes plausible when one realizes that the above basis elements of grade p correspond to the Slater determinants for a system of p electrons which can be formed from a basis set of n linearly independent spin-orbitals.

5.2 Polynomial Rings

For everyday applications there is little doubt that the integers $\mathbb{Z}$ constitute the most important ring which is not also a field. Perhaps the next most important is the ring of polynomials involving one or more variables. Suppose R is any ring; then we consider all expressions of the form $P(x) = a_s x^s$, where $0 \le s \le n$, x^s denotes the sth power of the variable x, and $a_s \in R$. If $a_n \neq 0$ we say that $P(x)$ is a polynomial of degree n in x. The set of all such polynomials of arbitrary finite degree will be denoted by $R[x]$. (Note the square bracket which distinguishes the ring from the field $R(x)$ of rational functions.) Assume that the powers of x commute with the elements of R and define addition and multiplication in the obvious way. Then $(R[x], +, \times)$ is a ring which is commutative if and only if R is commutative. For example, if $R = \mathbb{Z}$, $R[x]$ is the ring of all polynomials with integer coefficients. If R is the ring of 2×2 matrices with complex entries, $R[x]$ consists of all 2×2 matrices whose entries are polynomials in x with complex coefficients. In this case the variable is often called λ. The theory of this particular ring is discussed by Turnbull and Aitken (1932), for example, under the title λ-matrices.

An obvious extension of the preceding is to adjoin two or more variables to R. Thus $R[x, y]$ denotes the set of polynomials in x and y with coefficients in R. A term such as $3x^2y^5$, formed by multiplication without addition, is called a monomial. The sum of the powers of x and y is called the degree of the monomial. Thus the degree of the preceding monomial is $2 + 5 = 7$. Clearly there are 8 different monomials of degree 7 in two variables. Any sum of these with coefficients in R is a homogeneous polynomial in x and y of degree 7. When R is a field we see that the homogeneous polynomials of degree 7 form a linear space of dimension 8.

More generally, it is of considerable interest to determine how many distinct monomials of degree n can be obtained from r variables. It is not difficult to see that the possible such monomials occur as the coefficients of t^n in the expansion of the r-fold product $\Pi(1 - x_i t)^{-1}$, where $1 \le i \le r$ and x_i are distinct variables. Setting all $x_i = 1$, we see that the required number is the binomial coefficient $\binom{r+n-1}{n}$.

This is an opportune point at which to explain the concept of a graded ring which appeared in Sec. 5.1(f) and has recently entered quantum physics in connection with super-symmetry. It is clear that any element of $R[x, y]$ is a unique sum of homogeneous terms and that the product of two homogeneous terms of degree p and

q, respectively, is homogeneous of degree $p+q$.

A graded ring (R, Σ) is a ring together with a set of grades, Σ, closed under addition, such that R contains special homogeneous elements to which a grade from Σ is assigned; any element of R can be expressed in a unique manner as a sum of homogeneous elements; the product of two homogeneous elements of grade α and β is homogeneous of grade $\alpha+\beta$. For polynomial rings we usually take Σ to be the non-negative integers. For the so-called Laurent polynomials $\mathbb{C}[t, t^{-1}]$, Σ is the set $\mathbb{Z}$ of all integers. For a Grassmann ring of r generators, Σ is the set of non-negative integers. However in that case there are no terms of grade greater than r, and for $0 \le n \le r$ the subspace of homogeneous elements of grade n has dimension $\binom{r}{n}$.

5.2.1 Binomial and Multinomial Theorem

For $r \in \mathbb{Z}, r \ge 0$, $(x+y)^r = C_n^r x^{r-n} y^n$, where n is summed from 0 to r. The binomial coefficients C_n^r satisfy the recurrence relation $C_n^{r+1} = C_n^r + C_{n-1}^r$, with which starting from $C_0^0 = 1$ we can generate the famous Pascal triangle.

Define $n! := 1 \times 2 \times 3 \cdots \times n$, which we read as "$n$-factorial" or "factorial n". C_n^r is often denoted by $\binom{r}{n}$ and is given by

$$\binom{r}{n} = \frac{r!}{n!(r-n)!}.$$

In this and most other formulas 0! is interpreted as 1.

The binomial coefficient is also the number of subsets of n elements in a set of r elements, or the so-called number of combinations of r things taken n at a time.

In the preceding, r and n are non-negative integers but, as Newton realized, the binomial coefficient can be defined as follows for any real r:

$$\binom{r}{n} = \frac{r(r-1)(r-2)\cdots(r-n+1)}{1 \times 2 \times 3 \cdots \times n}.$$

Here the numerator is a product of n factors which begin with r and are decremented successively by 1. For example with $r = -1$, $\binom{r}{n} = (-1)^n$. Hence,

$$(1-x)^{-1} = 1 + x + x^2 + x^3 \cdots,$$

which is valid when the infinite sum exists, that is if the absolute value of x is less than 1. This form was used in the preceding section to obtain the number of distinct monomials of degree n in r variables, viz. $\binom{-r}{n}$.

The binomial coefficient is a particular case of the multinomial coefficient which arises in powers of sums of more than two variables. Thus,

$$(x_1 + x_2 + \cdots + x_n)^r = C_{r_1, r_2, \ldots, r_n} \times x_1^{r_1} x_2^{r_2} \ldots x_n^{r_n},$$

where $0 \le r_i \le r$ and the summation is over all $r_i \in \mathbb{N}$ such that $\Sigma r_i = r$. It is not difficult to see that

$$C_{r_1, r_2, \ldots, r_n} = \frac{r!}{\Pi_i r_i!},$$

with the product for $1 \le i \le n$.

Like the binomial coefficient, to which it specializes when $n = 2$, this number has a combinatorial interpretation which explains its appearance in certain arguments of statistical mechanics. Suppose a set S of r distinct elements is partitioned into the union of n subsets S_i such that S_i contains exactly r_i elements; then there are $C_{r_1, r_2, \ldots, r_n}$ distinct ways in which such a partitioning can be effected.

5.2.2 Fundamental Theorem of Algebra

Since the square of any real number is positive or zero, there is no $x \in R$

such that $x^2 = -1$. Suppose there is a commutative ring R which contains $\mathbb{R}$ and also an element i such that $i^2 = -1$; then R contains all elements of the form $x + iy$, where x and y are arbitrary real numbers. As a ring, together with i, R will contain $-i$. If we require that R has no divisors of zero, since $x^2 + 1 = (x + i)(x - i)$, there are two and only two possible solutions of $x^2 + 1 = 0$, namely, i and $-i$. Further since $(x + iy)(x - iy) = x^2 + y^2 \neq 0$, any element of R of the form $x + iy$ has a reciprocal of the same form. In particular $i^{-1} = -i$. Thus the ring $\mathbb{R}[i]$, generated by the reals together with i, is in fact a field. It is of course the complex numbers which we are denoting by $\mathbb{C}$. Thus any commutative ring R which contains the reals and one solution of the equation $x^2 + 1 = 0$ contains $\mathbb{C}$.

We found $\mathbb{C}$ by starting with $\mathbb{R}$ and demanding that a simple second degree equation have a solution. If one starts with $\mathbb{Z}$ and asks for solutions of equations such as $5x + 7 = 0$, one soon realizes that one needs more numbers than the integers. This leads us to the rationals $\mathbb{Q}$. If, like the ancient Greeks, we ask for solutions of equations such as $x^2 - 3 = 0$, we are forced beyond the rationals and are led to define the reals, $\mathbb{R}$.

The extraordinary property of the complex numbers, first proved by Carl Friedrich Gauss, is that any equation of finite degree, $a_j x^j = 0$, $0 \leq j \leq n$, whose coefficients a_j belong to $\mathbb{C}$ has a solution which also belongs to $\mathbb{C}$. This result is so important that it has been called the Fundamental Theorem of Algebra. A field $\mathbb{F}$ which contains the zeros of all finite-degree polynomials which can be formed with coefficients in $\mathbb{F}$ is called algebraically closed, as we noted in Sec. 2.2.

If $P(x)$ is a polynomial of degree n with coefficients in a field $\mathbb{F}$ such that $P(\alpha) = 0$ for some $\alpha \in \mathbb{F}$, then one easily proves that $P(x) = (x - \alpha)Q(x)$, where $Q(x)$ is a polynomial of degree $n - 1$. If, like $\mathbb{C}$, $\mathbb{F}$ is algebrically closed, nothing prevents us from continuing this process so that $P(x) = c\Pi_i(x - \alpha_i)$, $0 \leq i \leq n$, if P has degree n, and c is the coefficient of x^n, where the product is over the zeros of $P(x)$. If $\mathbb{F} = \mathbb{R}$ this process will not carry through in general; however, it can be shown that a polynomial with real coefficients can always be expressed as a product of factors of first or second degree with real coefficients.

The Fundamental Theorem of Algebra assures us of the existence of a zero for any polynomial $P(x) \in \mathbb{C}[x]$, but it does not give us an effective procedure for finding such a zero. Indeed, Evariste Galois showed that it is only for polynomials of degree less than five that an explicit formula, analogous to that for solving a quadratic equation, exists. It is worth looking in detail at equations of degree less than 5.

(i) $\deg[P(x)] = 1$. Take $P(x) = ax + b$, where $a, b \in \mathbb{F}$ and $a \neq 0$. Then there is a unique zero, $x = -b/a$, which belongs to $\mathbb{F}$.

(ii) $\deg[P(x)] = 2$. Take $P(x) = ax^2 + 2bx + c$, $a \neq 0$, $a, b, c \in \mathbb{F}$. Clearly $P(\alpha) = 0$ is equivalent to $aP(\alpha) = 0$, but

$$\begin{aligned} aP(x) &= a^2x^2 + 2abx + ac \\ &= (ax + b)^2 + ac - b^2. \end{aligned}$$

Thus if α is a zero of P, $(a\alpha + b)^2 = b^2 - ac = D$. To find α it will be necessary to find the square root of D. Since by various choices of a, b, and c, D will vary over the whole of our ground field $\mathbb{F}$, it follows that quadratic equations with coefficients

in $\mathbb{F}$ will always have solutions in $\mathbb{F}$ if and only if $\mathbb{F}$ is closed under the operation of taking square roots.

In elementary school we learn to manipulate with square and cube roots and come to take them quite casually. But note that even the world's largest super-computer is not able to find the square root of 2 exactly since it is a nonrecurring decimal. To find it requires an infinite process. In other words, moving from talking about square roots to actually calculating them takes us out of algebra as defined in the Introduction of this essay!

For equations of degree n, if $P(x) = a_j x^j$, $0 \le j \le n$, $a_n \neq 0$, we could divide $P(x)$ by a_n, so there is no loss of generality in assuming that $P(x)$ is monic; that is, it has leading coefficient equal to unity. It then follows that if we replace x by $y - b$, where $nb = a_{n-1}$, the resulting polynomial in y has zero as the coefficient of y^{n-1}.

(iii) $\deg[P(x)] = 3$. By the preceding argument we may assume that $P(x) = x^3 + px + q$. It is easy to see that if $(x-\alpha)^2$ or a higher power of $(x-\alpha)$ divides a polynomial then $x - \alpha$ also divides its derivative. By seeking for a common factor of $P(x)$ and its first derivative we find that two or more of the zeros of $P(x)$ will be equal if and only if $4p^3 + 27q^2 = 0$.

This conclusion is valid for p and q belonging to any field, $\mathbb{F}$. If $\mathbb{F} = \mathbb{R}$ we know from a simple graphical argument that $P(x)$ has at least one real zero. If $p, q \in \mathbb{R}$, it is possible to show that the roots will all be real or that there will be one real and two conjugate imaginary roots according as $4p^3 + 27q^2$ is, respectively, less than or greater than 0. When it is zero all the roots are real and at least two are equal.

It is interesting to note that if $pq \neq 0$, by setting $x = k\cos\vartheta$, the solution of $x^3 + px + q = 0$ can be obtained by comparison with the identity $4\cos^3\vartheta - 3\cos\vartheta - \cos(3\vartheta) = 0$. With k and ϑ such that $3k^2 + 4p = 0$ and $pk\cos(3\vartheta) = 3q$, the three roots are $k\cos\vartheta$, $k\cos(\vartheta + 2\pi/3)$, and $k\cos(\vartheta + 4\pi/3)$. When $pq = 0$, the solution is trivial.

(iv) $\deg[P(x)] = 4$. In this case the solution of the equation $P(x) = 0$ can be made to depend on the solution of a so-called resolvent equation of degree 3. Then four zeros are obtained as expressions involving two square roots of rather complicated combinations of the coefficients of $P(x)$ and one zero of the resolvent cubic equation. These formulas are of little theoretical value so we do not display them. Nowadays anyone wanting explicit values for the zeros would obtain them by a computer-implementable algorithm.

There are elegant and deep arguments connecting the solution of equations of degree 5 and 6 with the theory of elliptic functions.

Of course, even for arbitrarily high degree there are particular equations of simple form which have convenient explicit solutions. For example, $x^n = a$ has n solutions: $r\exp(\vartheta + 2\pi ki/n)$, $0 \le k < n$, with $r > 0$ and $r^n \exp(n\vartheta) = a$.

6 Algebras

The word "algebra" is used in two distinct senses. On the one hand it refers to the

subject which the reader began to study with more or less enthusiasm around the age of 13. Recall that when Einstein asked his uncle "What is algebra?" the latter replied "In algebra you let x be the unknown and then you find x." Bourbaki would not accept this as a satisfactory definition but it started Einstein on the right path. Who knows? Without that initial encouragement we would never have had General Relativity which contains a good deal of algebra!

But *an* algebra also refers to a rather specific type of mathematical structure: a ring, which is also a linear space over some field. It is therefore a very rich structure. As a linear space, an algebra has a dimension – finite or infinite. In quantum mechanics the observables generate an associative algebra of operators on an appropriate Hilbert space, which is finite or infinite over $\mathbb{C}$ depending on the physical system.

As we noted in Sec. 5.1, rings differ among themselves according to whether the "product" is or is not associative or commutative. Because an algebra has linear-space structure and therefore a basis, it is possible to characterize different types of algebras by their structure constants which are defined as follows. Let $\{e_i\}$ be a basis for the algebra over $\mathbb{F}$ and denote the product merely by juxtaposition. Then e_ie_j is a linear combination of the basis such that $e_ie_j = c_{ij}^k e_k$, where the coefficients belong to $\mathbb{F}$. It then follows by linearity that if $x = x^ie_i$ and $y = y^ie_j$ are any elements of the algebra, the product:

$$xy = x^iy^jc_{ij}^ke_k = z^ke_k,$$

where

$$z^k = c_{ij}^kx^iy^j.$$

The nature of the product is determined by the structure constants c_{ij}^k. For example, if for all i and j they are antisymmetric in i-j, then $xy + yx = 0$, as in a Lie ring [Sec. 5.1(e)].

In understanding and classifying groups there is a class called simple groups with a key role in the sense that they are the building blocks with which other groups are constructed. For complex Lie groups the classification of all possible simple groups was essentially completed by Wilhelm Killing in 1888. The solution of the analogous problem for finite groups was achieved only in 1980 as a result of gigantic efforts over decades by hundreds of mathematicians – undoubtedly one of the truly outstanding intellectual achievements of all time. A simple group is one that contains no normal subgroups other than the full group and the one-element subgroup consisting of the identity. Another way of defining a simple group G is to say that any homomorphism of G into another group either is an isomorphism or sends every element of G onto a single element. Finally, this is the same as saying that if we denote the homomorphism by π then its kernel, $K = \{g|\pi(g) = e\}$, is either G or $\{e\}$.

The concept of homomorphism – that is, structure-preserving maps – applies to rings and algebras as well as groups. If R and R' are rings and $\pi: R \to R'$ is a map of R into R' then π is a homomorphism if $\pi(x + y) = \pi(x) + \pi(y)$ and $\pi(xy) = \pi(x)\pi(y)$ for all x and y in R. These are the conditions for a ring-homomorphism. If R is in fact an algebra, then π will also have to satisfy the condition $\pi(\alpha x + \beta y) = \alpha\pi(x) + \beta\pi(y)$, where α and β are in $\mathbb{F}$, if π is to qualify as an algebra-homomorphism. If we then define $K = \{x \in R|\pi(x) = 0\}$, we easily see that K is a subring (or subalgebra) of R. But more! If $x \in K$ and $z \in R$ then both xz and zx belong to K.

A subring of R which satisfies this last condition is called an ideal. If K is the kernel of a homomorphism π, and we define an equivalence relation on R by $x\rho y \Leftrightarrow x - y \in K$, then the image $\pi(R)$ is isomorphic to R/ρ. Conversely, if K is an ideal in R there is a homomorphism of R onto R/ρ with K as kernel. A ring (or algebra) is simple if it has no ideals other than $\{0\}$ and itself.

The famous Wedderburn theorem which emerged between 1895 and 1905 asserts that a simple finite-dimensional associative algebra is isomorphic to a complete matrix algebra for some n – that is, the algebra of all $n \times n$ matrices which we considered in Sec. 5.1(b).

6.1 Examples of Algebras

The algebra of $n \times n$ matrices over some field $\mathbb{F}$ to which we have just alluded is undoubtedly one of the most important and useful types of algebras. We mention three other algebras which are widely applied by physicists.

(a) *Frobenius or Group Algebra*. Suppose G is a finite group of order n. We can decree that the elements g_i of G are a basis for a linear space over some field $\mathbb{F}$. Then the set $A = \{x^i g_i | x^i \in \mathbb{F}\}$ is closed under addition and also under multiplication by elements of $\mathbb{F}$ and therefore is a linear space of dimension equal to the order of G. If $x = x^i g_i$ and $y = y^i g_i$ are arbitrary elements of A, we can define $xy := x^i y^j g_i g_j = \Sigma_k z^k g_k$, $1 \leq i, j, k \leq n$, where $z^k = c_{ij}^k x^i y^j$. Here, c_{ij}^k is 1 if $g_i g_j = g_k$ and 0 otherwise. With these definitions of addition and multiplication A is an algebra over $\mathbb{F}$. It is named after the Berlin mathematician G. Frobenius who made basic contributions to the theory of group characters, differential equations, and other branches of mathematics.

There is a better way to display the product xy. Denote x^i by $x(g_i)$ so that $x = \Sigma_g x(g) g$ where the sum is over all $g \in G$. It then follows that $xy = \Sigma_g z(g) g$, where $z(g) = \Sigma_h x(h) y(h^{-1} g)$ for fixed g with summation on $h \in G$. Viewed in this way there is a bijection between the elements of A and functions $x(g)$ on G. The sum $x + y$ is mapped onto the function $x(g) + y(g)$, and the product xy onto the convolution, $\Sigma_h x(h) y(h^{-1} g)$, of the functions $x(g)$ and $y(g)$.

Now suppose that G is a continuous group on which there is a concept of volume or measure which is invariant under translation by elements of the group – a so-called Haar measure. Then it is possible to extend the notion of Frobenius algebra from finite to continuous groups where the elements of the algebra are functions on G, addition is point-wise addition of functions, and the product of $x(g)$ and $y(g)$ is the function $\int x(h) y(h^{-1} g)\, dh$. This convolution of functions occurs in the theory of Fourier and Laplace transforms. A rigorous treatment involves deep problems of continuity, convergence, and measure theory which analysts love. But the algebraic structure is apparent and explains why the idea of convolution is so useful.

(b) *Clifford Algebras* are generalizations of quaternions. They are the E-numbers of Sir A. S. Eddington's Fundamental Theory. They play a key role in relativistic quantum mechanics as the Dirac matrices. They are very useful in the discussion of the orthogonal group in n-dimensions.

Suppose that $E_i, 1 \leq i \leq n$, are elements of an associative algebra over the field $\mathbb{F}$ and satisfy the conditions

$$E_i E_j + E_j E_i = 2\delta_{ij} I,$$

where the Kronecker delta is 1 if $i = j$ and 0 otherwise and I is the identity of the algebra. Since $E_i^2 = I$ and distinct E's anticommute, it follows that the algebra generated by the E's has a basis consisting of I, E_i, $E_i E_j$, $E_i E_j E_k, \ldots 1 \leq i < j < k < \ldots \leq n$, and therefore like the Grassmann algebra has dimension 2^n. However, if $X = x^i E_i$ we see that $X^2 = \delta_{ij} x^i x^j I$, displaying the Euclidean metric! It follows that if T is any fixed invertible element of the Clifford algebra, the transformation $X \to TXT^{-1}$ gives rise to a linear transformation of the coordinates x^i under which the metric is invariant. It is therefore an orthogonal transformation.

In the case $n = 4$, $2^n = 16$ which is the number of entries in a 4×4 matrix. Indeed, there is then a faithful representation of the Clifford algebra by 4×4 matrices among which the Dirac matrices appear in a natural manner.

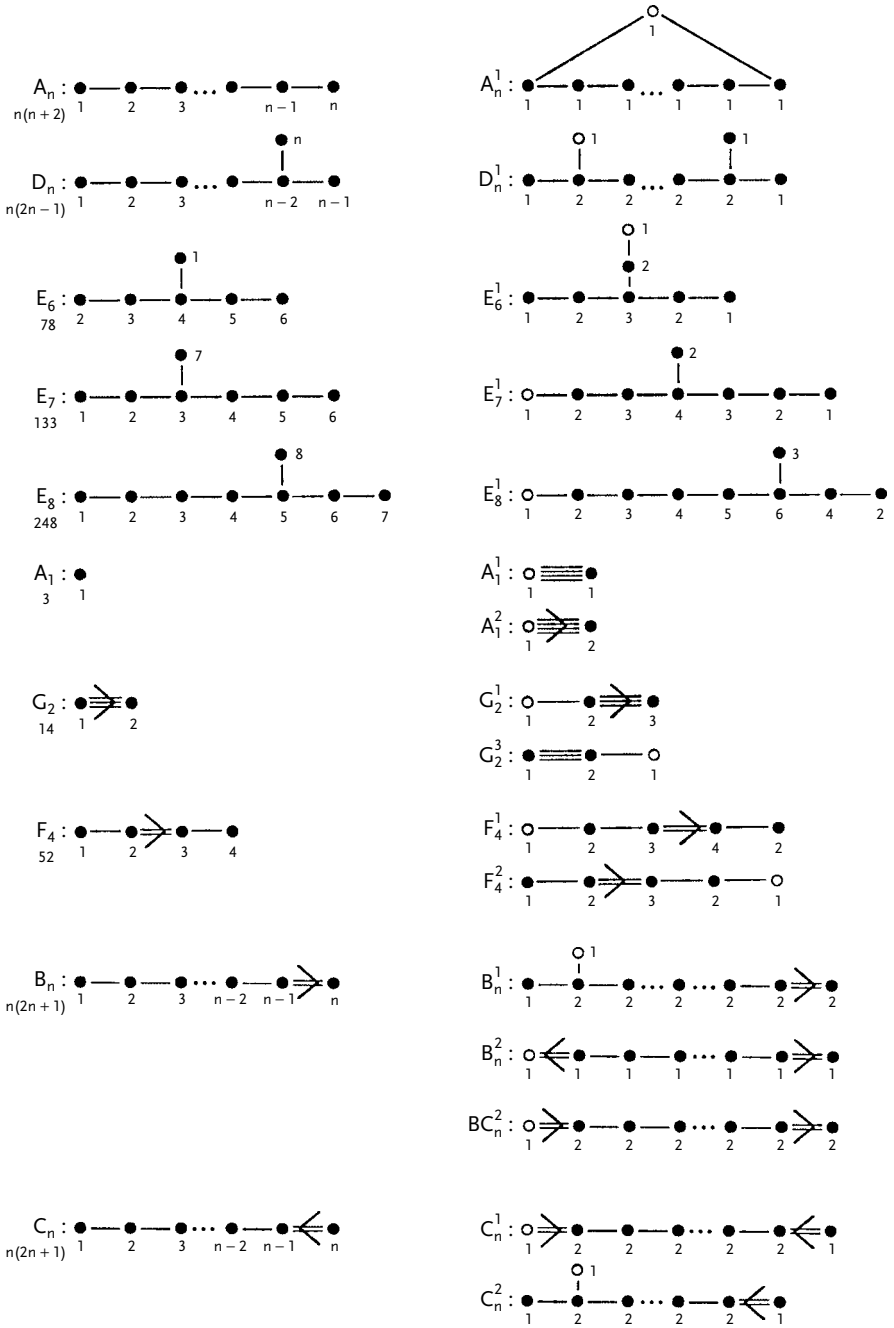
(c) *Lie Algebras*. A Lie algebra $(L, \mathbb{F}, +, \circ)$, abbreviated as LA, is a Lie ring as defined in Sec. 5.1, which is also a linear space over some field $\mathbb{F}$. As such it could have finite or infinite dimension. The smallest subalgebra of L which contains all the elements of the form $x \circ y$, for $x, y \in L$, is an ideal in L, which we denote by $L \circ L := L'$. Defining an equivalence relation ρ by $x \rho y \Leftrightarrow x - y \in L'$ leads to a quotient algebra L/ρ which is Abelian; that is the product of any two elements of L/ρ is 0. A Lie algebra is said to be solvable if the descending sequence of subalgebras $L^{p+1} \subset L^p$ where $L^0 = L$, $L^1 = L'$ and $L^{p+1} = (L^p)'$, terminates in $\{0\}$ in a finite number of steps. Every finite dimensional Lie algebra has a maximal solvable ideal, R, called the radical of L, which is such that the quotient algebra L/R is semisimple – that is, a direct sum of simple LA's. Recall that a simple ring is one with only trivial subideals. Thus simple and solvable Lie algebras are the basic building blocks needed for the analysis of any LA.

The complete classification of simple finite dimensional LA's over $\mathbb{C}$ was obtained by W. Killing (1888) and expounded clearly by Elie Cartan in his thesis (1894). The classification of simple finite dimensional LA's over $\mathbb{R}$ was achieved by Cartan (1913).

The concept of a simple LA was given a far-reaching extension by the Canadian R. V. Moody (1967) and the Russian Victor Kac (1967) quite independently of one another. These infinite dimensional algebras, which are now called Kac–Moody (K-M) algebras, quickly proved to play a key role in fundamental particle theory as does the closely related Virasoro algebra which can be thought of as a set of derivations of a K-M algebra (Goddard and Olive, 1988). The most accessible K-M algebras are the so-called affine or Euclidean algebras which, together with the finite-dimensional simple algebras, can be classified rather conveniently by means of the Coxeter-Dynkin diagrams, which are exhibited in Table 1. In this table the left-hand column names the finite-dimensional simple algebras and the right-hand column, the affine algebras. These diagrams encode so much useful information that it is worthwhile explaining them in some detail.

Up to isomorphism, a K-M algebra is characterized by a square matrix (a_{ij}),

Tab. 1 Coxeter-Dynkin diagram of the finite and affine Lie algebras



called a Cartan matrix even though it was first defined by Killing, which has the following properties: (i) $a_{ij} \in \mathbb{Z}, 1 \leq i, j \leq n$. (ii) $a_{ii} = 2$. (iii) For $i \neq j, a_{ij} \leq 0$ and if $a_{ij} = 0$ then $a_{ji} = 0$. Then L is a LA generated by $2n$ elements $\{e_i, f_i\}$ satisfying the following conditions: (a) $e_i \circ f_j = h_i \delta_{ij}$; (b) $h_i \circ e_j = a_{ij} e_j$; (c) $h_i \circ f_j = -a_{ij} f_j$; (d) $h_i \circ h_j = 0$; (e) the $h_i, 1 \leq i \leq n$, are linearly independent; (f) $\widehat{e}_i^{1-a_{ij}} e_j = \widehat{f}_i^{1-a_{ij}} f_j = 0$.

In (f) we use the notation $\widehat{x} = ad(x)$ to denote a linear map of L into L defined by $\widehat{x}y = x \circ y$. The h_i span an Abelian subalgebra, H, called the Cartan subalgebra of L of dimension n, which is also the rank of L. For $h = t^i h_i$, define $\alpha_j(h) = a_{ij} t^i$, so α_j is a linear function on H. Then (b) and (c) imply that $h \circ e_j = \alpha_j(h) e_j$ and $h \circ f_j = -\alpha_j(h) f_j$.

The algebra L is spanned by monomials which are products of the e's and f's but (a) – (d) imply that nonzero products involve only the e's or only the f's or belong to H. The algebra is graded by ascribing a grade α_i to e_i, 0 to $h \in H$, and $-\alpha_i$ to f_i. Hence the possible grades for nonzero elements of L are 0 or $\alpha = k^i \alpha_i$, where the k^i are all non-negative integers, in which case α is said to be positive and noted $\alpha > 0$, or where the k^i are all non-positive integers, in which case α is negative or $\alpha < 0$. Thus L is a direct sum of homogeneous terms which have positive, negative, or zero grade. That is, $L = L^- \oplus H \oplus L^+$ where L^- is spanned by products of f's and L^+ by products of e's.

The grades form a lattice $k^i \alpha_i$ with $k^i \in \mathbb{Z}$. This lattice is an additive Abelian group, generated by α_i for $1 \leq i \leq n$, whose elements can be pictured as the points of a crystallographic lattice. The grades for which there is a non-zero homogeneous element of L are called roots and span a sublattice of the lattice of grades. The set of roots other than 0 is generally denoted by Δ which has the obvious partition into the positive and negative roots $\Delta = \Delta^- \cup \Delta^+$, with $\Delta^- = -\Delta^+$. A nonzero element x_α of L with grade α is called an α-root vector.

As mentioned above, the K-M algebras and the simple finite LA's are classified by means of the Coxeter-Dynkin diagrams (CDD) of Table 1. These were first employed by Coxeter (1931) in his study of finite groups generated by reflections and applied by him to LA's in 1934. They were also introduced independently by Dynkin (1946). Bourbaki named them after Dynkin, having first learned about them from Dynkin's important work concerning LA's. It was only in 1949 that Claude Chevalley, a founding member of Bourbaki, learned of Coxeter's papers from the author of the present article.

In 1888 or earlier, Killing noticed that the operation $S_i: \alpha \to \alpha - \alpha(h_i)\alpha_i$ effects a permutation of Δ such that $S_i^2 = I$. The n operations $S_i, 1 \leq i \leq n$, generate a group which is now usually called the Weyl group because in 1923 Hermann Weyl popularized a particular representation of this group. I prefer to call it the Killing-Weyl group. A particular element of this group, $R = S_1 S_2 S_3 \ldots S_n$, which is usually called the Coxeter transformation, was used by Killing to effect the classification of the finite simple LA's over $\mathbb{C}$. He exhibited the orders of all the "Coxeter" transformations some years before Coxeter was born. It is perhaps only fair to follow B. Kostant's usage (1959) and call this operator the Killing-Coxeter transformation. The order of the R associated with a simple LA on the LHS of Table 1 is the sum of the digits appearing on the diagram immediately to the right. For example, the order of $R(G_2)$ is

$1 + 2 + 3 = 6$. The Killing-Coxeter transformation is discussed in Coleman (1989), where additional references may be found.

The CDD gives the Cartan matrix (a_{ij}) and also the relations among the generators of the Killing-Weyl group for the corresponding LA. The nodes of the diagram are numbered by the indices 1 to n for the finite algebras and 0 to n for the affine algebras. A branch between i and j indicates that a_{ij} and a_{ji} are different from zero and, therefore, both are negative integers. A simple branch indicates that the product $a_{ij}a_{ji}$ is 1 and, therefore, each factor is -1; a double branch indicates that the product is 2, so the factors are -1 and -2; a triple branch indicates that the factors are -1 and -3. An arrow pointing from i to j indicates that a_{ij} is numerically larger than a_{ji}. Thus in the graph G_2, $a_{12} = -3$ and $a_{21} = -1$. The symbols A_n, B_n, C_n, D_n were introduced by Killing and Cartan to denote the four infinite classes of Lie algebras corresponding, respectively, to $SL(n+1)$, the general linear group on $n+1$ variables with determinant 1; the orthogonal group on $2n+1$ variables; the symplectic group on $2n$ variables; and the orthogonal group on $2n$ variables. *E, F, G* denote the exceptional LA's and the subscripts 2, 4, 6, 7, 8, n denote the rank of the algebra. For the finite algebras their dimension is noted in Table 1 under the name of the algebra. The left-hand column thus encodes the Cartan matrix for all the finite simple LA's.

But these diagrams also enable us to infer the relations among n or $n+1$ generators, S_i, of the Killing-Weyl group of the finite simple LA's or the affine algebras. $S_i^2 = I$, and for $i \neq j$, $(S_iS_j)p = I$, where p equals 2, 3, 4, or 6 according as the i-node and the j-node are joined by a 0-, 1-, 2-, or 3-fold branch. For the simple LA's in the first column the determinants of the Cartan matrix and of all its principal subminors are strictly positive, whereas for the matrices in the second column the determinant of the matrix itself is zero but those of all principal subminors, positive. This corresponds to the fact that can be observed from the diagrams that if one node is removed from a diagram in the second column we obtain one or two diagrams of the first column. From this we infer immediately a class of finite semisimple LA's which are subalgebras of the affine algebras.

An open node is numbered 0 and the others retain their numbers from the diagram immediately to the left. The numbers attached to the nodes of the affine diagrams are the coefficients of the canonical null-root of the affine algebra. The affine algebras have infinite dimension, Killing-Weyl groups of infinite order, and an infinite number of roots. The dimension of a root space is finite and called the root multiplicity. Roots are distinguished according as they are in the orbit under the K-W group of a simple root α_i or not. The former are called real, the latter, imaginary. The real roots all have multiplicity 1, whereas imaginary roots can have arbitrarily high finite multiplicity.

By the height, $\mathrm{ht}(\alpha)$, of the positive root $\alpha = k^i\alpha_i$ we mean Σk^i. A finite dimensional simple LA has a unique root of greatest height which can be read from Table 1. The numbers attached to nodes 1 to n in the diagram X_n^1 give the coefficients of α_i for the root of greatest height in the corresponding finite algebra X_n. Thus $\alpha = 2\alpha_1 + 3\alpha_2 + 4\alpha_3 + 2\alpha_4$ is the highest root of $\Delta(F_4)$.

We should note that some authors use the symbols $A_2^{(2)}$, $D_4^{(3)}$, $E_6^{(2)}$, $D_{n+1}^{(2)}$, $A_{2n}^{(2)}$, and A_{2n-1}^2 for our algebras A_1^2, G_2^3, F_4^2, B_n^2, BC_n^2, and C_n^2, respectively.

7 Modules

So far we have not discussed exact sequences, commutative diagrams, or the game called diagram chasing which have played an increasingly important role in algebra in recent decades. The theory of modules provides a convenient context in which to introduce these ideas and is also significant in its own right.

Recall that a linear space is an additive group $(V, +)$, which is closed under multiplication by elements of a field $\mathbb{F}$ (usually commutative). We frequently denote the space $(V, \mathbb{F}, +, \times)$ by V. A module is an additive group $(M, +)$ which is closed under multiplication by elements of a ring, R (often non-commutative). Frequently, when the context is clear, a module $(M, R, +, \times)$ is denoted by M. Essential conditions in the definition of an R-module M are that for $r \in R$, and $m_1, m_2 \in M$, (i) $rm_1 \in M$, (ii) $r(m_1 + m_2) = rm_1 + rm_2$.

7.1 Examples of Modules

(a) Suppose $R = \mathbb{Z}$ and $M = \{3n | n \in \mathbb{Z}\} = 3\mathbb{Z}$. Clearly M is an additive group. Define the action of R on M as multiplication of integers. Obviously, if $n, m \in \mathbb{Z}$, $3n \in M$ and $m(3n) = 3mn \in M$.

(b) Let $M = \mathbb{Z}/3\mathbb{Z} := \mathbb{Z}_3$. M has three elements which we could denote [0], [1], and [2]. For example $[1] = \{.. -5, -2, 1, 4, 7, \ldots\}$. Take $R = \mathbb{Z}$. For $n \in R$ and $m \in \{0, 1, 2\}$, define $n[m] = [nm]$. For example, $7[2] = [14] = [2 + 3 \times 4] = [2]$. We easily check that M is an R-module.

(c) For the physicist perhaps the most important example of a module occurs when R is the Frobenius algebra of a group G and M is a linear space over $\mathbb{R}$ or $\mathbb{C}$. In this case the action of R on M is called a linear representation of G. This example will be treated in other articles in this book. Of course it was via the representations of groups that Hermann Weyl and Eugene Wigner introduced the "group pest" into physics around 1930. The algebraization of physics then took a dramatic leap forward, and we were able to understand the periodic table, relativistic wave equations, and the classification of fundamental particles.

7.2 Morphisms of Modules

A mapping $f: M_1 \to M_2$ of one R-module into another could merely map one set into the other set. As such it could be (i) injective, such that $x \neq y \Leftrightarrow f(x) \neq f(y)$, (ii) surjective such that for each $y \in M_2$ there is an $x \in M_1$ for which $f(x) = y$, or (iii) bijective, that is both injective and surjective – or, one-to-one and onto. However, f might also have the property that for all $x, y \in M_1, f(x + y) = f(x) + f(y)$. We would then say that f is a homomorphism, or, now more frequently, simply, a morphism of additive groups. Even better, f might not only be a morphism of additive groups but also have the property $f(rm) = rf(m)$ for all $r \in R$ and $m \in M_1$. It would then have the distinction of being called an R-module morphism.

The kernel $K = \{x \in M_1 | f(x) = 0\}$ of an R-module morphism f is an R-submodule of M_1. The image or range of $f, f(M_1) \subset M_2$, is an R-submodule of M_2 and is isomorphic to the quotient module M_1/K. When f is surjective, this situation is now frequently described by saying that

the following diagram portrays an exact sequence of mappings:

$$0 \longrightarrow K \longrightarrow M_1 \xrightarrow{f} M_2 \longrightarrow 0.$$

The preceding statement presupposes that the reader realizes that given our context (i) an arrow denotes a morphism of R-modules, (ii) the sequence is exact at K, M_1, and M_2. By exact is meant that the image of an incoming morphism is the kernel of the outgoing morphism. Hence, (i) the mapping from K to M_1 is injective since its kernel is 0, (ii) the mapping f is surjective since its image must be the whole of M_2 because this is the kernel of the final map. Since f is an R-module morphism the diagram implies that $M_2 = f(M_1) = M_1/K$.

Though we chose to introduce the concept of an exact sequence for modules, it has wider application. Consider the diagram:

$$1 \longrightarrow K \longrightarrow G \longrightarrow H \longrightarrow 1,$$

where now the arrows represent homorphism of groups. Then if the sequence is exact, K is injected into G and can be identified with the normal subgroup of G which is the kernel of the map from G to H. We conclude that H is isomorphic to G/K. So this particular exact sequence encapsulates the familiar and basic First Isomorphism Theorem of group theory.

Returning to modules or additive groups for which the neutral element is denoted by 0, we see that for any exact sequence:

$$\ldots \xrightarrow{\partial_{n-3}} M_{n-2} \xrightarrow{\partial_{n-2}} M_{n-1} \xrightarrow{\partial_{n-1}} M_n \xrightarrow{\partial_n}$$
$$M_{n+1} \xrightarrow{\partial_{n+1}} M_{n+2} \xrightarrow{\partial_{n+2}} \ldots,$$

$\partial_{n+1}\partial_n M_n = 0$, since $\partial_n M_n$ is the kernel of ∂_{n+1} and is mapped onto 0.

A topologist will inevitably recall that the boundary of a boundary is empty; a physicist will recall that the curl of a gradient is zero and that the divergence of a curl is zero. Indeed, these observations are the key to the algebraization of topology and of differential forms which led to homology and cohomology theory. The algebraic core of these topics has proliferated as Homological Algebra in which exact sequences and commutative diagrams are rampant.

Diagrams of maps can become quite complex and frequently admit two or more paths from object A to object B. If for all pairs A and B the compositions of maps along all different paths from A to B are identical, the diagram is said to be commutative. The game, art, or science of diagram chasing is the process of verifying that a given diagram is or is not commutative. Here, finally, is a simple but important example of a commutative diagram. It illustrates a method now widely used by algebraists to define universal objects.

Suppose G and $\hat{G}$ are given topological groups and $\pi: \hat{G} \to G$ a given homomorphism of $\hat{G}$ onto G. Suppose further that for any group H and any homomorphism τ such that $\tau(H) = G$ there exists a homomorphism σ which makes the accompanying diagram commutative:

$$\begin{tikzcd} H \arrow[dr, "\tau"'] & \hat{G} \arrow[l, "\sigma"'] \arrow[d, "\pi"] \\ & G \end{tikzcd}$$

Then $\hat{G}$ is the universal covering group of G, the kernel K of π is the Poincaré group of G, and the cardinality of K is the connectivity index of G. An example of this, which "explains" the occurrence of spin in physics, is that if G were

SO(3) then $\hat{G}$ would be isomorphic to SU(2).

Glossary

This article contains so many words which may be unfamiliar to the reader that if we gave their definition here the glossary would be almost as long as the article. Therefore, we list the most important concepts followed by the section number in which the concept can be found.

Abelian: see Sec. 1
Affine Lie Algebra: see Sec. 6.1(c)
Algebra: see Sec. 6
Algebraically Closed: see Sec. 2.2
Anti-Hermitian: see Sec. 3.6
Antisymmetric: see Sec. 3.6
Associativity: see Sec. 1
Automorphism: see Sec. 2.2
Basis: see Sec. 3.1
Bijective: see Sec. 7.2
Binary: see Sec. 1
Binomial Coefficient: see Sec. 5.2.1
Canonical Matrix: see Sec. 3.6
Cartan Matrix: see Sec. 6.1(c)
Cartan Subalgebra: see Sec. 6.1(c)
Chasing Diagrams: see sec. 7.2
Characteristic Function: see Sec. 3.5
Characteristic of a Field: see Sec. 2.1
Clifford Algebra: see Sec. 6.1(b)
Cogredient: see Sec. 3.7
Commutative: see Sec. 1
Commutative Diagram: see Sec. 7.2
Component: see Sec. 3.1
Connected Matrix: see Sec. 3.5
Contravariant: see Sec. 3.7
Contragredient: see Sec. 3.7
Contragredient Vector: see Sec. 3.7
Convolution: see Sec. 6.1(a)
Covariant: see Sec. 3.7
Covariant Vector: see Sec. 3.7
Coxeter-Dynkin Diagram: see Sec. 6.1(c)
Degree: see Sec. 2.3
Derivation: see Sec. 5.1(e)
Determinant: see Sec. 3.4
Diagonalizable: see Sec. 3.5
Dimension: see Sec. 3.1
Direct Sum: see Sec. 5
Distributivity: see Sec. 2
Domain: see Sec. 7
Dual Space: see Sec. 3.7
Eigenvalue: see Sec. 3.5
Eigenvector: see Sec. 3.5
Equivalence Relation: see Sec. 4
Exact Sequence: see Sec. 7.2
Field: see Sec. 2
Frobenius Algebra: see Sec. 6.1(a)
Fundamental Theorem of Algebra: see Sec. 5.2.2
Galois Field: see Sec. 2.1
Galois Group: see Sec. 2.2
Graded Ring: see Sec. 5.2
Grassmann Product: see Sec. 5.1
Grassmann Ring: see Sec. 5.1(f)
Group: see Sec. 1
Height of a Root: see Sec. 6.1(c)
Hermitian: see Sec. 3.6
Homogeneous Space: see Sec. 4
Homogeneous Subspace: see Sec. 5.1(f)
Homomorphism: see Secs. 3.3 and 6
Ideal: see Sec. 6
Identity Element: see Sec. 1
Imaginary Root: see Sec. 6.1(c)
Indecomposable: see Sec. 3.5
Injective: see Sec. 7.2
Invertible: see Introduction
Isomorphism: see Sec. 2.2
Jacobi Identity: see Sec. 5.1(e)
Jordan Block: see Sec. 3.5
Jordan Canonical Form: see Sec. 3.6(i)
Kernel: see Secs. 3.3 and 7.2
Killing-Coxeter Transformation: see Sec. 6.1(c)
Killing-Weyl Group: see Sec. 6.1(c)
Laurent Polynomials: see Sec. 5.2
Lie Algebra: see Sec. 6.1(c)
Lie Ring: see Sec. 5.1(e)

Linear Dependence and Independence: see Sec. 3.1
Linear Map: see Sec. 3.3
Linear Space: see Sec. 3
Matrix Ring: see Sec. 5.1(b)
Minimal Polynomial: see Sec. 3.5
Module: see Sec. 7
Monic: see Sec. 3.5
Monomial: see Sec. 5.2
Morphism: see Sec. 7.2
Multinomial Coefficient: see Sec. 5.2.1
Neutral Element: see Sec. 1
Nilpotent Matrix: see Sec. 3.5
Normal Subgroup: see Sec. 4
Nullity: see Sec. 3.3
Order of a Group: see Sec. 1
Partial Fractions: see Sec. 2.3
Pauli Matrices: see Sec. 5.1(c)
Polynomial of Degree *n*: see Sec. 5.2
Quaternion: see Secs. 2.2 and 5.1(c)
Quotient Set: see Sec. 4
Radical: see Sec. 6.1(c)
Range: see Sec. 3.3
Rank: see Sec. 3.3
Rational Functions $\mathbb{F}(x)$: see Sec. 2.3
Real Roots: see Sec. 6.1(c)
Reflexive Relation: see Sec. 4
Ring: see Sec. 5
Root: see Sec. 6.1(c)
Root Multiplicity: see Sec. 6.1(c)
Root Vector: see Sec. 6.1(c)
Scalar: see Sec. 3
Semisimple Lie Algebra: see Sec. 6.1(c)
Simple Group: see Sec. 6
Skew Field: see Sec. 5.1(c)
Solvable Lie Algebra: see Sec. 6.1(c)
Span: see Sec. 3.1
Stabilizer: see Sec. 4
Structure Constants: see Sec. 6
Sum of Maps: see Sec. 3.3
Surjective: see Sec. 7.2
Symmetric Matrix: see Sec. 3.6
Symmetric Relation: see Sec. 4
Tensor: see Secs. 3.5 and 3.8
Tensor Product: see Sec. 3.8 and 4
Transition Matrix: see Sec. 3.2
Transitive Group: see Sec. 4
Transitive Relation: see Sec. 4
Transpose: see Sec. 3.6
Unipotent: see Sec. 3.5
Universal Covering Group: see Sec. 7.2
Wedderburn Theorem: see Sec. 6

List of Works Cited

Cartan, E. (1984), *Sur la structure des groupes de transformation finis et continus*, Thèse, Paris: *Oeuvres Complètes* (1984) **I**, 137–257.

Coleman, A. J. (1989), *Invent. Math.* **95**, 447–477.

Coxeter, H. S. M. (1931), *J. London Math. Soc.* **6**, 132–136.

Dynkin, E. B. (1946), *Mat. Sbornik* **18**, 347–352.

Jacobson, J. (1964), *Lectures in Abstract Algebra* (3 Vols.), New York: van Nostrand.

Kac, V. (1967), *Func. Anal. Appl.* **1**, 328–329.

Killing, W. (1888–90), *Math. Annalen* **31**, 252–290; **33**, 1–48; **34**, 57–122; **36**, 161–189.

Kostant, B. (1959), *Am. J. Math.* **81**, 973–1032.

MacDuffee, C. C. (1943), *Vectors and Matrices*, Oberlin, OH: Mathematical Association of America.

Moody, R. V. (1967), *Bull. Am. Math. Soc.* **73**, 217–221.

Senata, E. (1973), *Non-negative matrices and Markov chains*, New York: Springer.

Turnbull, H. W., Aitken, A. C. (1932), *An Introduction to the Theory of Canonical Matrices*. London: Blackie and Sons.

Witt, E. (1937), *J. Reine Angew. Math.* **177**, 152–160.

Further Reading

Albert, A. A. (1956), *Fundamental Concepts of Higher Algebra*, Chicago: University of Chicago.

Bourbaki, N. (1942), *Algebra*, Paris: Hermann.

Goddard, P., Olive, D. (1988), *Kac-Moody and Virasoro Algebras*, Singapore: World Scientific.

Kac, V. (1985), *Infinite Dimensional Lie Algebras*, New York: Cambridge University.

Lidl, R., Niederreiter, H. (1985), *Finite Fields*, Reading, MA: Addison-Wesley.
MacDuffee, C. C. (1943), *Vectors and Matrices*, Oberlin, OH: Mathematical Association of America.
Maclane, S., Birkhoff, G. (1967), *Algebra*, New York: MacMillan.
Maltsev, A. I. (1963), *Foundations of Linear Algebra*, San Francisco: Freeman.
Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists*, London: Academic.
Shaw, R. (1983), *Linear Algebra and Group Representations*, London: Academic.

Analytic Methods

Charlie Harper
Department of Physics, California State University, Hayward, California, USA

Introduction ***35***
1 Functions of a Complex Variable ***36***
1.1 Introduction *36*
1.2 Complex Variables and Their Representations *36*
1.3 Analytic Functions of a Complex Variable *37*
1.4 Contour Integrals *38*
1.5 The Taylor and Laurent Expansions *39*
1.6 The Cauchy Residue Theorem *41*
1.7 The Cauchy Principal Value and Dispersion Relations *43*
1.8 Conformal Transformations *44*
2 Ordinary Differential Equations ***46***
2.1 Introduction *46*
2.2 First-Order Linear Differential Equations *47*
2.2.1 Separable Differential Equations *47*
2.2.2 Exact Differential Equations *48*
2.2.3 Solution of the General Linear Differential Equation *49*
2.3 Second-Order Linear Differential Equations *49*
2.3.1 Homogeneous Differential Equations with Constant Coefficients *50*
2.3.2 Nonhomogeneous Differential Equations with Constant Coefficients *51*
2.3.3 Homogeneous Differential Equations with Variable Coefficients *52*
2.3.4 Nonhomogeneous Differential Equations with Variable Coefficients *53*
2.4 Some Numerical Methods for Ordinary Differential Equations *54*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

2.4.1 The Improved Euler Method for First-Order Differential Equations *54*
2.4.2 The Runge–Kutta Method for First-Order Differential Equations *56*
2.4.3 Second-Order Differential Equations *56*
3 Partial Differential Equations *57*
3.1 Introduction *57*
3.2 The Time-Independent Schrödinger Wave Equation *58*
3.3 One-Dimensional Mechanical Wave Equation *58*
3.4 One-Dimensional Heat Conduction Equation *60*
3.5 The Two-Dimensional Laplace Equation *61*
3.6 Fourier Transform Method *62*
3.7 Green's Functions in Potential Theory *64*
3.8 Numerical Methods for Partial Differential Equations *64*
3.8.1 Fundamental Relations in Finite Differences *65*
3.8.2 Two-Dimensional Laplace Equation: Elliptic Equation *65*
3.8.3 One-Dimensional Heat Conduction Equation: Parabolic Equation *67*
3.8.4 One-Dimensional Wave Equation: Hyperbolic Equation *67*
4 Integral Equations *67*
4.1 Introduction *67*
4.2 Integral Equations with Degenerate Kernels *68*
4.3 Integral Equations with Displacement Kernels *69*
4.4 The Neumann Series Method *70*
4.5 The Abel Problem *70*
5 Applied Functional Analysis *70*
5.1 Introduction *70*
5.2 Stationary Values of Certain Definite Integrals *71*
5.3 Hamilton's Variational Principle in Mechanics *74*
5.3.1 Introduction *74*
5.3.2 Generalized Coordinates *74*
5.3.3 Lagrange's Equations *75*
5.3.4 Format for Solving Problems by Use of Lagrange's Equations *76*
5.4 Formulation of Hamiltonian Mechanics *76*
5.4.1 Derivation of Hamilton's Canonical Equations *76*
5.4.2 Format for Solving Problems by Use of Hamilton's Equations *77*
5.4.3 Poisson's Brackets *77*
5.5 Continuous Media and Fields *77*
5.6 Transitions to Quantum Mechanics *78*
5.6.1 The Heisenberg Picture *78*
5.6.2 The Schrödinger Picture *79*
5.6.3 The Feynman Path-Integral Approach to Quantum Mechanics *79*

Glossary *80*
Further Reading *80*

Introduction

The article on analytic methods is subdivided into the following five broad and interrelated subjects: functions of a complex variable, ordinary differential equations, partial differential equations, integral equations, and applied functional analysis. Throughout the article, emphasis is placed on methods of application involving physical problems and physical interpretations of solutions rather than on a rigorous mathematical presentation. Special cases of linear relations in one and two Cartesian dimensions are used to explain techniques. Extensions to more general cases and different coordinate systems are straightforward, in principle.

Section 1 is devoted to some aspects of complex variable theory needed in mathematical physics. The section begins with a discussion of complex variables and their representations, analytic and singular functions of a complex variable, important integral relations, and the Taylor and Laurent expansions. The Cauchy residue theorem is applied to obtain the Cauchy principal value of an integral and dispersion relations. A discussion of the uses of dispersion relations throughout physics is also given. The section is concluded with a brief discussion of physical applications of conformal transformations and Riemann surfaces.

Section 2, on ordinary differential equations, treats classes of physical problems that lead to first- and second-order ordinary linear differential equations. Procedures for obtaining solutions for first- and second-order ordinary linear differential equations are presented. Methods of applying initial and boundary conditions are discussed. Green's functions are introduced in connection with the variation of parameters method for solving second-order nonhomogeneous differential equations with variable coefficients. A brief introduction to numerical methods for solving first- and second-order ordinary differential equations is also presented.

In the section on partial differential equations (Sec. 3), some important partial differential equations involving the Laplacian operator are presented and explained. Separation of variables and Fourier transform methods for solving partial differential equations are illustrated. Green's functions for three-dimensional problems are discussed in this section. Extensions to cylindrical and spherical coordinates and to certain special functions in mathematical physics are discussed. The section is concluded with a brief presentation of numerical methods for solving partial differential equations.

An introduction to one-dimensional linear integral equations is given in Sec. 4. Discussions of classifications and methods of solution of integral equations are given. The essential difference between an integral- and a differential-equation formulation of a physical problem is discussed. The Abel problem is presented as an example of a physical problem that leads directly to an integral equation.

The focus of Sec. 5 is on applied functional analysis. The method of the calculus of variations is introduced in

connection with finding the extremum of the definite integral of a functional, and techniques of variational calculus are applied to Hamilton's variational principle of mechanics. The Feynman path integral approach to quantum mechanics is presented as an example of functional integration.

1 Functions of a Complex Variable

1.1 Introduction

The imaginary number, $i = \sqrt{-1}$, was introduced into mathematics during the latter part of the sixteenth century. Imaginary numbers are needed since certain equations, for example, $x^2 + 1 = 0$, have no solutions that involve only real numbers. In physics, one writes the solution of the equation of motion for the linear harmonic oscillator, $\ddot{x} + \omega^2 x = 0$, in the form $x(t) = A\exp(i\omega t)$. Index of refraction is written in complex (containing real and imaginary parts) form in modern optics, and the wave function in quantum mechanics is often a complex quantity. How physical results are obtained from complex numbers or functions of a complex variable will be explained below. Complex variables are used throughout physics, and this section is devoted to discussions of some properties of complex variables that are useful in physical applications.

1.2 Complex Variables and Their Representations

A complex variable may be written in the general form

$$z = x + iy = re^{i\theta}. \qquad (1)$$

In Eq. (1), x and y are the respective real and imaginary parts of z and are written as $x = \mathrm{Re}z$ and $y = \mathrm{Im}z$; θ is the argument (phase) of z and is written as $\theta = \arg z = \theta_p + 2\pi n$ for $n = 0, 1, 2, \ldots$; θ_p is the principal argument of z and varies from 0 to 2π; $e^{i\theta} = \cos\theta + i\sin\theta$ (Euler's formula); and $r = |z|$ is the absolute value (magnitude, modulus) of z where $r = (x^2 + y^2)^{1/2}$. The complex conjugate of z is denoted as z^* (for notational convenience, $\bar{z}$ is sometimes used to denote complex conjugate) and is obtained by changing the sign of the imaginary part (or imaginary terms) of z, $z^* = x - iy$. It is clear, and useful to note for physical purposes, that z^*z is a real quantity. Complex variables are subject to the same algebraic laws as real variables. The Argand diagram (z-plane diagram) is a convenient geometrical representation of a complex variable and is illustrated in Fig. 1.

On raising Eq. (1) to the nth power, one obtains

$$z^n = r^n(\cos\theta + i\sin\theta)^n = r^n e^{in\theta}$$
$$(n = 0, \pm 1, \pm 2, \ldots \text{ and } z \neq 0). \qquad (2)$$

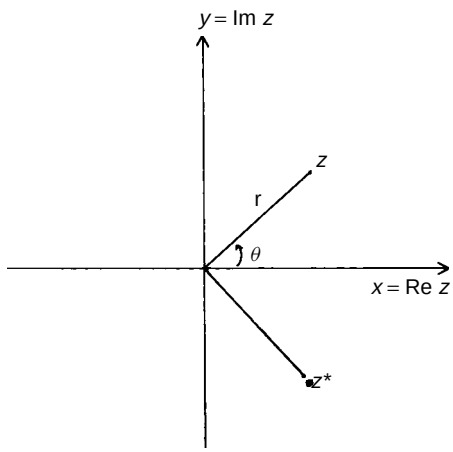


Fig. 1 Argand diagram

Equation (2) is de Moivre's theorem and is often written as

$$(\cos\theta + i\sin\theta)^n = \cos(n\theta) + i\sin(n\theta). \tag{3}$$

de Moivre's theorem may be used to obtain relations involving sines and cosines of multiple angles. On considering de Moivre's theorem for $n = 2$, expanding the left-hand side, and equating corresponding real and imaginary parts, the following well-known relations are obtained: $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\cos\theta\sin\theta$. For $n > 2$, the binomial expansion may be used to expand the left-hand side of de Moivre's theorem, Eq. (3).

By use of de Moivre's theorem, the nth root of z may be written as

$$z^{1/n} = r^{1/n}\left[\cos\left(\frac{\theta + 2\pi k}{n}\right) + i\sin\left(\frac{\theta + 2\pi k}{n}\right)\right]; \quad k = 0, 1, 2, \ldots, n-1. \tag{4}$$

The quantity $r^{1/n}$ represents the positive nth root of r. The square root of i, where $r = 1$ and $\theta_p = \pi/2$, is found to illustrate the procedure for applying Eq. (4). Roots for $k = 0$ and $k = 1$, respectively, are

$$z = \cos\left(\frac{\pi}{4}\right) + i\sin\left(\frac{\pi}{4}\right) = \frac{1+i}{\sqrt{2}} \tag{5}$$

and

$$z = \cos\left(\frac{3\pi}{4}\right) + i\sin\left(\frac{3\pi}{4}\right) = -\frac{1+i}{\sqrt{2}}. \tag{6}$$

The above two roots may be checked for correctness. The procedure used to calculate the square root of i can be applied to calculate the nth root of any quantity $z (z \neq 0)$.

1.3 Analytic Functions of a Complex Variable

A function $f(z)$ of a complex variable is itself a complex quantity and may be written in terms of real and imaginary parts in the following manner:

$$f(z) = u(x, y) + iv(x, y). \tag{7}$$

The Argand diagram representations of z and $f(z)$ are respectively called z-plane and w-plane diagrams. The number $w = f(z)$ is the value of $f(z)$ at z. A single-valued function $f(z)$ is analytic (regular, holomorphic) at z_0 if it has a unique derivative at z_0 and at every point in the neighborhood of z_0. If a function fails to be analytic at some point z_0 but is analytic at points in the neighborhood of z_0, then z_0 is said to be a singular point (singularity) of $f(z)$. In this connection, note that the function $1/z$ is analytic everywhere except at $z = 0$ (singular point). Liouville's theorem states that a function which is analytic for all z (including infinity) must equal a constant.

By analogy with the case of real variables, the derivative of a function of a complex variable is defined as

$$f'(z) = \lim_{\Delta z \to 0}\left(\frac{f(z + \Delta z) - f(z)}{\Delta z}\right) = \lim_{\Delta z \to 0}\left(\frac{\Delta u + i\Delta v}{\Delta x + i\Delta y}\right). \tag{8}$$

The evaluation of Eq. (8) for the two paths (a) $\Delta x = 0$ and $\Delta y \to 0$ and (b) $\Delta x \to 0$ and $\Delta y = 0$ leads to

$$\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = -i\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}. \tag{9}$$

The Cauchy–Riemann conditions for analytic $f(z)$ result from equating corresponding real and imaginary parts of Eq. (9); the

results are

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{10}$$

If u and v possess continuous partial derivatives to second order, Eq. (10) leads to

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \text{ and } \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0. \tag{11}$$

The equations in Eq. (11) are two-dimensional Laplace equations, and functions u and v (called harmonic or conjugate functions) are, therefore, solutions of the two-dimensional Laplace equations. The theory of analytic functions is extremely useful in solving problems in electrostatics, fluid mechanics, or whenever the two-dimensional Laplace equation occurs. The function $f(z)$ also satisfies the two-dimensional Laplace equation.

1.4 Contour Integrals

The integral of a function of a complex variable $f(z)$ is defined in a manner analogous to the case of real variable theory and may be written as

$$\int_C f(z)dz \equiv \lim_{\substack{n \to \infty \\ \max|z_j - z_{j-1}| \to 0}} \left(\sum_{j=1}^{n} f(\xi_j) \times (z_j - z_{j-1}) \right) \equiv \int_{z_0}^{z'} f(z)\, dz. \tag{12}$$

The path (contour) of integration C is divided into n segments by points z_j, and ξ_j is a point between z_j and z_{j-1} (see Fig. 2). In complex variable theory, the integral in Eq. (12) is referred to as the contour integral of $f(z)$ along the path C from z_0 to z'. The integral around a closed path is denoted as $\oint f(z)dz$. The sign convention for contour integrals is as follows: When the path of integration is traversed such that the region of interest is on the left, the integral is considered positive. Regions in the complex plane are classified as either simply connected or multiply connected. Simply connected regions possess the following three equivalent properties: every closed path within the region contains only points that belong to the region, every closed path within the region can be shrunk to a point, and every scissors cut starting at an arbitrary point on the boundary and finishing at another point on the boundary separates the region into two unconnected pieces. Regions that are not simply connected are said to be multiply connected.

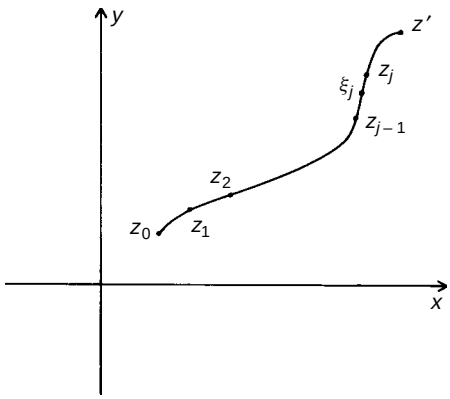


Fig. 2 Path for the contour integral in Eq. (12)

Two extremely important relations involving integrals of a function of a complex variable, the Cauchy integral theorem and the Cauchy integral formula, will now be discussed.

The *Cauchy Integral Theorem*: If $f(z)$ is analytic throughout a simply connected region Γ and C is a closed path within Γ, then

$$\oint_C f(z)dz = 0. \tag{13}$$

Cauchy's integral theorem applies to special cases that are important in physics where the value of the integral of a function depends only on end points and is independent of the path taken between end points. The inverse of this theorem is also valid.

The *Cauchy Integral Formula* is written as

$$\oint_C \frac{f(z)dz}{z - z_0} = 2\pi\, i f(z_0). \tag{14}$$

The function $f(z)$ in Eq. (14) is analytic within C, z_0 is within C, and the integrand is not analytic at $z = z_0$.

By use of the definition of $f'(z)$ and Cauchy's integral formula, the nth derivative of $f(z)$ evaluated at $z = z_0$ may be written as

$$f^{(n)}(z_0) = \frac{n!}{2\pi\, i} \oint_C \frac{f(z)dz}{(z - z_0)^{n+1}}. \tag{15}$$

Equation (15) will be used below in developing the Taylor expansion for $f(z)$.

1.5 The Taylor and Laurent Expansions

Two important series expansions, Taylor's series and Laurent's series, are valid for a function of a complex variable. If $f(z)$ is analytic in a region Γ and C is a circle within Γ with center at z_0 (see Fig. 3), then the Taylor expansion of $f(z)$ where $f^{(n)}(z_0) = n!a_n$ is

$$f(z) = \sum_{n=0}^{\infty} \frac{(z - z_0)^n}{n!} f^{(n)}(z_0)$$
$$= \sum_{n=0}^{\infty} a_n (z - z_0)^n. \tag{16}$$

The Taylor expansion of $f(z)$ is obtained and applied in a manner similar to that in

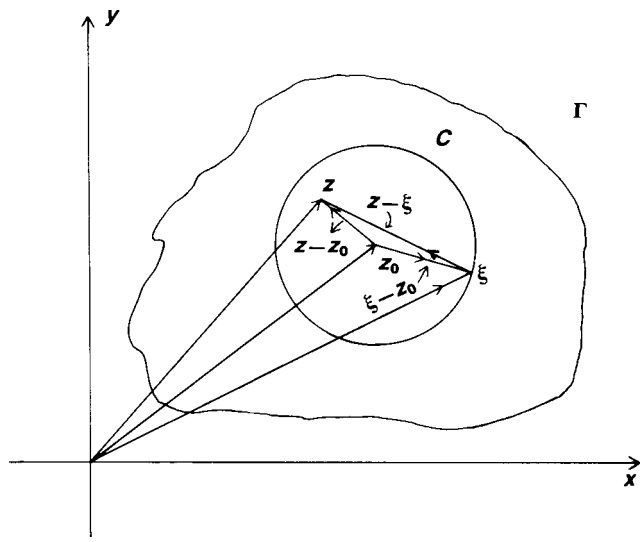


Fig. 3 Diagram for the Taylor expansion

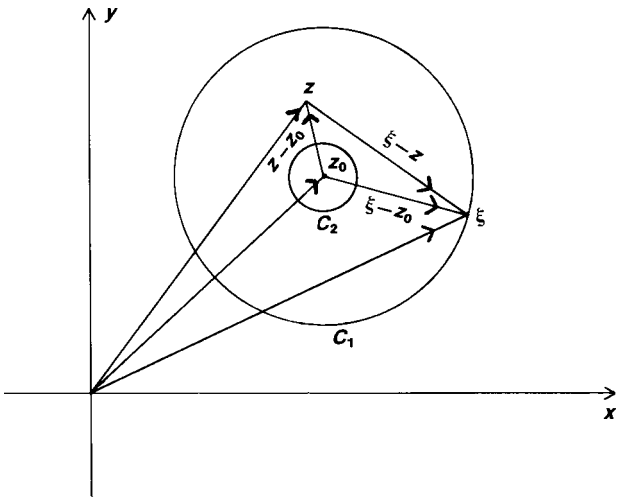


Fig. 4 Diagram for the Laurent expansion

real variable theory. Classification of the zeros of $f(z)$ is made by use of Taylor's expansion of $f(z)$ as follows: (a) If $f(z) = 0$ at $z = z_0$, the point z_0 is said to be a zero of $f(z)$. (b) If $a_0 = a_1 = \cdots = a_{m-1} = 0$ but $a_m \neq 0$, then $f(z)$ has a zero of order m at $z = z_0$. It is therefore clear that the conditions $f(z_0) = 0$ and $f'(z_0) \neq 0$ indicate the existence of a zero of order $m = 1$ (called simple zero) at $z = z_0$.

The Laurent expansion of $f(z)$ has no real-variable counterpart and is key in the discussion of singularities and residues. If $f(z)$ is analytic in the interior and on the boundary of a circular ring between two circles C_1 and C_2 (see Fig. 4), it may be represented as a Laurent expansion which has the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n + \sum_{n=1}^{\infty} \frac{a_{-n}}{(z - z_0)^n}. \tag{17}$$

In Eq. (17), the coefficients a_n and a_{-n} have the forms

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z)dz}{(z - z_0)^{n+1}}, \qquad n = 0, 1, 2, \ldots, \tag{18}$$

and

$$a_{-n} = \frac{1}{2\pi i} \oint_C (z - z_0)^{n-1} f(z)\, dz, \qquad n = 1, 2, \ldots. \tag{19}$$

The first series in the Laurent expansion, Eq. (17), is called the analytic part, and it converges everywhere within C_1. The second series in the Laurent expansion is the principal part which converges everywhere outside C_2. The quantity a_{-1} is the residue of $f(z)$ at $z = z_0$ and is given by

$$a_{-1} = \frac{1}{2\pi i} \oint_C f(z)\, dz. \tag{20}$$

In the above three equations, C is any circle between C_1 and C_2 that encloses z_0. Note that $2\pi i a_{-1}$ is the value of the integral in Eq. (20). For cases where the residue can be determined directly, an indirect method of evaluating definite integrals may be developed. First, the classification of isolated singularities and calculations of corresponding residues are considered.

A singularity at z_0 is said to be isolated if a circle of radius ε, containing no

other singularities, can be drawn with z_0 as its center. Singularities are classified using the principal part of the Laurent expansion. If the first m terms in the principal part are different from zero but the remaining terms equal zero, then $f(z)$ has a singularity (a pole) of order m at z_0. When $m = 1$, the pole is called a simple pole. If m is infinite, the singularity at z_0 is said to be an essential singularity. The residue of $f(z)$ at z_0 may be obtained by use of the following three methods.

1. The Laurent expansion directly [the coefficient of the $1/(z - z_0)$ term]. In the Laurent expansion

$$f(z) = \frac{1}{z^3} - \frac{1}{3!z} + \frac{z}{3!} - \frac{z^3}{7!} + \ldots, \tag{21}$$

there is a third-order pole at $z = 0$ with residue $a_{-1} = 1/3!$.

2. The general formula

$$a_{-1} = \lim_{z \to z_0} \left(\frac{1}{(m-1)!} \frac{d^{m-1}\phi(z)}{dz^{m-1}} \right), \tag{22}$$

where

$$\phi(z) = (z - z_0)^m f(z)$$

for $\lim_{z \to z_0}[\phi(z)]$ analytic and nonzero. To illustrate the procedure for applying the general formula, let us classify the singularities and calculate the residues of

$$f(z) = \frac{1}{(z^2 + a^2)^2} = \frac{1}{(z + ia)^2 (z - ia)^2} \quad \text{for} \quad a > 0. \tag{23}$$

There are singularities at $z = \pm ia$. Note that $m = 2$, $\phi(ia)$ is nonzero and analytic, and the residue is $1/(4ia^3)$ when $z = ia$. In a similar manner, the residue for the singularity at $-ia$ is $-1/(4ia^3)$.

3. If $f(z) = g(z)/h(z)$ where $g(z_0) \neq 0$, $h(z_0) = 0$, but $h'(z_0) \neq 0$, then

$$a_{-1} = \frac{g(z_0)}{h'(z_0)}. \tag{24}$$

For analytic $A(z)$ in $f(z) = A(z)/\sin z$, method **3** may be used to calculate the residue. There are singularities at $z = n\pi$ for $n = 0, \pm 1, \pm 2, \ldots$; here the quantity $h(n\pi)$ equals zero, but $h'(n\pi)$ is different from zero. These poles are therefore simple poles and the residue is $a_{-1} = (-1)^n A(n\pi)$.

1.6 The Cauchy Residue Theorem

The Cauchy residue theorem and Cauchy principal value of an integral result from the applications of the Cauchy integral theorem and the Cauchy integral formula. The residue theorem and principal value of an integral are extremely important in physics.

Cauchy's Residue Theorem: If $f(z)$ is analytic within and on a closed region Γ (except at a finite number of isolated

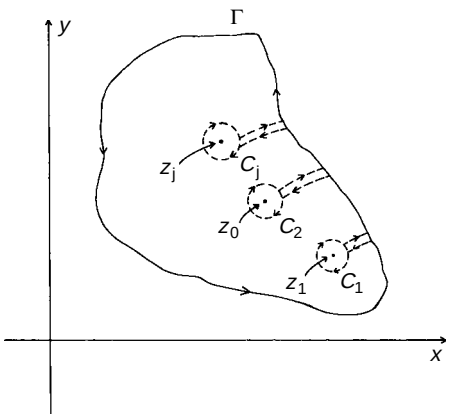


Fig. 5 Diagram for the Cauchy residue theorem

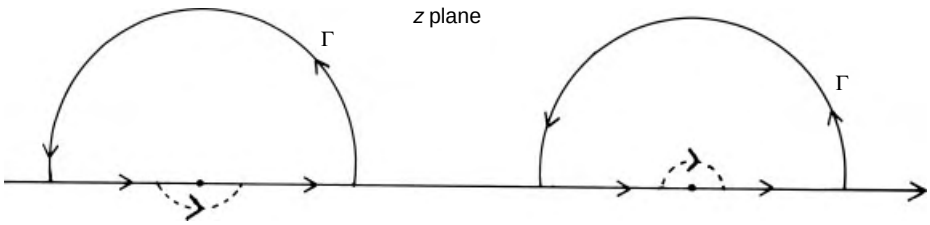


Fig. 6 Simple pole on the path

singular points z_j within Γ), then (see Fig. 5)

$$\oint_\Gamma f(z)dz = 2\pi i \sum_{j=1}^{n} a_{-1} z_j = 2\pi i \sum_{j=1}^{n} [\text{enclosed residue of } f(z)]. \tag{25}$$

For a simple pole on the path of integration (see Fig. 6), the residue theorem yields

$$\oint_C f(z)dz = \pi i a_{-1}. \tag{26}$$

The extension of the residue theorem to cases of simple poles on the path of integration is important in physics, and the residue theorem is written as

$$\oint_C f(z)dz = 2\pi i \sum_{j=1}^{n} (\text{enclosed residue}) + \pi i \sum_{k=1}^{m} (\text{residue of simple poles on path}). \tag{27}$$

The residue theorem may be used to evaluate certain definite integrals that occur when solving physical problems, and the procedure for evaluating four types of integrals will now be illustrated.

Type 1 Integrals:

$$I_1 = \int_0^{2\pi} f(\sin\theta, \cos\theta) d\theta.$$

It is assumed that the integrand $f(\sin\theta, \cos\theta)$ contains no singularities other than poles. If $z = e^{i\theta}$ (unit circle), then

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{z^2 - 1}{2iz}$$

and

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{z^2 + 1}{2z}.$$

In terms of z, the integral I_1 becomes

$$I_1 = -i \oint_{\substack{\text{unit}\\ \text{circle}}} f\left(\frac{z^2 - 1}{2iz}, \frac{z^2 + 1}{2z}\right) \frac{dz}{z} = 2\pi \sum (\text{residue within the unit circle}). \tag{28}$$

The analysis for Type 1 integrals may be used to evaluate the following integral:

$$I_1 = \int_0^{2\pi} \frac{d\theta}{5 + 4\cos\theta} = -i \oint_{\substack{\text{unit}\\ \text{circle}}} \frac{dz}{(2z+1)(z+2)} = \frac{2\pi}{3}. \tag{29}$$

Type 2 Integrals: If (a) $f(z)$ is analytic in the upper-half plane except for a finite number of enclosed singularities z_j and/or simple poles x_k on the real axis, and

(b) $zf(z)$ approaches zero as $|z|$ approaches infinity, then

$$I_2 = \int_{-\infty}^{\infty} f(x)dx$$
$$= 2\pi i \sum_{j=1}^{n} a_{-1,z_j} + \pi i \sum_{k=1}^{m} a_{-1,x_k}. \quad (30)$$

By use of the analysis for Type 2 integrals, let us evaluate

$$I_2 = \int_{-\infty}^{\infty} \frac{dx}{1+x^2}$$
$$= \oint_{\text{semicircle}} \frac{dz}{(z-i)(z+i)} = \pi. \quad (31)$$

Types 3 and 4 Integrals: If (a) $f(z)$ is analytic in the upper-half plane except at a finite number of enclosed singular points and/or simple poles on the real axis and (b) $f(z)$ approaches zeros as $|z|$ approaches infinity, then integrals of the form

$$\int_{-\infty}^{\infty} f(x)\exp(imx)\,dx$$

yield I_3 and I_4 where

$$I_3 = \int_{-\infty}^{\infty} f(x)\cos mx\,dx$$
$$= -2\pi \sum_{\text{enclosed}} \text{Im}\{\text{residue}[f(z)\exp(imz)]\}$$
$$-\pi \sum_{\substack{\text{on the}\\ \text{path}}} \text{Im}\{\text{residue}[f(z)\exp(imz)]\} \quad (32)$$

and

$$I_4 = \int_{-\infty}^{\infty} f(x)\sin mx\,dx$$
$$= 2\pi \sum_{\text{enclosed}} \text{Re}\{\text{residue}[f(z)\exp(imz)]\}$$
$$+\pi \sum_{\substack{\text{on the}\\ \text{path}}} \text{Re}\{\text{residue}[f(z)\exp(imz)]\}. \quad (33)$$

The application of Type 3 integrals is similar to that for Type 4 integrals. Type 4 may be used to evaluate

$$I_4 = \int_{-\infty}^{\infty} \frac{\sin x\,dx}{x}$$
$$= \pi \sum \text{Re}\{\text{residue}[\exp(iz)/z]\} = \pi. \quad (34)$$

1.7 The Cauchy Principal Value and Dispersion Relations

On returning to the case of a simple pole on the real axis, note that it is useful to express the result in terms of the Cauchy principal value of an integral. The integral of a function $f(x)$ which has a simple pole at $x = x_0$ for x_0 within $[a, b]$ may be written as

$$\int_a^b f(x)dx = \lim_{\varepsilon\to 0}\left\{\int_a^{x_0-\varepsilon} f(x)dx + \int_{x_0+\varepsilon}^{b} f(x)\,dx\right\}$$
$$\equiv \text{P}\int_a^b f(x)\,dx. \quad (35)$$

The symbol P in front of an integral indicates the Cauchy principal value of the integral and means carry out the limiting process in Eq. (35). Note that the Cauchy principal value may exist even if the regular value of the integral does not exist; for example, $\text{P}\int_{-1}^{1} dx/x^3 = 0$.

Dispersion relations (also known as spectral representations, Kronig-Kramers relations, and Hilbert transforms) result from the analytic properties of the complex representation of physical quantities and the Cauchy residue theorem. Originally, Kronig and Kramers were concerned with the dispersion of light and the relation between the real (dispersive) and

imaginary (absorptive) parts of the index of refraction at different frequencies. The basic idea of dispersion relations is applied in areas ranging from electronic design to quantum field theory. Here, general forms for dispersion relations will be presented. For a physical quantity $\chi(\omega)$ which approaches zero as ω approaches infinity and is analytic in the upper-half plane (see Fig. 7), consider the evaluation of the integral

$$\oint_{\Gamma} \frac{\chi(\omega)d\omega}{\omega - \omega_0}. \tag{36}$$

By use of the Cauchy residue theorem for a simple pole at ω_0 on the contour and the physical property that $\chi(\omega)$ approaches zero as ω approaches infinity, Eq. (36) yields

$$-\chi(\omega_0) = \frac{i}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{\chi(\omega)d\omega}{\omega - \omega_0}. \tag{37}$$

On equating corresponding real and imaginary parts in Eq. (37), the dispersion relations are obtained:

$$\mathrm{Re}\chi(\omega_0) = \frac{1}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{\mathrm{Im}\chi(\omega)d\omega}{\omega - \omega_0},$$
$$\mathrm{Im}\chi(\omega_0) = -\frac{1}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{\mathrm{Re}\chi(\omega)d\omega}{\omega - \omega_0}. \tag{38}$$

The equations in Eq. (38) express one part of an analytic function in terms of an integral involving the other part and are called dispersion relations. In electronics,

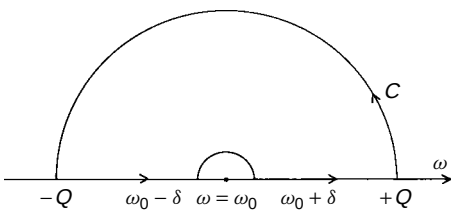


Fig. 7 Contour for Eq. (36)

one has $Z(\omega) = R(\omega) + iX(\omega)$ where Z is impedance, R is resistance, and X is reactance. Dispersion relations may be used to express resistance in terms of reactance. Dispersion relations for light (complex index of refraction $\eta = n_c + iaa$) yield relations between dispersive power and absorption. In addition, a large number of definite integrals may be evaluated by use of the dispersion relations. Dispersion relations applied to $f(z) = \cos x + i \sin x$ lead to values of integrals with integrands of forms $(\sin x)/x$ and $(\cos x)/x$ for limits of integration from minus infinity to plus infinity.

1.8 Conformal Transformations

An analytic function $w = f(z) = u(x, y) + iv(x, y)$ for $z = x + iy$ is completely characterized by two pairs of variables (x, y) and (u, v). Riemann developed a mode of visualizing the relation $w = f(z)$ which uses two separate complex planes, z plane for (x, y) and w plane for the corresponding (u, v). By use of the two-plane picture, the equation $w = f(z)$ defines the transformation (relation, correspondence, or mapping) between the two planes. That is to say, $w = f(z)$ may be used for mapping a set of points (locus, figure) in the z plane (or w plane) into the corresponding figure in the w plane (or z plane). For physical problems, the basic idea involves transforming the geometry of a complicated problem in the z plane into a simpler geometry in the w plane, solving the problem with the simpler geometry, and inverting the transformation to obtain the desired solution in the z plane. The most important class of transformations used in solving physical problems are those that preserve the angle between two straight

lines (conformal transformations). The angle-preserving property of conformal transformations will now be illustrated: Assume that two lines intersect at $z = a$ in the z plane and at $w = f(a)$ in the w plane, with elements of length along two lines given respectively by $dz_1 = |dz_1| \exp(i\theta_1)$ and $dz_2 = |dz_2| \exp(i\theta_2)$. The corresponding elements of length in the w plane are $dw_1 = |dz_1||f'(z)| \exp[i(\phi + \theta_1)]$ and $dw_2 = |dz_2| \times |f'(z)| \exp[i(\phi + \theta_2)]$ since $dw = dz|f'(z)| \times \exp(i\phi)$. Finally, note that the direction of the corresponding lines in the w plane is rotated by ϕ, but the angle between the lines in the z plane $(\theta_2 - \theta_1)$ equals the angle between the lines in the w plane $[(\phi + \theta_2) - (\phi + \theta_1)]$.

Four often used elementary transformations are the following.

1. *Translation*: $w = z + z_0$ for z_0 constant. The transformation equations are

$$\begin{aligned} w &= (x + x_0) + i(y + y_0), \\ u &= x + x_0, \\ v &= y + y_0. \end{aligned} \tag{39}$$

2. *Magnification*: $w = az$ for constant and real a. The transformation equations are

$$w = ax + iay, \;\; u = ax, \;\; \text{and } v = ay. \tag{40}$$

3. *Rotation*: $w = z_0 z$. Here one may write

$$w = \rho \exp(i\phi) = r_0 r \exp[i(\theta + \theta_0)]. \tag{41}$$

The angle in the w plane is $\phi = \theta + \theta_0$ where θ_0 is the angle of rotation and r_0 corresponds to the magnification.

4. *Inversion*: $w = 1/z$. In polar form, w may be written as

$$w = \rho \exp(i\phi) = \frac{1}{r} \exp(-i\theta). \tag{42}$$

The transformation equations for inversion are

$$\begin{aligned} u &= \frac{x}{x^2 + y^2}, \; v = -\frac{y}{x^2 + y^2}, \\ x &= \frac{u}{u^2 + v^2}, \; y = -\frac{v}{u^2 + v^2}. \end{aligned} \tag{43}$$

The following transformation is also useful in physical applications: $w = z^2$ which yields $w = \rho \exp(i\phi) = r^2 \exp(i2\theta)$or $\rho = r^2$ and $\phi = 2\theta$ with transformation equations given by $u = x^2 - y^2$ and $v = 2xy$. Here one finds that a circle with radius r_0 is mapped into a corresponding circle with radius $R = r_0^2$, and θ_0 is mapped into $2\theta_0$. In potential theory, the two-dimensional Laplace equation is to be solved with appropriate boundary conditions. Note that the transformation $w = z^2$ maps the right angle in the z plane into a straight line in the w plane (see Fig. 8) where boundary conditions may be applied more conveniently.

In connection with the transformation $w = z^2$ (and other multivalued functions), note that the transformation is conformal except at $w = 0$ which is called a branch point, and separately maps the upper- and lower-half planes of the z plane into the whole w plane (points z and $-z$ are mapped

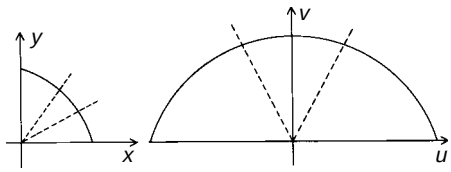


Fig. 8 Diagram for the transformation $w = z^2$

into the same points in the w plane). The inverse transformation $z = \sqrt{w}$ cannot be unique. The quantity z may be written as

$$z = \sqrt{\rho}\exp\left(i\frac{\phi}{2}\right) = \sqrt{\rho}\exp\left(i\frac{\phi_p}{2} + i\pi k\right). \tag{44}$$

Odd and even values of k in Eq. (44) yield opposite signs for z. In describing values of a unit circle about the origin in the z plane for $k = 0$, it is found that (a) $z = 1$ for $\phi_p = 0$ and (b) $z = -1$ when $\phi_p = 2\pi$. When $k = 1$, the values become (a) $z = -1$ when $\phi_p = 0$ and (b) $z = 1$ when $\phi_p = 2\pi$. One may avoid the double values by assuming a cut, which may not be crossed, from zero to infinity along the u axis in the w plane. Riemann introduced the scheme of two planes (sheets, surfaces) joined edge to edge at the cut as a way to combine cases for $k = 0$ (all evens) and $k = 1$ (all odds) and to eliminate the cut. For example, a lower sheet contains the set of values for $k = 0$ and an upper sheet contains the values for $k = 1$ (see Fig. 9).

The function $\sqrt{w}$ is analytic over the whole Riemann surface (two sheets) except at the branch point, $w = 0$. In summary, it is found that the w plane is mapped into two sheets (Riemann surface). The concept of Riemann surfaces has broad application in physics.

2 Ordinary Differential Equations

2.1 Introduction

A differential equation is an equation which contains derivative(s), and it may be either an ordinary or a partial differential equation. Ordinary differential equations contain derivative(s) with respect to one independent variable, and partial differential equations contain partial derivatives

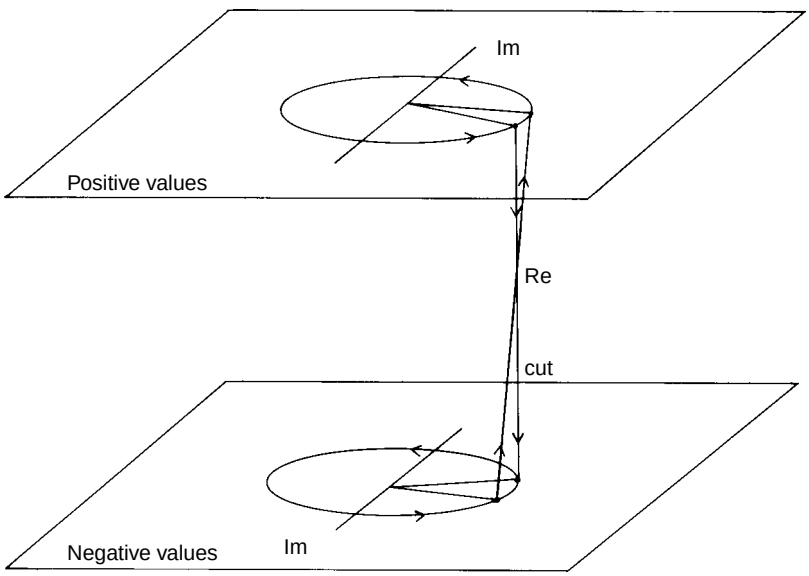


Fig. 9 Riemann surface

with respect to two or more independent variables.

The order of a differential equation is the order of the highest derivative appearing in the equation. The degree of a differential equation is the power of the highest derivative after fractional powers of all derivatives have been removed. If the dependent variable and all of its derivatives are to first power without a product of the dependent variable and a derivative, the differential equation is said to be linear. If a differential equation is not linear, it is classified as nonlinear.

Applications of appropriate physical laws to a large number of physical problems lead to differential equations. In general, a physical process is described by use of a differential equation with appropriate boundary conditions in space and/or initial conditions in time and/or an integral equation. The boundary and/or initial conditions determine from the many possible solutions the one that describes the specific physical phenomenon involved.

The main purpose here concerns the development of solutions for differential equations which adequately describe physical processes under investigation. The mathematical subjects of existence and uniqueness theorems for solutions of differential equations will not be discussed.

An elementary introduction to the subject of ordinary differential equations, as it relates to the needs in solving physical problems, can be reduced to that of treating linear (or reducible to the linear form) first- and second-order differential equations. This presentation is devoted to the construction of solutions and physical applications of such ordinary differential equations.

First- and second-order linear ordinary differential equations have the following standard forms, respectively:

$$\frac{dy}{dx} + p(x)y = Q(x) \text{ or } y' + p(x)y = Q(x) \tag{45}$$

and

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = f(x)$$

or

$$y'' + p(x)y' + q(x)y = f(x). \tag{46}$$

In Eqs. (45) and (46), the notations $y' = dy/dx$ and $y'' = d^2y/dx^2$ have been used. When time t is the independent variable, one writes $\dot{y} = dy/dt$ and $\ddot{y} = d^2y/dt^2$. If the right-hand side of Eqs. (45) or (46) equals zero, the equation is classified as homogeneous; otherwise, the equation is classified as nonhomogeneous (inhomogeneous).

2.2 First-Order Linear Differential Equations

The formulation of many physics problems leads to first-order differential equations, and this section is devoted to solutions of such problems.

2.2.1 Separable Differential Equations

Differential equations that can be put in the form $g(y)dy = f(x)\,dx$ are called separable differential equations since the left-hand side is a function of y only and the right-hand side is a function of x only. For $dy = f(x)\,dx$, the general solution is

$$y = \int f(x)dx + C. \tag{47}$$

Since the general solution of a first-order differential equation results from one integration, it will contain one arbitrary constant. Similarly, the general solution of a second-order ordinary differential equation will contain two arbitrary constants.

Values of arbitrary constants are determined by use of physical boundary or initial conditions.

EXAMPLE 2.1: In the radioactive decay of nuclei, the process is governed by the following differential equation: $dN/dt = -\lambda N$ with initial condition $N(t = 0) = N_0$. The number of parent nuclei present at time t is represented by $N(t)$, and the decay constant λ is characteristic of the particular nuclei involved. The negative sign is used to indicate that the number of nuclei decreases with time. Let us find $N(t)$ subject to the indicated initial condition.

Solution: The differential equation may be written in the form

$$\frac{dN}{N} = -\lambda\, dt$$

with general solution

$$\ln N = -\lambda t + C_1 \text{ or } N(t) = C_2 \exp(-\lambda t). \tag{48}$$

The value of the constant of integration is determined by use of the initial condition, $N(t = 0) = N_0$; the initial condition leads to $N(0) = N_0 = C_2$. The specific (particular) solution of the problem is the familiar relation

$$N(t) = N_0 e^{-\lambda t}. \tag{49}$$

2.2.2 Exact Differential Equations

The general first-order differential equation, $dy/dx = f(x, y)$, may be written in the form

$$M(x, y)dx + N(x, y)dy = 0. \tag{50}$$

The total (exact) differential of $F(x, y) = C$ (where F is continuous with continuous derivatives) is

$$dF = \left(\frac{\partial F}{\partial x}\right)_y dx + \left(\frac{\partial F}{\partial y}\right)_x dy = 0.$$

Note that the general differential equation in Eq. (50) is exact if

$$M(x, y) = \left(\frac{\partial F}{\partial x}\right)_y \text{ and}$$

$$N(x, y) = \left(\frac{\partial F}{\partial y}\right)_x. \tag{51}$$

Since it is assumed that $F(x, y)$ has continuous first derivatives, note that

$$\left(\frac{\partial M}{\partial y}\right)_x = \left(\frac{\partial N}{\partial x}\right)_y. \tag{52}$$

The condition indicated in Eq. (52) is both necessary and sufficient for Eq. (50) to be an exact differential equation.

EXAMPLE 2.2: Determine whether the following differential equation is exact and find its solution if it is exact: $(4x^3 + 6xy + y^2) \times dx/dy = -(3x^2 + 2xy + 2)$.

Solution: The standard form of this differential equation is $(4x^3 + 6xy + y^2)dx + (3x^2 + 2xy + 2)dy = 0$; it is exact since the condition in Eq. (52) is satisfied. The solution of the original differential, therefore, has the form $F(x, y) = C$. The function $F(x, y)$ is obtained as follows:

$$\frac{\partial F}{\partial x} = 4x^3 + 6xy + y^2$$

or

$$F(x, y) = x^4 + 3x^2y + y^2x + f(y) \tag{53}$$

and

$$\frac{\partial F}{\partial y} = 3x^2 + 2xy + 2$$

or

$$F(x, y) = 3x^2y + y^2x + 2y + g(x). \tag{54}$$

Functions $f(y)$ and $g(x)$ arise from integrating with respect to x and y, respectively. For consistency, it is required

that $f(y) = 2y$ and $g(x) = x^4$. The solution of the original differential equation is $x^4 + 3x^2y + xy^2 + 2y = C$.

2.2.3 Solution of the General Linear Differential Equation

A good feature of first-order linear differential equations is that the general equation in this category can be solved. It can be shown that the general solution of Eq. (45) may be obtained from the formula

$$y(x) = \exp\left(-\int p(x)\,dx\right) \times \int Q(x)\exp\left(\int p(x)\,dx\right)dx + C\exp\left(-\int p(x)\,dx\right). \qquad (55)$$

If the first-order linear differential equation is separable, the method of Sec. 2.2.1 for separable equation should be followed, and the method of Sec. 2.2.2 for exact differential equations yields solutions for exact differential equations. The formula in Eq. (55) will now be applied to obtain the general solution of the differential equation generated by applying Kirchhoff's loop method to the circuit in Fig. 10.

EXAMPLE 2.3: The appropriate differential equation and initial condition for the circuit in Fig. 10 are

$$L\frac{dI}{dt} + RI = E, \text{ where } I(0) = 0. \qquad (56)$$

On applying the formula in Eq. (55) for $p(x) = R/L$ and $Q(x) = E/L$ and initial condition, the solution of Eq. (56) reduces to

$$I(t) = \frac{E}{R}\left[1 - \exp\left(\frac{-Rt}{L}\right)\right]. \qquad (57)$$

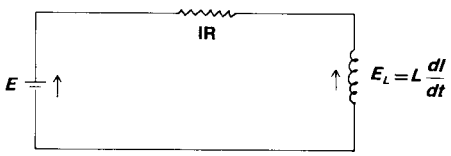


Fig. 10 Diagram for Example 2.3

Differential equations of the form $y' + p(x)y = Q(x)y^n$ where $n > 1$ (Bernoulli's equation) often occur in physical problems. Bernoulli's type of nonlinear first-order differential equation can be reduced to the linear form by use of the transformation $z = y^{1-n}$, $z' + (1-n)p(x)z = (1-n)Q(x)$. The differential equation with dependent variable z can be solved by use of the formula in Eq. (55).

EXAMPLE 2.4: The motion of a particle in a viscous fluid with Stokes damping, av, and Newtonian damping, bv^2, is characterized by an equation of motion of the form $\dot{v} + av = -\beta v^2$ subject to $v(0) = v_0$. This equation of motion is of the Bernoulli form where $n = 2$, $Q(t) = -\beta$, and $p(t) = \alpha$, and the general solution is obtained from

$$z = v^{-1} = \frac{-\beta}{\alpha} + C\exp(\alpha t). \qquad (58)$$

The particular solution is

$$v(t) = \frac{\alpha v_0}{(\alpha + v_0\beta)\exp(\alpha t) - \beta v_0}. \qquad (59)$$

A graph of Eq. (59) characterizes the speed of the particle as a function of time. To obtain position as a function of time, replace v with $\dot{x}$ and solve the resulting first-order differential equation for $x(t)$.

2.3 Second-Order Linear Differential Equations

The superposition of solutions principle, stated here in the form of two theorems, will be assumed valid for second-order linear homogeneous differential equations.

THEOREM 1: The set of all solutions of an nth-order linear homogeneous differential equation forms an n-dimensional vector space (see ALGEBRAIC METHODS).

For a second-order differential equation, Theorem 1 means that $y = y_1 + y_2$ is a solution of $y'' + p(x)y' + q(x)y = 0$ if y_1 and y_2 are two linearly independent solutions of the original differential equation.

THEOREM 2: A necessary and sufficient condition that solutions y_1 and y_2 of a second-order linear differential equation be linearly independent is that the Wronskian of these solutions be different from zero. The Wronskian of y_1 and y_2 is the determinant with elements $a_{11} = y_1$, $a_{12} = y_2$, $a_{21} = y_1'$, and $a_{22} = y_2'$.

2.3.1 Homogeneous Differential Equations with Constant Coefficients

The standard form for the general second-order homogeneous differential equation with constant coefficients is $y'' + p_0 y' + q_0 y = 0$ which may be written as $(D^2 + p_0 D + q_0)y = 0$ where $D = d/dx$. The procedure for solving differential equations in this category involves treating $D^2 + p_0 D + q_0 = 0$, the auxiliary or characteristic equation, algebraically and using techniques for solving first-order differential equations. The roots of the auxiliary equation (quadratic) may be real and unequal, real and equal, or a complex-conjugate pair. For real and unequal roots a and b of the auxiliary equation, the differential equation may be written in the symbolic form $(D - a)u = 0$ where $u = (D - b)y$. The form of the general solution becomes $y(x) = c_1 e^{ax} + c_2 e^{bx}$ when the two indicated first-order differential equations are solved. If the roots of the auxiliary equation are a complex-conjugate pair $a^* = b$, the solution of the differential equation has the same form as the case for real and unequal roots with a^* replacing b. The solution of the differential equation for real and equal roots of the auxiliary equation is obtained from solving the two indicated first-order differential equations $(D - a)u = 0$ where $u = (D - a)y$; the form of the general solution is $y(x) = (c_1 x + c_2)e^{ax}$.

EXAMPLE 2.5: Consider the motion of a particle of mass m initially at rest and subject to a restoring force of $-kx$ and a damping force of $-a\dot{x}$. The equation of motion of this particle is $m\ddot{x} = -kx - a\dot{x}$. The equation of motion in standard form is $\ddot{x} + 2\delta\dot{x} + \omega^2 x = 0$ where $\omega^2 = k/m$ and $2\delta = a/m$ (the factor 2 is used for convenience). Find the solution of the equation of motion for the following cases:

1. $\delta = 0$ (no damping);
2. $\delta = \omega$ (critical damping);
3. $\delta < \omega$ (light damping); and
4. $\delta > \omega$ (heavy damping).

Case 1: The equation of motion for $\delta = 0$ reduces to $\ddot{x} + \omega^2 x = 0$ with solution

$$\begin{aligned} x(t) &= c_1 e^{i\omega t} + c_2 e^{-i\omega t} \\ &= A\cos\omega t + B\sin\omega t \\ &= X_0 \cos\omega t \text{ for } x(0) = X_0 \\ &\quad \text{and } \dot{x}(0) = 0. \end{aligned} \tag{60}$$

The motion without damping is oscillatory and periodic (with constant amplitude X_0).

Case 2: For $\delta \neq 0$ and $\omega \neq 0$, the solutions of the corresponding auxiliary equation are $-\delta + \Delta$ and $-\delta - \Delta$ where $\Delta = \sqrt{\delta^2 - \omega^2}$. The solution of the equation of motion for critical damping $\delta = \omega$ is

$$\begin{aligned} x(t) &= (c_1 t + c_2)e^{-\delta t} \\ &= X_0(\delta t + 1)e^{-\delta t} \\ \text{for } x(0) = X_0 &\quad \text{and} \quad \dot{x}(0) = 0. \end{aligned} \tag{61}$$

Here the motion is not oscillatory and approaches equilibrium at a rapid rate.

Case 3: The solution for light damping $\delta < \omega$ using $\Delta' = \sqrt{\omega^2 - \delta^2}$ is

$$x(t) = (A\cos\Delta' t + B\sin\Delta' t)e^{-\delta t}$$
$$= X_0\left(\cos\Delta' t + \frac{\delta}{\Delta'}\sin\Delta' t\right)e^{-\delta t}$$
$$\text{for } x(0) = X_0 \quad \text{and} \quad \dot{x}(0) = 0. \tag{62}$$

In this case, the motion is oscillatory with decreasing amplitude (not periodic).

Case 4: The solution for heavy damping $\delta > \omega$ is

$$x(t) = \frac{(\delta + \Delta)}{2\Delta}X_0\exp[(-\delta + \Delta)t] + \frac{(\Delta - \delta)}{2\Delta}X_0\exp[-(\delta + \Delta)t]. \tag{63}$$

The motion in this case is not oscillatory and approaches equilibrium at a rate less rapid than for critical damping.

2.3.2 Nonhomogeneous Differential Equations with Constant Coefficients

The standard form for second-order nonhomogeneous differential equations with constant coefficients is $y'' + p_0 y' + q_0 y = f(x)$, and the two widely used methods for solving differential equations in this category are (a) $y = y_h + y_p$ where y_h is the solution of the corresponding homogeneous equation and y_p is any solution of the original nonhomogeneous differential equation, and (b) successive integration. The method of successive integration involves writing the differential equation in the factored form $(D - a)u = f(x)$ where $u = (D - b)y$, and solving the two indicated first-order differential equations.

Physical problems are often solved by use of the first method since y_p can often be obtained without difficulty. Systematic methods for finding y_p for three types of nonhomogeneous terms (polynomial, exponential, and sine and/or cosine) will now be given.

1. The nonhomogeneous term $f(x)$ is a polynomial of degree $n \geq 0$.
   (A) If zero is not a root of the characteristic equation, then assume
   $$y_p = A_0 + A_1 x + \cdots + A_n x^n.$$
   (B) If zero is a single root of the characteristic equation, then assume
   $$y_p = x(A_0 + A_1 x + \cdots + A_n x^n).$$
   (C) If zero is a double root of the characteristic equation, then assume
   $$y_p = x^2(A_0 + A_1 x + \cdots + A_n x^n).$$
2. The nonhomogeneous term $f(x)$ is of the form $C\exp(kx)$.
   (A) If k is not a root of the characteristic equation, then assume
   $$y_p = A\exp(kx).$$
   (B) If k is a single root of the characteristic equation, then assume
   $$y_p = Ax\exp(kx).$$
   (C) If k is a double root of the characteristic equation, then assume
   $$y_p = Ax^2\exp(kx).$$
3. The nonhomogeneous term $f(x)$ is of the form $\sin kx$, $\cos kx$, or $\sin kx + \cos kx$.
   (A) If ik is not a root of the characteristic equation, then assume
   $$y_p = A\cos kx + B\sin kx.$$

(B) If ik is a single root of the characteristic equation, then assume

$$y_p = Ax\cos kx + Bx\sin kx.$$

Values for constants in the assumed expression for y_p are obtained when that expression is substituted into the original nonhomogeneous differential equation.

EXAMPLE 2.6: The equation of motion for a mass attached to the end of a vertical spring fixed at the other end is $\ddot{y} + \omega^2 y = -g$ where g is the acceleration due to gravity. The general solution of the homogeneous equation, $\ddot{y} + \omega^2 y = 0$, is $y_h = A\cos\omega t + B\sin\omega t$. By use of inspection, it is clear that $y_p = -g/\omega^2$ is a solution of the original nonhomogeneous equation. The solution of the equation of motion for $y(0) = Y_0$ and $\dot{y}(0) = 0$ is

$$\begin{aligned} y(t) &= y_h + y_p \\ &= A\cos\omega t + B\sin\omega t - \frac{g}{\omega^2} \\ &= \left(Y_0 + \frac{g}{\omega^2}\right)\cos\omega t - \frac{g}{\omega^2}. \end{aligned} \quad (64)$$

A graph of Eq. (64) characterizes the motion, position as a function of time, of this particle.

2.3.3 Homogeneous Differential Equations with Variable Coefficients

The general procedure used to solve differential equations of the form $y'' + p(x)y' + q(x)y = 0$ is the power-series method. The power-series method due to Frobenius and Fuchs yields the following two kinds of information concerning the nature of the solution for $x \neq 0$: form of the solution as a result of the nature of $p(x)$ and $q(x)$, and form of the solution as indicated by the nature of the solution of the indicial equation. As normally needed in solving physical problems, the general form of the power series solution is

$$y(x) = \sum_{\lambda=0}^{\infty} a_\lambda x^{\lambda+k} \text{ for } a_0 \neq 0. \quad (65)$$

EXAMPLE 2.7: Consider the differential equation $xy'' + 2y' + xy = 0$. By use of the power-series method, obtain the indicial equation and its two solutions, recursion formula, and general solution of the differential equation. On substituting Eq. (65) into the differential equation to be solved, one obtains

$$\sum_{\lambda=0}^{\infty} a_\lambda(\lambda + k + 1)(\lambda + k)x^{\lambda+k-2} + \sum_{\lambda=0}^{\infty} a_\lambda x^{\lambda+k} = 0. \quad (66)$$

The basic plan at this stage is to write the result using a single sum. On replacing λ with $\lambda' + 2$ in the first sum, the power of x in the first sum becomes the same as that in the second sum. Equation (66) now becomes

$$a_0 k(k+1)x^{k-2} + a_1(k+1)(k+2)x^{k-1} + \sum_{\lambda=0}^{\infty}\{a_{\lambda+2}(\lambda + k + 3)(\lambda + k + 2) + a_\lambda\} \times x^{\lambda+k} = 0. \quad (67)$$

Since terms in Eq. (67) are linearly independent, it is required that

$$a_0 k(k+1) = 0 \text{ (indicial equation)}, \quad (68)$$

$$a_1(k+1)(k+2) = 0, \quad (69)$$

and

$$a_{\lambda+2}(\lambda + k + 3)(\lambda + k + 2) + a_\lambda = 0 \quad \text{(recursion formula)}. \quad (70)$$

The indicial equation results from equating the coefficient of the lowest power of the variable to zero. In this case, the solutions of the indicial equation are $k = 0$ and $k = -1$. When $k = 0$, $a_1 = 0$ because of Eq. (69). The coefficient a_1 is arbitrary when $k = -1$, and two independent solutions of the original differential equation may be obtained by use of $k = -1$ since a_0 is arbitrary by hypothesis. The form of the solution becomes

$$y(x) = \sum_{\lambda=0}^{\infty} a_\lambda x^{\lambda-1}. \tag{71}$$

Coefficients in Eq. (71) are obtained from the recursion formula using $k = -1$. The general expressions for even and odd expansion coefficients, respectively, are

$$a_{2j} = \frac{(-1)^j a_0}{(2j)!},$$
$$a_{2j+1} = \frac{(-1)^j a_1}{(2j+1)!}, \quad j = 0, 1, 2, \ldots. \tag{72}$$

The general solution of the original differential equation is obtained by substituting coefficients in Eq. (72) into Eq. (71).

2.3.4 Nonhomogeneous Differential Equations with Variable Coefficients

Variation of parameters and Green's-function methods are normally used to solve nonhomogeneous linear differential equations with variable coefficients that occur in physics. The standard form for these differential equations is

$$y'' + p(x)y' + q(x)y = f(x). \tag{73}$$

The method of variation of parameters due to Lagrange will now be used to solve Eq. (73) subject to the conditions given below. Assume the solution has the form

$$\begin{aligned} y(x) &= C_1 y_1 + C_2 y_2 \\ &= v_1(x) y_1 + v_2(x) y_2. \end{aligned} \tag{74}$$

In Eq. (74), y_1 and y_2 are two linearly independent solutions of the corresponding homogeneous differential equation, and constant parameters C_1 and C_2 are replaced with functions v_1 and v_2. Functions v_1 and v_2 are unknown parameters to be determined. If $v_1' y_1 + v_2' y_2 = 0$, and $f(x)$ is continuous in the region of interest, then the solution of the original differential equation, Eq. (73), is obtained by use of

$$\begin{aligned} y(x) = -y_1 \int \frac{f(x) y_2 dx}{W(y_1, y_2)} \\ + y_2 \int \frac{f(x) y_1 dx}{W(y_1, y_2)}. \end{aligned} \tag{75}$$

The quantity $W(y_1, y_2)$ is the Wronskian of y_1 and y_2.

On using Eq. (75) to solve $y'' - (2y'/x) + (2y/x^2) = (\ln x)/x$ for $x \neq 0$, it is found that $y_1 = x$ and $y_2 = x^2$ are two linearly independent solutions of the corresponding homogeneous equation, the Wronskian equals x^2, and the solution becomes

$$\begin{aligned} y(x) = -x \left[\frac{(\ln x)^2}{2} + \ln x + 1 \right] \\ - C_1 x + C_2 x^2. \end{aligned}$$

Equation (75) will now be put in the form of a definite integral that is useful in solving initial or boundary value problems. Let x be a point in the closed interval $[a, b]$ such that the first term in Eq. (75) is replaced by a definite integral from x to b and the second term in Eq. (75) is replaced by a definite integral from a to x. In terms of the indicated two definite integrals, Eq. (75) becomes

$$y(x) = \int_a^x \frac{y_1(t)y_2(x)f(t)dt}{W(t)} + \int_x^b \frac{y_1(x)y_2(t)f(t)dt}{W(t)} = \int_a^b G(x,t)f(t)\,dt. \tag{76}$$

The function $G(x, t)$ in Eq. (76) is called the Green's function for Eq. (73) subject to the appropriate boundary conditions. The Green's function is defined by

$$G(x,t) = \begin{cases} \dfrac{y_1(t)y_2(x)}{W(t)} \equiv G_1 & \text{for } a \le t \le x, \\ \dfrac{y_1(x)y_2(t)}{W(t)} \equiv G_2 & \text{for } x \le t \le b. \end{cases} \tag{77}$$

Note that the Green's function depends only on y_1, y_2, and the Wronskian. The quantity $W(t)$ means $W(y_1(t), y_2(t))$. The value of the Green's-function approach is related to the fact that initial or boundary conditions are incorporated in the formulation of the problem in a natural manner. At $t = a$, $G_1(x, t)$ satisfies the boundary condition imposed on $y(x)$, and $G_2(x, t)$ satisfies the boundary condition for $y(x)$ at $t = b$. On applying the Green's function method to solve $y'' = 6x$ subject to $y(0) = y(1) = 0$, it is found that $y_1 = x$ and $y_2 = x - 1$ are two linearly independent solutions of the homogeneous equation, the Wronskian equals unity, the Green's functions become $G_1(x, t) = t(x - 1)$ for $0 \le t \le x$ and $G_2(x, t) = x(t - 1)$ for $x \le t \le 1$, and the solution of the differential equation is $y(x) = \int_0^1 G(x, t)6t\,dt = x^3 - x$.

2.4 Some Numerical Methods for Ordinary Differential Equations

Numerical methods are treated in detail elsewhere in this book (see NUMERICAL METHODS), and a summary of essential features related to solutions of ordinary differential equations is given in this section. In general, the numerical solution of a differential equation consists of a table of values of the dependent variable for corresponding values of the independent variable.

2.4.1 The Improved Euler Method for First-Order Differential Equations

The basic idea of Euler's method for solving first-order ordinary differential equations is to convert the differential equation (continuous) to a difference equation (discrete). The general form for a first-order ordinary differential equation will now be written as

$$\frac{dy}{dx} = f(x, y). \tag{78}$$

By use of the definition of a derivative, one may write

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \left(\frac{y(x + \Delta x) - y(x)}{\Delta x} \right) = f(x, y). \tag{79}$$

The scheme of the finite difference method involves writing Eq. (79) as

$$y(x_{n+1}) = y(x_n) + f(x_n, y_n)\Delta x. \tag{80}$$

Equation (80) is the Euler algorithm for solving first-order ordinary differential equations. The notations in Eq. (80) have the following meanings: $x_{n+1} = x_n + \Delta x$, and $y(x_{n+1}) = y_{n+1}$. To apply Euler's method, select the interval size Δx, evaluate $y(x)$ at x_0, and evaluate $f(x, y)$ at x_0, y_0; the result for $y(x_1)$ is

$$y(x_1) = y(x_0) + f(x_0, y_0)\Delta x. \tag{81}$$

A second iteration with inputs $y(x_1)$ from Eq. (81) and $f(x_1, y_1)$ yields $y(x_2)$; the result is

$$y(x_2) = y(x_1) + f(x_1, y_1)\Delta x. \tag{82}$$

The iteration is continued to yield a numerical solution of the required first-order ordinary differential equation in the region of interest. A systematic procedure for calculating the error involved during each iteration does not exist for Euler's method.

To improve the simple Euler method, the class of first-order differential equations is restricted to those whose solutions can be expanded in a Taylor series. Neglecting terms of order $O((\Delta x)^3)$, one obtains

$$\begin{aligned} y(x_{n+1}) = y(x_n) &+ \Delta x f(x_n, y_n) \\ &+ \frac{(\Delta x)^2}{2}\left(\frac{\partial f(x_n, y_n)}{\partial x}\right. \\ &\left. + f(x_n, y_n)\frac{\partial f(x_n, y_n)}{\partial y}\right). \end{aligned} \tag{83}$$

Equation (83) is referred to as the improved Euler method and will be used to obtain the solution of first-order ordinary differential equations.

EXAMPLE 2.8: The equation of motion for a certain particle is $\dot{v} + \alpha v = g$ where $\alpha = 0.01\ \mathrm{s}^{-1}$, $g = 9.8\ \mathrm{m\ s}^{-2}$, and $v(0) = 0$. The analytical solution of this equation of motion is

$$v(t) = \frac{g}{\alpha}(1 - e^{-\alpha t}). \tag{84}$$

Find the numerical solution of this equation of motion by use of the improved Euler method.

Solution: The general form of the improved Euler method for the differential equation $\dot{v} = g - \alpha v$ is

$$\begin{aligned} v(t_{n+1}) = v(t_n) &+ f(t_n, v_n)\Delta t + \frac{(\Delta x)^2}{2} \\ &\times \left(\frac{\partial f(t_n, v_n)}{\partial t} + f(t_n, v_n)\frac{\partial f(x_n, v_n)}{\partial v}\right). \end{aligned} \tag{85}$$

For arbitrary Δt in Eq. (85), the quantities reduce to

$$\begin{aligned} f(t_n, v_n) &= g - \alpha v(t_n), \\ \frac{\partial f(t_n, v_n)}{\partial t} &= 0, \\ \frac{\partial f(t_n, v_n)}{\partial v} &= -\alpha. \end{aligned} \tag{86}$$

The essential programming statement for calculating the numerical solution of the original differential equation is

$$\begin{aligned} v(n+1) = v(n) &+ [g - \alpha v(n)] \\ &\times \left[\Delta t - \frac{\alpha(\Delta t)^2}{2}\right]. \end{aligned} \tag{87}$$

2.4.2 The Runge–Kutta Method for First-Order Differential Equations

There exist many methods for finding numerical solutions of first-order ordinary differential equations, and the fourth-order Runge–Kutta method is probably the most often used method. As with the Euler and the improved Euler methods, the essential problem is to generate a table of values for x and y for the differential equation $y' = f(x, y)$ when $y(x)$ at $x = x_0$ is given. The task is to develop a method for finding y_1 at $x_0 + \Delta x$, y_2 at $x_0 + 2\Delta x$, and successive values for y_n throughout the range of interest. For calculating successive values of $y(x_n)$ in the differential equation $y' = f(x, y)$, Runge–Kutta methods use a recurrence formula in the form

$$y_{i+1} = y_i + \Delta x \sum_{i=1}^{n} a_i k_i. \tag{88}$$

Of the many parameters a_i and k_i in Eq. (88), some are chosen arbitrarily and others are obtained by use of the Taylor series involving one and two variables. The order of the Runge–Kutta approximation is indicated by the value of n in Eq. (88). Evaluation of the parameters in Eq. (88) for $n > 4$ in the Runge–Kutta approximation is straightforward but involves tedious algebraic manipulations. For $h = \Delta x$, the formula for the fourth-order Runge–Kutta method reduces to

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) + O(h^5). \tag{89}$$

The parameters in Eq. (89) are determined by use of

$$k_1 = hf(x_i, y_i),$$
$$k_2 = hf\left(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}\right),$$
$$k_3 = hf\left(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}\right),$$
$$k_4 = hf(x_i + h, y_i + k_3).$$

EXAMPLE 2.9: Find the numerical solution of the differential equation in Example 2.8 by use of the fourth-order Runge–Kutta method.

Solution: The general form of the Runge–Kutta method for the differential equation $\dot{v} = g - \alpha v$ is

$$v(n+1) = v(n) + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \tag{90}$$

The k parameters reduce to $k_1 = h[g - \alpha v(n)]$, $k_2 = h\{g - \alpha[v(n + h/2) + k_1/2]\}$, $k_3 = h\{g - \alpha[v(n + h/2) + k_2/2]\}$, and $k_4 = h\{g - \alpha[v(n + h) + k_3]\}$.

2.4.3 Second-Order Differential Equations

Numerical solutions of second-order differential equations are obtained by first reducing them to a system of first-order differential equations and applying the methods for solving first-order differential equations. The general second-order differential equation may be written as

$$\frac{d^2y}{dx^2} = f(x, y, y'). \tag{91}$$

For $z = dy/dx$, Eq. (91) reduces to the following pair of first-order differential equations:

$$\frac{dz}{dx} = f(x, y, z) \text{ and}$$
$$\frac{dy}{dx} = z = g(x, y, z). \tag{92}$$

The procedure for solving Eq. (91) is to solve the first equation in Eq. (92) with condition $y'(0)$ and use that result as an input for the second equation in Eq. (92) to obtain the solution $y(x)$ with condition $y(0)$.

3
Partial Differential Equations

3.1
Introduction

Physical problems involving two or more independent variables are often described by use of partial differential equations. Partial differential equations contain partial derivatives with respect to two or more independent variables. The procedures for determining order, degree, whether linear or nonlinear, and whether homogeneous or nonhomogeneous for partial differential equations are the same as for ordinary differential equations. Some methods for solving partial differential equations are direct integration, characteristics, separation of variables, Fourier and Laplace transforms, and Green's functions. Appropriate boundary (space) and/or initial (time) conditions must be applied to the general solution of a partial differential equation to obtain a suitable solution for the problem under investigation. Three common types of boundary conditions are Dirichlet, specification of the solution at each point on the boundary; Neumann, specification of the normal derivative of the solution at each point on the boundary; and Cauchy, specification of both initial value(s) and the Dirichlet or Neumann condition.

The following equations are examples of important partial differential equations in physics involving the Laplacian operator, $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$:

$$\nabla^2 u = 0; \text{ Laplace's equation.} \tag{93}$$

The function $u(x, y, z)$ in Eq. (93) may represent electric potential in a charge-free region, gravitational potential in a region free of matter, or steady-state (time-independent) temperature in a region without a heat source.

$$\nabla^2 u = f(x, y, z); \quad \text{Poisson's equation.} \tag{94}$$

The function $u(x, y, z)$ in Eq. (94) may represent electric potential, gravitational potential, or steady-state temperature in regions with respective sources denoted by $f(x, y, z)$.

$$\nabla^2 u = \frac{1}{\sigma}\frac{\partial u}{\partial t}; \text{ heat conduction (or diffusion) equation.} \tag{95}$$

In Eq. (95), the function $u(x, y, z, t)$ may represent a time-dependent temperature in a region without a heat source or concentration of a diffusing substance. The constant σ is called the diffusivity.

$$\nabla^2 u = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2}; \quad \text{mechanical wave equation.} \tag{96}$$

The function $u(x, y, z, t)$ in Eq. (96) may represent the motion of a vibrating string or membrane, and v is the speed of the wave motion.

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V(x, y, z)\right\}\Psi = i\hbar\frac{\partial \Psi}{\partial t}; \quad \text{Schrödinger's equation.} \tag{97}$$

Schrödinger's wave equation is the basic equation of motion of a microscopic particle of mass m, and $\Psi(x, y, z, t)$ is called the wave function. The potential energy of the particle is represented by $V(x, y, z)$, and other quantities in this equation have their usual meaning.

This section on partial differential equations is mainly devoted to the physical

applications of linear second-order homogeneous partial differential equations in two independent variables; the general form for equations in this category is

$$A\frac{\partial^2 u}{\partial x^2} + 2B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = 0. \quad (98)$$

In Eq. (98), the coefficients may be functions of x and y, and properties of the solution of the differential equation depend on the relative magnitudes of the coefficients. Based on the coefficients, partial differential equations are classified as elliptic, hyperbolic, or parabolic for $AC - B^2$ greater than zero, less than zero, or equal to zero, respectively. This classification is related to the general equation of a conic section ($Ax^2 + 2Bxy + Cy^2 = 1$) representing an ellipse, a hyperbola, or a parabola. According to these classifications, note that Laplace's equation is elliptic, the mechanical wave equation is hyperbolic, and the heat conduction (diffusion) and Schrödinger equations are parabolic. The geometrically related classifications are not of primary importance when solving the differential equation by use of analytical methods but do reflect the nature of the boundary conditions. Solutions of elliptic equations must satisfy conditions on a closed boundary. In this section, the focus will be on separation of variables and Fourier transforms as methods for solving the partial differentials involved in physical applications. The method of separation of variables is illustrated in the following four sections. The Fourier transform method is presented in Sec. 3.6, and Sec. 3.7 is devoted to the Green's-function method for three-dimensional problems.

3.2 The Time-Independent Schrödinger Wave Equation

The method of separation of variables will now be used to obtain the time-independent Schrödinger wave equation. Assuming that $\Psi(x, y, z, t) = \psi(x, y, z)T(t)$ in Eq. (97) and dividing both sides of the resulting equation by ψT, the result obtained is

$$\frac{1}{\psi}\left(-\frac{\hbar^2}{2m}\right)\nabla^2\psi + V(x, y, z) = \frac{i\hbar}{T}\frac{dT}{dt} \equiv E. \quad (99)$$

Since the left-hand side of Eq. (99) is a function of space only and the right-hand side is a function of time only (time has been separated from the space variables), each side must equal a constant (separation constant) that is independent of space and time. The separation constant is a physical parameter when solving physical problems and has the dimensions of energy, denoted by E, in Eq. (99). Equation (99) leads to

$$T(t) = C\exp\left(\frac{-iEt}{\hbar}\right), \quad (100)$$

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(x, y, z)\right)\psi = E\psi. \quad (101)$$

Equation (101) is the time-independent (steady-state) Schrödinger wave equation. Analyses of solutions of Eq. (101) for various potentials and use of fundamental postulates of quantum theory form the major part of the study of quantum mechanics.

3.3 One-Dimensional Mechanical Wave Equation

Here the one-dimensional mechanical wave equation characterizing the motion

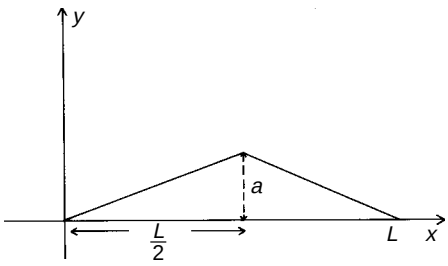


Fig. 11 Initial configuration of the string

of a string fixed at the ends $u(0, t) = u(L, t) = 0$ with initial configuration such that $u(x, 0) = 2hx/L$ for x in the closed interval $[0, L/2]$ and $u(x, 0) = 2h(L - x)/L$ for x in the closed interval $[L/2, L]$ is solved. The string is initially at rest which means that the partial derivative of $u(x, t)$ with respect to t evaluated at $t = 0$ equals zero, $u_t(x, 0) = 0$ (see Fig. 11). The method of separation of variables is applied to the equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2}. \tag{102}$$

Assume $u(x, t) = X(x)T(t)$ in Eq. (102) and divide the resulting equation by XT. The result is

$$\frac{1}{X}\frac{d^2X}{dx^2} = \frac{1}{v^2}\frac{1}{T}\frac{d^2T}{dt^2}. \tag{103}$$

Since the left-hand side of Eq. (103) is a function of x only and the right-hand side is a function of time only, the two sides must equal a constant (separation constant). The separation constant is denoted by $-\lambda^2$. The square is used for convenience as will be seen below. The negative sign is selected since an oscillatory solution is anticipated. Boundary conditions, however, will determine the required sign for the separation constant. Equation (103) leads to the following two ordinary differential equations:

$$\frac{d^2X}{dx^2} + \lambda^2 X = 0 \text{ with solution}$$
$$X(x) = A\cos\lambda x + B\sin\lambda x \tag{104}$$

and

$$\frac{d^2T}{dt^2} + \lambda^2 v^2 T = 0 \text{ with solution}$$
$$T(t) = C\cos\lambda vt + D\sin\lambda vt. \tag{105}$$

The general solution of Eq. (102) is

$$u(x, t) = (A\cos\lambda x + B\sin\lambda x)(C\cos\lambda vt + D\sin\lambda vt). \tag{106}$$

Boundary and initial conditions will now be used to determine the values of the arbitrary constants in Eq. (106). The first end-point condition $u(0, t) = 0$ in Eq. (106) leads to $A = 0$. The second end-point condition $u(L, t) = 0$ requires that $\sin\lambda L = 0$ for a nontrivial solution or $\lambda_n = n\pi/L$ where n ranges from unity to infinity. The solution now reduces to

$$u(x, t) = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi x}{L}\right) \times \left[C_n\cos\left(\frac{n\pi vt}{L}\right) + D_n\sin\left(\frac{n\pi vt}{L}\right)\right]. \tag{107}$$

Condition $u_t(x, 0) = 0$ substituted into the partial derivative of $u(x, t)$ with respect to t requires that $D_n = 0$ for all n, and the resulting solution becomes

$$(x, t) = \sum_{n=1}^{\infty} B'_n \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi vt}{L}\right);$$
$$B'_n = B_n C_n. \tag{108}$$

The B'_n coefficients in Eq. (108) are evaluated by use of the Fourier sine series. A detailed discussion of the Fourier series method is given elsewhere in this book (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). Here a summary

of Fourier series concepts needed in solving boundary valued problems is presented.

The Fourier representation of $f(x)$ in the closed interval $[-L, L]$ is

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\frac{n\pi x}{L}\right]. \tag{109}$$

Coefficients in Eq. (109) are determined by use of (Euler's formulas)

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi x}{L}\right)\,dx, \quad n = 0, 1, 2, \ldots, \tag{110}$$

and

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi x}{L}\right)\,dx, \quad n = 1, 2, 3, \ldots. \tag{111}$$

Equation (109) is valid in $[-L, L]$ when $f(x)$ is single valued, is bounded, has at most a finite number of maxima and minima, and has at most a finite number of discontinuities. If $f(x)$ is an even function, $f(x) = f(-x)$, the Fourier cosine series results, and the Fourier sine series results when $f(x)$ is odd, $f(x) = -f(-x)$.

The final condition for Eq. (108), $u(x, 0) = 2hx/L$ for $[0,L/2]$ and $u(x, 0) = 2h(L - x)/L$ for $[L/2,L]$, leads to a Fourier sine series from which the B'_n may be obtained. The expression for the B'_n coefficients is

$$\begin{aligned} B'_n &= \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx \\ &= \frac{4h}{L^2}\left[\int_0^{L/2} x\sin\left(\frac{n\pi x}{L}\right)dx + L\int_{L/2}^{L}\sin\left(\frac{n\pi x}{L}\right)dx - \int_{L/2}^{L} x\sin\left(\frac{n\pi x}{L}\right)dx\right] \\ &= \frac{8h}{n^2\pi^2}\sin\left(\frac{n\pi}{2}\right) \text{ for } n \text{ odd}, \\ &= 0 \text{ for } n \text{ even}. \end{aligned} \tag{112}$$

The particular solution of Eq. (102) reduces to

$$u(x, t) = \frac{8h}{\pi^2}\sum_{\text{odd}}^{\infty}\left[\frac{(-1)^{(n-1)/2}}{n^2} \times \sin\left(\frac{n\pi x}{L}\right)\cos\left(\frac{n\pi vt}{L}\right)\right]. \tag{113}$$

The motion of the string is such that only odd-harmonics occur and is symmetrical about the midpoint.

3.4 One-Dimensional Heat Conduction Equation

The method of separation of variables will now be applied to solve the one-dimensional heat conduction equation for the temperature distribution $u(x, t)$ in a rod of length L such that $u(0, t) = u(L, t) = 0$ and $u(x, 0) = T_0 \times \exp(-ax^2)$. The one-dimensional heat conduction equation is

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\sigma}\frac{\partial u}{\partial t}. \tag{114}$$

In Eq. (114), substitute $u(x, t) = X(x)T(t)$ and divide the resulting equation by XT; the resulting two ordinary differential equations for separation constant $-\lambda^2$ are

$$\frac{d^2X}{dx^2} + \lambda^2 X = 0 \text{ with solution}$$
$$X(t) = A\cos\lambda x + B\sin\lambda x \tag{115}$$

and

$$\frac{dT}{dt} + \lambda^2\sigma T = 0 \text{ with solution}$$

$$T(t) = C\exp(-\lambda^2\sigma t). \qquad (116)$$

The general solution of Eq. (114) is

$$u(x,t) = (A\cos\lambda x + B\sin\lambda x) \times [C\exp(-\lambda^2\sigma t)]. \qquad (117)$$

Conditions $u(0,t) = u(L,t) = 0$ lead to $A = 0$ and $\lambda_n = n\pi/L$ for $n = 1, 2, \ldots$, respectively. The final condition yields

$$u(x,0) = T_0\exp(-ax^2) = \sum_{n=1}^{\infty} B'_n \sin\left(\frac{n\pi x}{L}\right). \qquad (118)$$

Equation (118) is just a Fourier sine series, and the B'_n coefficients are given by

$$B'_n = \frac{2}{L}\int_0^L T_0\exp(-ax^2)\sin\left(\frac{n\pi x}{L}\right)dx = \frac{4T_0}{n\pi} \text{ for } n \text{ odd}, \qquad (119)$$
$$= 0 \text{ for } n \text{ even}.$$

The particular relation for the temperature distribution in the rod is therefore given by

$$u(x,t) = \frac{4T_0}{\pi}\sum_{\text{odd}}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi x}{L}\right) \times \exp\left(-\frac{n^2\pi^4\sigma t}{L}\right). \qquad (120)$$

3.5 The Two-Dimensional Laplace Equation

Laplace's equation is an example of an elliptic differential equation, and solutions of Laplace's equations are called harmonic functions. The electric potential $u(x,y)$ at points inside a rectangle (see Fig. 12) will now be determined from the solution of the two-dimensional Laplace equation with

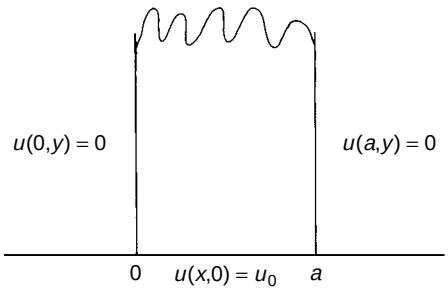


Fig. 12 Boundary configuration for Eq. (121)

the indicated boundary conditions:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0;$$
$$u(0,y) = u(a,y) = u(x,\infty) = 0 \text{ and } u(x,0) = u_0. \qquad (121)$$

Separation of variables with separation constant $-\lambda^2$ yields

$$\frac{d^2X}{dx^2} + \lambda^2 X = 0 \text{ with solution}$$
$$X(x) = A\cos\lambda x + B\sin\lambda x \qquad (122)$$

and

$$\frac{d^2Y}{dy^2} - \lambda^2 Y = 0 \text{ with solution}$$
$$Y(y) = C\exp(\lambda y) + D\exp(-\lambda y). \qquad (123)$$

The general solution of Eq. (121) is

$$u(x,y) = (A\cos\lambda x + B\sin\lambda x)[C\exp(\lambda y) + D\exp(-\lambda y)]. \qquad (124)$$

Condition $u(x,\infty) = 0$ requires that $C = 0$, condition $u(0,y) = 0$ leads to $A = 0$, and condition $u(a,y) = 0$ gives $\lambda_n = n\pi/a$ for $n = 1, 2, 3, \ldots$. The general solution now reduces to

$$u(x,y) = \sum_{n=1}^{\infty} B'_n \sin\left(\frac{n\pi x}{a}\right)\exp\left(-\frac{n\pi y}{a}\right),$$
$$B'_n = B_n D_n. \qquad (125)$$

The final condition is used to determine the values of B'_n as follows:

$$u(x, 0) = u_0 = \sum_{n=1}^{\infty} B'_n \sin\left(\frac{n\pi x}{a}\right). \tag{126}$$

Equation (126) is just a Fourier sine series, and the B'_n are given by

$$\begin{aligned} B'_n &= \frac{2}{a}\int_0^a u_0 \sin\left(\frac{n\pi x}{a}\right) dx \\ &= -\frac{4u_0}{n\pi} \text{ for } n \text{ odd,} \\ &= 0 \text{ for } n \text{ even.} \end{aligned} \tag{127}$$

The particular solution, expression for the potential at points within the rectangle in Fig. 12, is therefore

$$\begin{aligned} u(x, y) = -\frac{4u_0}{\pi} \sum_{\text{odd}}^{\infty} \frac{1}{n} \sin\left(\frac{n\pi x}{a}\right) \\ \times \exp\left(-\frac{n\pi y}{a}\right). \end{aligned} \tag{128}$$

The extension to more than two independent variables is straightforward. While the presentation has been restricted to Cartesian coordinates, inclusion of other coordinate systems (for example, cylindrical and spherical) may be carried out in the usual manner. In general, time-independent equations involving the Laplacian operator may be put in the form of Helmholtz's differential equation, $\nabla^2 u + k^2 u = 0$, when the appropriate k is used. Hence, solutions of Helmholtz's equation in various coordinate systems apply to all problems involving the Laplacian operator. In spherical coordinates (r, θ, ϕ), use of separation of variables, the power-series method, and the appropriate k for Helmholtz's equation lead to the following special functions: spherical harmonics, associated Legendre polynomials and Legendre polynomials, associated Laguerre polynomials and Laguerre polynomials, and spherical Bessel functions. Bessel functions result when cylindrical coordinates (ρ, ϕ, z) are used in Helmholtz's differential equation.

3.6 Fourier Transform Method

Methods of integral transforms are treated in detail elsewhere in the *Encyclopedia* (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). This section is devoted to the technique for solving differential equations (ordinary and partial) by use of the Fourier transform method. The one-dimensional Fourier transform pairs in symmetrical notation are given by

$$\begin{aligned} F(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{ikx}\, dx, \\ f(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(k) e^{-ikx}\, dk. \end{aligned} \tag{129}$$

In Eq. (129), $F(k)$ is referred to as the Fourier transform of $f(x)$, and $f(x)$ is the inverse transform of $F(k)$. The convention for quantum-mechanical problems involves a sign change in the exponents. Relations in Eq. (129) may be extended to multiple dimensions in a natural manner. The basic idea of the Fourier transform method for solving differential equations (ordinary or partial) is to transform the original equation (ordinary or partial) into a simpler equation (algebraic or ordinary differential) that can be easily solved. The required solution of the original differential equation is then obtained by finding the inverse transform of the solution of the simpler equation which is in transform space.

EXAMPLE 3.1: By use of the Fourier transform method, solve the ordinary differential equation, $\ddot{x} + 2\alpha\dot{x} + w_0^2 x = f(t)$

subject to conditions that $x(t)$ and $\dot{x}(t)$ go to zero as t goes to plus and minus infinity.

Solution: On taking the Fourier transform of each term in the original differential, the result is

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\ddot{x}(t)e^{i\omega t}dt + \frac{2\alpha}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\dot{x}(t)e^{i\omega t}dt + \frac{\omega_0^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x(t)e^{i\omega t}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}f(t)e^{i\omega t}\,dt. \quad (130)$$

By use of partial integration and the conditions that $x(t)$ and $\dot{x}(t)$ approach zero as t approaches plus and minus infinity, Eq. (130) reduces to the algebraic equation

$$-\omega^2 X(\omega) - 2\alpha i\omega X(\omega) + \omega_0^2 X(\omega) = F(\omega). \quad (131)$$

On solving the algebraic equation in Eq. (131) for $X(\omega)$ and inverting the transform, the solution $x(t)$ is obtained:

$$x(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{F(\omega)e^{-i\omega t}d\omega}{\omega_0^2 - \omega^2 - 2i\alpha\omega}. \quad (132)$$

The integral in Eq. (132) can be evaluated by use of the methods of calculus of residues when $f(t)$ is known.

EXAMPLE 3.2: By use of the Fourier transform method, solve the one-dimensional heat conduction equation for the temperature distribution $T(x, t)$ such that $T(x, t)$ and $T_x(x, t)$ approach zero as x approaches plus and minus infinity and $T(x, 0) = T_0 \exp(-ax^2)$ for constant a.

Solution: Here, one transforms out the space variable so that the resulting equation will be a first-order ordinary differential equation in t. On taking the Fourier transform of each term in the one-dimensional heat conduction equation, one obtains

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\partial^2 T}{\partial x^2}e^{ikx}dx = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial t}\int_{-\infty}^{\infty}T(x, t)e^{ikx}\,dx. \quad (133)$$

By use of partial integration and the conditions that $T(x, t)$ and $T_x(x, t)$ approach zero as x approaches plus and minus infinity, Eq. (133) reduces to

$$\frac{\partial T(k, t)}{\partial t} + \sigma k^2 T(k, t) = 0. \quad (134)$$

The solution of Eq. (134) is

$$T(k, t) = A(k)e^{-\sigma k^2 t} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}T(x, t)e^{ikx}\,dx. \quad (135)$$

Substituting the condition $T(x, 0) = T_0 \times \exp(-ax^2)$ into Eq. (135) yields

$$A(k) = \frac{T_0}{\sqrt{2a}}\exp\left(-\frac{k^2}{4a}\right). \quad (136)$$

The solution in transform space (k space) is therefore

$$T(k, t) = \frac{T_0}{\sqrt{2a}}\exp\left(-\frac{1 + 4\sigma at}{4a}\right)k^2. \quad (137)$$

The solution in x space is obtained when the k-space solution in Eq. (137) is inverted; the result is

$$T(x, t) = \frac{T_0}{\sqrt{1 + 4\sigma at}}\exp\left(\frac{-ax^2}{1 + 4\sigma at}\right). \quad (138)$$

The convolution theorem,

$$\int_{-\infty}^{\infty}f(x - \xi)g(\xi)d\xi = \int_{-\infty}^{\infty}F(k)G(k)\exp(-ikx)\,dk,$$

may be used to find the inverse transform when the solution in transform space is the product of two functions $F(k)G(k)$.

3.7 Green's Functions in Potential Theory

Here the three-dimensional Fourier transform method will be used to solve Poisson's equation for the electric potential $\phi(\mathbf{r})$ due to a volume charge density $\rho(\mathbf{r})$, and the three-dimensional Green's function will be defined. Poisson's equation is written as

$$\nabla^2\phi(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\varepsilon_0}. \tag{139}$$

The quantity ε_0 is the permittivity of free space. The Fourier transform of $\phi(\mathbf{r})$ has the form

$$\Phi(\mathbf{k}) = \frac{1}{(2\pi)^{2/3}}\int_{-\infty}^{\infty}\phi(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})d^3r. \tag{140}$$

A shorthand notation for triple integral, $d^3r = dx\ dy\ dz$, is used in Eq. (140). On taking the Fourier transform of both sides of Eq. (139), the solution in transform space [subject to the conditions that $\phi(\mathbf{r})$ and $\partial\phi/\partial r$ approach zero as r approaches plus and minus infinity] becomes

$$\Phi(\mathbf{k}) = \frac{\rho(\mathbf{k})}{k^2}\varepsilon_0. \tag{141}$$

The inverse transform of $\Phi(\mathbf{k})$ yields the solution $\phi(\mathbf{r})$, and the result is

$$\begin{aligned}\phi(\mathbf{r}) &= \frac{1}{(2\pi)^{3/2}}\int_{-\infty}^{\infty}\frac{\rho(\mathbf{k})}{k^2\varepsilon_0}\exp(-i\mathbf{k}\cdot\mathbf{r})d^3r \\ &= \frac{1}{(2\pi)^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\rho(\mathbf{r}')}{k^2\varepsilon_0} \\ &\quad\times\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')]\,d^3k\,d^3r' \\ &= \frac{1}{(2\pi)^2\varepsilon_0}\int_{-\infty}^{\infty}\rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r'.\end{aligned} \tag{142}$$

The function $G(\mathbf{r},\mathbf{r}')$, Green's function for the operator ∇^2, is given by

$$G(\mathbf{r},\mathbf{r}') = \int_{-\infty}^{\infty}\frac{\exp[-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')]d^3k}{k^2}.$$

When spherical polar coordinates are chosen where $d^3k = -k^2d(\cos\theta)d\phi dk$ and $\mathbf{r}-\mathbf{r}'$ is assumed to be along the polar axis, the expression for the Green's function reduces to $G(\mathbf{r},\mathbf{r}') = 2\pi^2/|\mathbf{r}-\mathbf{r}'|$.

Physically, the Green's function $G(\mathbf{r},\mathbf{r}')$ is the electric potential at point $\mathbf{r}$ due to a point charge located at $\mathbf{r}'$. For a volume charge density $\rho(\mathbf{r}')$, the potential at $\mathbf{r}$ is given by $\int\rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r'$. In differential equation form, this analysis may be written as $\nabla^2G(\mathbf{r},\mathbf{r}') = -4\pi\delta(\mathbf{r}-\mathbf{r}')$ subject to appropriate boundary conditions for $G(\mathbf{r},\mathbf{r}')$. The Dirac delta function $\delta(\mathbf{r}-\mathbf{r}')$ means

$$\delta(x-x')\delta(y-y')\delta(z-z')$$

with properties $\delta(\mathbf{r}-\mathbf{r}') = 0$ for $\mathbf{r}-\mathbf{r}' \neq 0$ and $\int_{-\infty}^{\infty}\delta(\mathbf{r}-\mathbf{r}')d^3r' = 1$. For Dirichlet boundary conditions, $G(\mathbf{r},\mathbf{r}') = 0$ on the boundary surface enclosing the charge distribution $\rho(\mathbf{r}')$. It can be shown that the Neumann problem requires appropriate nonzero values for the normal derivative of the Green's function on the boundary surface. Use of the Green's function method simplifies the problem of applying boundary conditions.

3.8 Numerical Methods for Partial Differential Equations

Numerical methods in partial differential equations form a vast subject and are treated in detail elsewhere in this book (see NUMERICAL METHODS). Here the focus is on essential concepts involved in converting a partial differential equation to its corresponding difference equation

by use of finite difference methods. One should consult the references for a detailed discussion of the various special techniques for finding numerical solutions, convergence of solutions, and stability of the various methods.

3.8.1 Fundamental Relations in Finite Differences

First differences $\Delta_x u$ and $\Delta_y u$ for positive h and k are defined by

$$\Delta_x u = \frac{u(x+h, y) - u(x, y)}{h},$$

$$\Delta_y u = \frac{u(x, y+k) - u(x, y)}{k}.$$

The corresponding second differences are defined by

$$\Delta_{xx} u = \frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2}$$

and

$$\Delta_{yy} u = \frac{u(x, y+k) - 2u(x, y) + u(x, y-k)}{k^2}.$$

For notational convenience, Δx and Δy are replaced with h and k, respectively, in the above finite difference equations, and k replaces Δt in Secs. 3.8.3 and 3.8.4.

3.8.2 Two-Dimensional Laplace Equation: Elliptic Equation

The two-dimensional Laplace equation in terms of finite differences reduces to

$$u(x, y) = \frac{1}{4}[u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h)].$$

The computational procedure involves replacing $u(x, y)$, for example, a potential, at a particular grid point (see Fig. 13) by the average value of its four closest neighbors. The function $u(x, y)$ or its derivative must

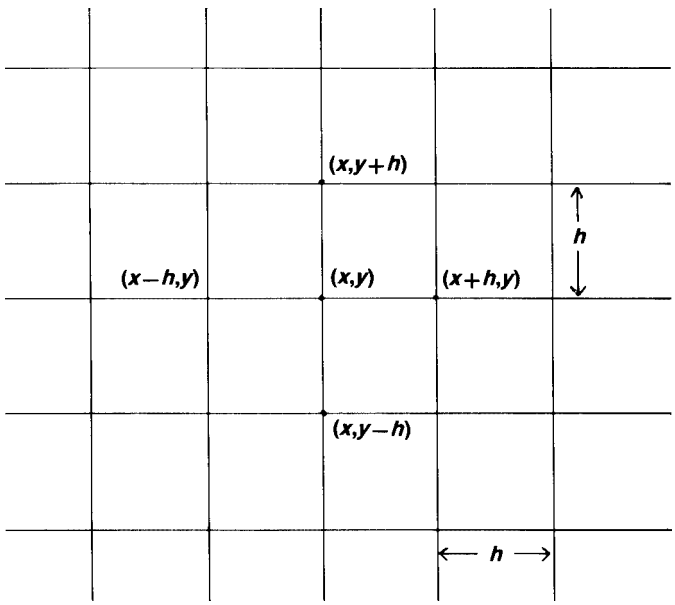


Fig. 13 Grid representation for Laplace's equation

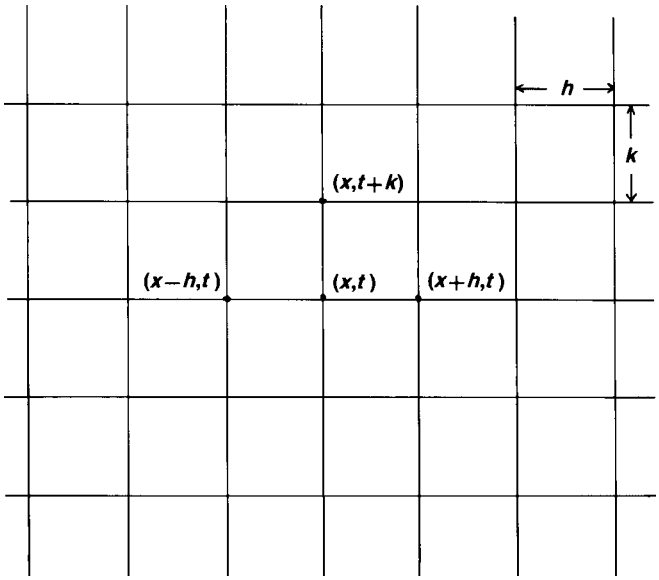


Fig. 14 Space-time grid for the heat equation

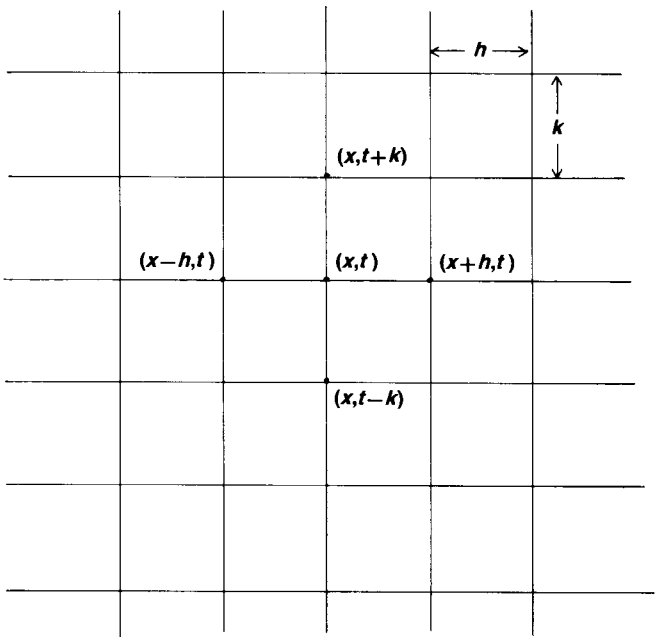


Fig. 15 Space-time grid for the wave equation

be specified at all points surrounding a given region.

3.8.3 One-Dimensional Heat Conduction Equation: Parabolic Equation

In terms of finite differences, the one-dimensional heat conduction equation reduces to

$$u(x, t+k) = \frac{\sigma k}{h^2}[u(x+h, t) - 2u(x, t) + u(x-h, t)] + u(x, t). \quad (143)$$

The numerical solution involves determining the initial values of $u(x, t)$ at various x locations (see Fig. 14 for the space-time grid) and applying Eq. (143) to obtain the $u(x, t)$ at other times.

3.8.4 One-Dimensional Wave Equation: Hyperbolic Equation

The finite-difference representation of the one-dimensional wave equation reduces to

$$u(x, t+k) = \frac{k^2 v^2}{h^2}[u(x+h, t) - 2u(x, t) + u(x-h, t)] + 2u(x, t) - u(x, t-k). \quad (144)$$

The starting value for $u(x, t+k)$ is determined from the initial conditions (see Fig. 15), and remaining values are determined by use of Eq. (144).

4 Integral Equations

4.1 Introduction

This section is devoted to a discussion of solutions and applications of one-dimensional linear integral equations of the first and second kinds. The formulations of many problems in physics lead to either differential or integral equations. Certain problems can only be represented by integral equations of the general form

$$u(x) = f(x) + \lambda \int_a^b k(x, s)u(s)\, ds. \quad (145)$$

Equation (145) is an integral equation since the unknown function $u(x)$ appears in the integrand. Functions $f(x)$ and $k(x, s)$ are to be given, and λ is a known parameter used here for convenience. The function $f(x)$ is called the free term, and $k(x, s)$ is referred to as the kernel (nucleus). Quantities $f(x)$, $k(x, s)$, and λ may be either real or complex but are considered real in this section. Equation (145) is a linear integral equation since u is linear. An integral equation is singular if either (or both) of the limits of integration is infinite and/or if the kernel becomes infinite in the range of integration. When $f(x)$ equals zero, Eq. (145) is classified as a homogeneous integral equation. If the kernel is continuous in the closed region $[a, b]$, then Eq. (145) is classified as a Fredholm-type integral equation of the second kind. The equation (where the upper limit is a variable)

$$u(x) = f(x) + \lambda \int_a^x k(x, s)u(s)ds \quad (146)$$

is known as a Volterra-type integral equation of the second kind. Fredholm integral equations of the first kind have the form

$$f(x) = \int_a^b k(x, s)u(s)\, ds. \quad (147)$$

Volterra-type integral equations of the first kind have the form

$$f(x) = \int_a^x k(x, s)u(s)\, ds. \quad (148)$$

In summary, classifications are Fredholm type if the limits of integration are fixed and Volterra type if one limit is variable, and first kind if the unknown function appears only in the integrand and second kind if the unknown function appears both in the integrand and outside the integrand.

Physical problems may be formulated as differential equations with appropriate boundary and/or initial conditions, integral equations, or either differential or integral equations. An essential difference in the formulation is that boundary conditions are imposed on general solutions of differential equations while boundary conditions are incorporated within the formulation of integral equations. While there exist procedures of converting differential equations to integral equations, use of integral equations seems more appropriate when formulations of problems lead directly to integral equations, or when solutions of the corresponding integral equations are easier to obtain than those for the corresponding differential equations. Laplace and Fourier transforms as well as dispersion relations are examples of singular integral equations of the first kind.

It is important to note that certain problems in classical mechanics, transport and diffusion phenomena, scattering theory, and other areas of physics can be formulated only by use of integral equations; the number of such problems is very small when compared to those leading to differential equations. In general, the theory of solution techniques needed in solving integral equations is not as familiar to physicists as techniques for solving differential equations. Integral equations are seldom treated in detail in introductory mathematical physics textbooks but are, however, discussed in advanced books in theoretical physics and mathematical physics. See Further Reading for some excellent books on integral equations. Many integral equations encountered in physics are normally solved by use of intuitive analytical methods, intuitive approximation methods and numerical techniques, or Laplace or Fourier transform methods.

Some systematic methods for solving nonsingular and linear integral equations are transform theory, Neumann series, separable kernel, Schmidt–Hilbert theory, Wiener–Hopf theory, and numerical. The Wiener–Hopf method is a different type of transform method which may be applied to certain integral equations with displacement kernels, $k(x, s) = k(x - s)$. Schmidt–Hilbert theory is an approach that applies to integral equations with Hermitian kernels, $k(x, s) = k^*(s, x)$. Fredholm theory involves representing the kernel as an infinite series of degenerate kernels (Sec. 4.2) and reducing the integral equation to a set of algebraic equations. Numerical solutions of Volterra equations involve reducing the original equations to linear algebraic equations, successive approximations and numerical evaluation of integrals. Numerical techniques for Fredholm equations involve solving a system of simultaneous equations.

4.2 Integral Equations with Degenerate Kernels

A subset of Fredholm equations of the first and second kinds with degenerate (separable) kernels can be solved by reducing them to a system of algebraic equations. In general, degenerate kernels may be written as

$$K(x, s) = \sum_{j=1}^{N} g_j(x)\phi_j(s). \qquad (149)$$

In Eq. (149), it is assumed that $g_j(x)$ and $\phi_j(s)$ are linearly independent quantities, respectively. Substituting Eq. (149) into Eq. (145) yields

$$u(x) = f(x) + \lambda \sum_{j=1}^{N} g_j(x) C_j. \tag{150}$$

The coefficients C_j are given by

$$C_j = \int_a^b \phi_j(s) u(s)\, ds. \tag{151}$$

The solution of Eq. (145) has now been reduced to finding the C_j from the indicated algebraic equations and substituting the C_j into Eq. (150).

EXAMPLE 4.1: By use of the degenerate kernel method, find the solution of $u(x) = x + \lambda \int_0^1 xsu(s)\, ds$. The integral equation becomes

$$u(x) = x + \lambda x \int_0^1 su(s) ds = x + \lambda x C. \tag{152}$$

The coefficient C reduces to

$$C = \int_0^1 su(s) ds = \int_0^1 s(s + \lambda s C) ds = \frac{1}{3-\lambda}. \tag{153}$$

The second step in Eq. (153) results when the second step of Eq. (152) is substituted into the first step of Eq. (153). From Eqs. (152) and (153), the solution of the original equation is $u(x) = 3x/(3-\lambda)$. It is seen that solutions exist for values of λ different from 3.

EXAMPLE 4.2: By use of the degenerate kernel method, find the solution of $u(x) = x + \frac{1}{2}\int_{-1}^{1}(s+x)\, ds$. The equation becomes

$$u(x) = x + \frac{C_1}{2} + \frac{xC_2}{2}. \tag{154}$$

The coefficients C_1 and C_2 are

$$C_1 = \int_{-1}^{1} su(s) ds = \frac{2 - C_2}{3} \text{ and}$$

$$C_2 = \int_{-1}^{1} u(s) ds = C_1 \quad \text{or} \quad C_1 = C_2 = 1. \tag{155}$$

On substituting the values for C_1 and C_2 into Eq. (154), the solution of the original equation becomes $u(x) = (3x+1)/2$.

4.3 Integral Equations with Displacement Kernels

If the kernel is of the form $k(x-s)$, it is referred to as a displacement kernel. Fredholm equations of the first and second kinds with displacement kernels and limits from minus infinity to plus infinity or from zero to plus infinity can normally be solved by use of Fourier and Laplace transform methods, respectively. Here the Fourier transform approach for solving integral equations with displacement kernels will be illustrated. Taking the Fourier transform of each term in Eq. (145) yields

$$\begin{aligned}\int_{-\infty}^{\infty} u(x) \exp(ikx) dx &= \int_{-\infty}^{\infty} f(x) \exp(ikx) dx \\ &\quad + \lambda \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(x-s) u(s)\, ds \right) \\ &\quad \times \exp(ikx)\, dx. \end{aligned} \tag{156}$$

In transform space, Eq. (156) is $u(k) = F(k) + \lambda K(k) u(k)$. The solution in x space is obtained when the inverse transform of $u(k)$ is taken, and the result becomes

$$u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{F(k) \exp(-ikx) dk}{1 - \lambda K(k)}. \tag{157}$$

4.4 The Neumann Series Method

Consider the set of Fredholm equations of the second kind such that

$$\int_a^b |f(x)|^2 dx \quad \text{and} \quad \int_a^b |K(x,s)|^2\, ds$$

are bounded. Assume the solution may be written as a power series, Neumann series (also known as Liouville–Neumann series), with form

$$u(x) = \sum_{n=0}^{\infty} \lambda^n u_n(x). \tag{158}$$

Terms in the successive approximation are obtained by substituting the Neumann series into the Fredholm equation, Eq. (145), and equating coefficients of like powers of λ; the results are

$$\begin{aligned} u_0 &= f(x); \\ u_1(x) &= \int_a^b K(x,s) u_0 ds; \ldots; \\ u_n(x) &= \int_a^b K(x,s) u_{n-1}(s)\, ds. \end{aligned} \tag{159}$$

It can be shown that the Neumann series converges for all values of λ for Volterra equations and converges for small values of λ for Fredholm equations; techniques exist for improving the convergence in the Fredholm case. Numerical techniques may be used to evaluate terms in the Neumann series.

4.5 The Abel Problem

The section on integral equations is concluded with the earliest application of integral equations to a physical problem, Abel's problem. The Abel problem is as follows: Consider a bead sliding on a smooth wire under the influence of gravity and find the curve for which the time of descent is a given function of the initial position.

Let the starting position of the bead be (x_0, y_0) and position of the bead at time t be (x, y) such that y equals zero at the end of the fall. The speed of the bead at (x, y) for ds an element of arc length along the path is determined from the conservation of energy principle and is given by

$$\frac{ds}{dt} = \sqrt{2g(y_0 - y)}.$$

If the shape of the curve is $u(y)$, then $ds = u(y)\, dy$ and the time of descent is given by

$$T = \int_0^{y_0} \frac{u(y) dy}{\sqrt{2g(y_0 - y)}}.$$

The Abel problem is to find the curve $u(y)$ for which the time T of descent is a given function $f(y_0)$ of the initial vertical position, and the result is obtained from the integral equation (Abel's equation):

$$f(y_0) = \int_0^{y_0} \frac{u(y) dy}{\sqrt{2g(y_0 - y)}}. \tag{160}$$

It can be shown that the curve in question is a portion of a cycloid.

5 Applied Functional Analysis

5.1 Introduction

Concepts of functions (of one variable) and operators were introduced into mathematics in connection with the development of calculus during the latter part of the seventeenth century. In general, an operator applied to a given function yields a

new function. The problem of finding an extremum (maximum or minimum) of a function is carried out in the usual manner by use of ordinary calculus, but the general problem of finding the stationary value (an extremum) of certain definite integrals that occur in mathematical physics is the subject matter of the branch of mathematics called the calculus of variations.

In relation to the calculus of variations, the process of connecting (mapping) each function $y(x)$ in $[a, b]$ with a number represented by the definite integral $\int_a^b F(y, y', x)\, dx$ (where $y' = dy/dx$) which depends on $y(x)$ was given the name functional during the end of the nineteenth century. The basic idea of functional analysis is that problems are often easier to solve if a function is considered to be a member of a whole space of functions, X. The space X is assumed to carry a metric, have a linear space structure, and be infinite dimensional. The concept of a metric involves topological and geometrical language while linear operators on X involve concepts of linear algebra, and relations among these concepts constitute linear functional analysis.

A function which depends on one or more functions rather than on discrete variables is referred to as a functional. The domain of a functional is a space of admissible functions. More precisely, functionals are continuous linear maps, from a normed space into itself or into some other normed space. The basic ingredient of the various definitions of a functional and of functional analysis is the existence of a linear space with a topology.

The main topics in Secs. 1–4 (functions of a complex variable and analytic functions, ordinary and partial differential equations, Fourier series and Fourier transform theory, and integral equations) are technically topics in functional analysis even though the topology and geometry of the linear spaces involved were not stressed. Mathematically, a valid argument can be made that concluding this article with a discussion of functional analysis is analogous to putting the cart before the horse. This argument, however, neglects the applications-of-techniques approach emphasized throughout the article.

In mathematical physics, functional analysis often involves discussions connected with the calculus of variations; theory of ordinary and partial differential equations; integral equations and transform theory; spectral theory involving eigenvalues, eigenfunctions, and Fourier series expansion theory involving orthogonal functions; functional calculus used in the path integral formulation of quantum mechanics, quantum field theory, and statistical mechanics; C^* algebra; and the theory of distributions. In mathematics, functional analysis often involves the general theory of linear normed spaces, the topological structure of linear spaces and continuous transformations, measure spaces and general theories of integration, spectral theories, C^* algebra, distribution theory, and number theory.

In this section, the original problem of functional analysis (the calculus of variations) and applications of functional integration to quantum mechanics, quantum field theory, and statistical mechanics will be discussed.

5.2 Stationary Values of Certain Definite Integrals

Consider the following definite integral of the functional $F(y, y', x)$ where F is a known function of y, y' (where $y' = dy/dx$),

and x, but $y(x)$ is unknown:

$$J = \int_{x_1}^{x_2} F(y, y', x)\, dx. \tag{161}$$

A fundamental problem in the calculus of variations (a problem which occurs frequently in mathematical physics) is that of finding a function $y(x)$ such that the functional J is stationary (an extremum; a minimum in most cases of physical interest). The basic procedure here is to evaluate the integral for a slightly modified path $y(x, a) = y(x, 0) + \alpha\eta(x)$ where $\eta(x_1) = \eta(x_2) = 0$ (all paths pass through the end points) and show that the change in the value of the integral due to the change in the path becomes zero. The function $\eta(x)$ is an arbitrary differentiable function, and α is a small scale factor (see Fig. 16). The function $y(x, \alpha)$ describes neighboring paths where $\delta y = y(x, a) - y(x, 0) = \alpha\eta(x)$ is the variation (hence the name calculus of variations) of $y(x,0)$ at some x. The delta symbol, δ, was introduced by Lagrange to denote a variation (a virtual change) and means a change made in an arbitrary manner. Both dy and δy denote infinitesimal changes in y, but dy means an infinitesimal change in $y(x)$ produced by dx while δy is an infinitesimal change which produces $y + \delta y$. It is straightforward to show that $d\delta y/dx = \delta dy/dx$ and $\delta \int_{x_1}^{x_2} F(y, y', x)dx = \int_{x_1}^{x_2} \delta F(y, y', x)\, dx$. On substituting $y(x, \alpha) = y(x, 0) + \alpha\eta(x)$ into Eq. (161) and differentiating both sides of the resulting equation with respect to α, one obtains

$$\frac{dJ(\alpha)}{d\alpha} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y}\eta(x) + \frac{\partial F}{\partial y'}\eta'(x) \right) dx. \tag{162}$$

Integrating the second term in Eq. (162) by parts and using the fact that $\eta(x)$ is arbitrary yield

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0. \tag{163}$$

Equation (163) is known as Euler's equation and its solution yields the $y(x)$ which makes J an extremum (minimum). An alternative and often used approach for obtaining Euler's equation makes use of expanding the right-hand side of δF in a Taylor's series with two variables; the result becomes

$$\begin{aligned} \delta F &= F(y + \alpha\eta, y' + \alpha\eta', x) - F(y, y', x) \\ &= \alpha\left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta'\right). \end{aligned} \tag{164}$$

Higher-order terms in the Taylor expansion may be used to determine the nature of the extremum (maximum or minimum), and neglected here since α is a small parameter. As a result of substituting Eq. (164) into the integrand for δJ, integrating the second term by parts as before, and setting $\delta J/\alpha = 0$, one obtains the Euler equation in Eq. (163).

The above processes of obtaining Euler's equation, Eq. (163), may be extended to functionals involving several dependent and/or independent variables; for example, the variational process applied to

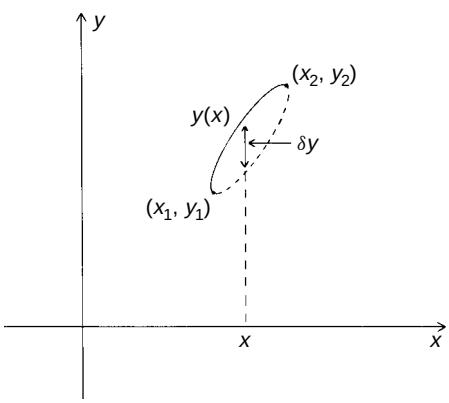


Fig. 16 A varied path between (x_1, y_1) and (x_2, y_2)

$F(y_1, \ldots, y_n, y_1', \ldots, y_n', x)$ yields the following set of Euler's equations:

$$\frac{\partial F}{\partial y_k} - \frac{d}{dx}\left(\frac{\partial F}{\partial y_k'}\right) = 0 \quad (k = 1, 2, \ldots, n). \tag{165}$$

EXAMPLE 4.3: By use of the variational calculus method (Euler's equation), determine the equation of the shortest path between two points (x_1, y_1) and (x_2, y_2) in Cartesian coordinates.

Solution: The element of distance along the path between the two points is given by

1.
$$ds = \sqrt{dx^2 + dy^2}.$$

The expression for the distance between the two points is therefore

2.
$$s = \int_{x_1}^{x_2} \sqrt{1 + (y')^2}\, dx,$$

where
$$y' = \frac{dy}{dx}.$$

For $F(y, y', x) = \sqrt{1 + (y')^2}$, the differential equation for the equation of the shortest path between the two points, Euler's equation, reduces to

3.
$$\frac{dy}{dx} = A \quad \text{since} \quad \frac{\partial F}{\partial y} = 0 \quad \text{and}$$
$$\frac{\partial F}{\partial y'} = \frac{y'}{\sqrt{1 + (y')^2}}.$$

The equation of the shortest path between the two points is therefore that of a straight line, $y(x) = Ax + B$.

EXAMPLE 4.4 (the *Brachistochrone* Problem): The brachistochrone (shortest time) problem, first formulated and solved by Johann Bernoulli in 1696, is one of the first variational problems. The problem is as follows: Consider a bead of mass m which slides, under the influence of gravity, down a frictionless wire bent into the appropriate shape. The goal is to find the equation (shape of the wire) of the path along which the bead travels so that the time is a minimum.

Solution: For convenience, it is assumed that the bead starts from rest at the origin of a coordinate system (see Fig. 17). Since this is a conservative system, the following relations are valid: $T_1 + V_1 = T_2 + V_2$, $V_2 = -mgy$, $T_1 = V_1 = 0$, $T_2 = \frac{1}{2}mv^2$, and $v = \sqrt{2gy}$. The expression for the time required for the bead to travel from the origin to point (x, y) is therefore given by

1.
$$t = \int \frac{\sqrt{dx^2 + dy^2}}{\sqrt{2gy}}$$
$$= \int_0^{y_2} \frac{\sqrt{1 + (x')^2}\, dy}{\sqrt{2gy}}, \quad x' = \frac{dx}{dy}.$$

The unknown function $y(x)$ must be determined such that the time is a minimum. On applying Euler's equation

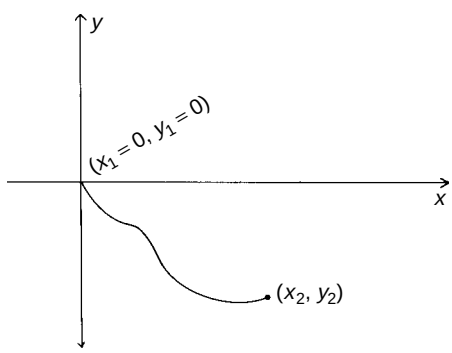


Fig. 17 Diagram for the brachistochrone example

with

$$F = \left(\frac{1 + (x')^2}{y}\right)^{1/2}$$

and independent variable y, one obtains

2.

$$x = \int \frac{Ay\,dy}{\sqrt{y - A^2y^2}} \text{ since}$$

$$\frac{\partial F}{\partial x} = 0 \text{ and}$$

$$\frac{\partial F}{\partial x'} = \frac{x'}{\sqrt{y[1 + (x')^2]}} = A.$$

On letting $A = 1/\sqrt{2a}$ and making the change of variable $y = a(1 - \cos\theta)$, the above integral reduces to $x = a(\theta - \sin\theta) + \text{const}$. The path that yields a minimum time of travel is in the form of parametric equations $x = a(\theta - \sin\theta)$ and $y = a(1 - \cos\theta)$, equations for a cycloid that passes through the origin.

5.3 Hamilton's Variational Principle in Mechanics

5.3.1 Introduction

Mechanics is the study of the motions (including rest) of physical objects. The laws of classical mechanics are valid for macroscopic objects (size larger than 10^{-10} m), and the laws of quantum mechanics are valid in the microworld (object size smaller than 10^{-10} m). In this section, the focus is on the study of classical mechanics. Widely used equivalent formulations of classical mechanics are Newtonian mechanics (1687), Lagrangian mechanics (1788), Hamiltonian mechanics (1834), and Hamilton-Jacobi theory (1837).

Formulations of classical mechanics developed since Newtonian mechanics are generalizations and equivalent representations of Newtonian mechanics. These generalizations do not lead to new information but offer different ways of approaching problems. Certain problems can be solved by use of all four approaches with equal amounts of ease (or difficulty). Other problems are more amenable to solution by use of one approach than by use of the others. The specific nature of the problem to be solved usually dictates the approach that should be used.

Newton's second law is the basic equation of motion in the Newtonian picture of mechanics. In Lagrangian mechanics, Lagrange's equations are the required set of equations of motion for the system (particle or group of particles) under investigation. Hamilton's canonical equations are basic to Hamiltonian mechanics, and the Hamilton–Jacobi equation is the foundation of the Hamilton–Jacobi theory.

The approach in this section begins with Hamilton's variational principle for conservative systems (where the forces acting on the system may be derived from a potential function) from which Lagrange's equations will be developed by use of the variational calculus method. By use of a Legendre transformation, the Hamiltonian and subsequently Hamilton's canonical equations are obtained.

The variational technique used in mechanics was developed mainly by Euler and Lagrange and is a mathematical formulation of mechanics in which kinetic energy and potential energy play an essential role. In Newtonian mechanics, forces play the central role.

5.3.2 Generalized Coordinates

Linearly independent quantities $\{q_k\} = q_1, \ldots, q_k$ that completely define the position (configuration) of a system as a

function of time are called generalized coordinates. Quantities $\{q_k\}$ are said to be linearly independent if $\sum_k \alpha_k q_k = 0$ implies that $\alpha_k = 0$ for all k. Generalized coordinates may be selected to match the conditions of the problem to be solved. The number of generalized coordinates that must be used to define uniquely the position of a system represents the number of degrees of freedom for the system. The corresponding quantities $\{\dot{q}_k\}$ are called generalized velocities.

The simultaneous specification of $\{q_k\}$ and $\{\dot{q}_k\}$ for a system determines the mechanical state of the system at that time, and subsequent motion is obtained from the solutions $q_k(t)$ of the appropriate equations of motion. The appropriate second-order differential equations expressing the relations among generalized coordinates q_k, generalized velocities $\{\dot{q}_k\}$, and generalized accelerations $\{\ddot{q}_k\}$ are called equations of motion for the system under investigation.

Although the set of generalized coordinates used to solve a problem is not unique, a proper set of generalized coordinates is that set which leads to an equation of motion whose solution has a straightforward physical interpretation. No general rule exists for obtaining a proper set of generalized coordinates.

5.3.3 Lagrange's Equations

Hamilton's variational principle asserts that the actual motion of a particle or system of particles (conservative system) from its initial configuration at time t_1 to its configuration at time t_2 is such that

$$\delta S = \delta \int_{t_1}^{t_2} L(q_k, \dot{q}_k) dt = 0. \qquad (166)$$

In Eq. (166), $q_k = q_k(t)$, $L = T - V$ is defined as the Lagrangian for the system under investigation, $L\,dt$ is called the action, and

$$S = \int_{t_1}^{t_2} L\,dt$$

denotes the action integral. The quantities T and V are kinetic and potential energy, respectively.

Among the infinite number of trajectories $q(t)$ that connect the end points $q(t_1)$ and $q(t_2)$, the physical (actual) path yields a stationary value for the action integral. The action is therefore a functional of the functions $q_k(t)$ satisfying the boundary conditions that all trajectories pass through the end points. By use of the variational technique leading to Eq. (165), one finds that $q(t)$ is obtained from the following set of differential equations:

$$\frac{\partial L}{\partial q_k} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) = 0, \quad k = 1, 2, \ldots, n. \qquad (167)$$

The equations in Eq. (167) are called Lagrange's (or Euler–Lagrange) equations. Lagrange's equations, the equations of motion for the system under investigation, are a set of n second-order differential equations. The general solutions of these equations contain $2n$ arbitrary constants of integration. The values of these $2n$ arbitrary constants are determined when the initial state (initial values for the q_k and $\{\dot{q}_k\}$ at $t = 0$) of the system is specified.

Quantities $\partial L/\partial \dot{q}_k$ and $\partial L/\partial q_k$ are defined to be canonical momenta (also called conjugate or generalized momenta) and generalized forces, respectively,

$$p_k = \frac{\partial L}{\partial \dot{q}_k} \quad \text{and} \quad F_k = \frac{\partial L}{\partial q_k}. \qquad (168)$$

By use of the definitions in Eq. (168), it is observed that Lagrange's equations may be considered a generalized version of Newton's second law where generalized

force equals the rate of change of canonical momentum.

Basic conservation laws of mechanics result from invariance of the Lagrangian under time translation–conservation of energy, coordinate translation–conservation of canonical momentum, translation in space–conservation of total linear momentum, and rotation in space–conservation of angular momentum. In spite of the important role of the Lagrangian, it is not a unique function for a system since the equations of motion for the system, Lagrange's equations, are unchanged if $df(q_k, t)/dt$ is added to the Lagrangian.

5.3.4 Format for Solving Problems by Use of Lagrange's Equations

The following steps should be used when applying Lagrange's equations.

1. Draw a detailed diagram. Specify the degrees of freedom and the level where potential energy V equals zero.
2. Write down the appropriate expressions for T, V, and L.
3. Write down the specific set of Lagrange's equation(s).
4. Work out the terms in the set of equations in Step 3.
5. Solve the resulting equation(s) of motion subject to the given initial conditions.

5.4 Formulation of Hamiltonian Mechanics

It has been shown that Hamilton's variational principle combined with techniques of the calculus of variations transforms the process of finding the solution of a mechanical problem to that of obtaining solutions for Lagrange's equations. Hamilton developed a procedure for transforming Lagrange equations to a simpler (canonical) form by replacing them (a set of n second-order differential equations) with a set of $2n$ first-order differential equations now called Hamilton's canonical equations of motion.

5.4.1 Derivation of Hamilton's Canonical Equations

The Lagrangian is a function of q_k and $\dot{q}_k$; now the change of variable $\dot{q}_k \to p_k$ where $p_k = \partial L/\partial \dot{q}_k$ will be made. By use of a Legendre transformation [new function equals the old function minus (the derivative of the old function with respect to the old variable) times the old variable; the physical and geometrical content of the new and old functions is the same], one obtains

$$-H \equiv L - \sum_{k=1}^{n} p_k \dot{q}_k. \tag{169}$$

The negative sign in Eq. (169) is by convention. The new function $H(q_k, p_k)$ contains the same geometrical and physical content as $L(q_k, \dot{q}_k)$ and is called the Hamiltonian of the system. Note that the action integral may now be written as

$$S = \int_{t_1}^{t_2} \left(\sum_{k=1}^{n} p_k \dot{q}_k - H \right) dt. \tag{170}$$

Applying the variational techniques of Sec. 5.2 to Eq. (170) yields

$$\dot{q}_k = \frac{\partial H}{\partial p_k} \quad \text{and} \quad \dot{p}_k = -\frac{\partial H}{\partial q_k}. \tag{171}$$

The equations in Eq. (171) are referred to as Hamilton's canonical equations of motion (or simply Hamilton's equations). Hamilton's equations can be used to develop the specific set of equations of motion for the system under investigation in terms of the phase space variables q_k and p_k. Note that Lagrange's equations consist

of n second-order differential equations whereas Hamilton's equations form a set of $2n$ first-order differential equations. For a conservative system, it can be shown that the Hamiltonian equals the total energy of the system ($H = T + V$).

5.4.2 Format for Solving Problems by Use of Hamilton's Equations

In solving problems by use of Hamiltonian mechanics, the following five-step procedure is highly recommended.

1. Write out the Lagrangian as in Lagrangian mechanics, $L = T - V$.
2. Solve the equation $p_k = \partial L/\partial \dot{q}_k$ for $\dot{q}_k$ and eliminate $\dot{q}_k$ from the Lagrangian.
3. Construct the Hamiltonian for the system, $H = \sum_{k=1}^{n} \dot{q}_k p_k - L$.
4. Obtain Hamilton's equations, $\dot{q}_k = -\partial H/\partial p_k$ and $\dot{p}_k = -\partial H/\partial q_k$.
5. Solve the $2n$ first-order differential equations (equations of motion) developed in step 4.

5.4.3 Poisson's Brackets

The total time derivative of a function $f(q_k, p_k)$ is

$$\frac{df}{dt} = \sum_{k=1}^{n} \left(\frac{\partial f}{\partial q_k} \dot{q}_k + \frac{\partial f}{\partial p_k} \dot{p}_k \right) = \sum_{k=1}^{n} \left(\frac{\partial f}{\partial q_k} \frac{\partial H}{\partial p_k} - \frac{\partial f}{\partial p_k} \frac{\partial H}{\partial q_k} \right) = \{f, H\}. \tag{172}$$

Hamilton's equations were used in obtaining Eq. (172). The last quantity in Eq. (172) is called a Poisson bracket. A Poisson bracket is defined by

$$\{f, g\} = \sum_{k=1}^{n} \left(\frac{\partial f}{\partial q_k} \frac{\partial g}{\partial p_k} - \frac{\partial f}{\partial p_k} \frac{\partial g}{\partial q_k} \right). \tag{173}$$

Hamilton's canonical equations in terms of Poisson brackets are given by

$$\dot{q}_k = \frac{\partial H}{\partial p_k} = \{q_k, H\}, \qquad \dot{p}_k = -\frac{\partial H}{\partial q_k} = \{p_k, H\}. \tag{174}$$

Two variables ξ_i and ϕ_i are said to be canonically conjugate if

$$\{\xi_i, \xi_k\} = \{\phi_i, \phi_k\} = 0 \quad \text{and} \quad \{\xi_i, \phi_k\} = \delta_{ik}. \tag{175}$$

The Kronecker delta function is defined by

$$\delta_{ik} = \begin{cases} 1; & i = k, \\ 0; & i \neq k. \end{cases} \tag{176}$$

The quantities q_j and p_j are canonically conjugate variables since $\{q_j, p_k\} = \delta_{jk}$ and $\{q_j, q_k\} = \{p_j, p_k\} = 0$; these three Poisson brackets are referred to as fundamental Poisson brackets.

5.5 Continuous Media and Fields

Thus far, only conservative systems composed of discrete particles have been considered. The Lagrangian of a system composed of N free particles may be written as

$$L = \sum_{i=1}^{N} L_i. \tag{177}$$

The extension of the above analysis to a system with an infinite number of degrees of freedom (a continuous medium) is achieved by replacing the subscript k with a continuous variable (say $\mathbf{x}$), q_k with a new function $q_k \to Q(\mathbf{x}, t)$, the sum with an integral $\sum_i \to \int d^3x$, and canonical momenta with canonical momentum

density given by $\pi(\mathbf{x}) = \partial\mathscr{L}/\partial\dot{Q}$ where $\mathscr{L}$ is the Lagrangian density. The quantity $Q(\mathbf{x}, t)$ is called a field. To denote several fields, the notation $Q_\alpha(\mathbf{x}, t)$ may be used. The parameter α distinguishes among the different fields. From a mathematical point of view, a field is a set of functions of space-time, and these functions satisfy a set of partial differential equations. The corresponding Hamilton's variational principle is

$$0 = \delta \int_{t_1}^{t_2} \sum_{i=1}^{N} L_i(q_k, \dot{q}_k)dt$$
$$= \delta \int_{t_1}^{t_2} \int_{\substack{\text{physical}\\ \text{space}}} \mathscr{L}\{Q_\alpha(\mathbf{x}, t), \dot{Q}_\alpha(\mathbf{x}, t)\}d^4x;$$
$$d^4x = dx\, dy\, dz\, dt. \tag{178}$$

Assuming that fields interact only with infinitesimal neighbors, the Lagrangian density should be a function of $Q_\alpha(\mathbf{x}, t)$, $\dot{Q}_\alpha(\mathbf{x}, t)$, and $\partial Q_\alpha(\mathbf{x}, t)/\partial x_k$ or $Q_\alpha(x^\mu)$ and $\partial_\mu Q_\alpha$ in four-vector notation. By use of appropriate boundary conditions, the variation in Eq. (178) leads to the following set of equations of motion:

$$\frac{\partial\mathscr{L}}{\partial Q_\alpha} - \partial_\mu\left(\frac{\partial\mathscr{L}}{\partial(\partial_\mu Q_\alpha)}\right) = 0,$$
$$\mu = 0, 1, 2, 3. \tag{179}$$

The equations in Eq. (179) are the Lagrange's equations for classical fields.

5.6 Transitions to Quantum Mechanics

The laws of classical mechanics are not in general valid for the microworld, and new laws (quantum theory) that are appropriate for the microworld were developed during the period 1900–1927. In this section, the transition from classical mechanics to quantum mechanics in the Heisenberg picture, in the Schrödinger picture, and by use of the action functional (path integral) approach due to Dirac and Feynman will be made. For notational convenience, the discussion is restricted to the case of one nonrelativistic particle. The starting point in both the Heisenberg and Schrödinger pictures is Hamiltonian mechanics while the Feynman (Dirac–Feynman) approach begins with Lagrangian mechanics.

The postulates of quantum mechanics may be stated as follows.

1. Each state of a physical system corresponds to a normalized vector in Hilbert space called the state vector, Ψ or $|\Psi\rangle$.
2. Physical quantities are represented by linear Hermitian operators in Hilbert space.
3. If a system is in a state $|\Psi\rangle$, then the probability that a measurement (consistent with quantum theory) of the quantity corresponding to $\widehat{A}$ will yield one of the eigenvalues a_k (where $\widehat{A}|\Psi\rangle = a_k|\Psi\rangle$) is given by $|\langle a_k|\Psi\rangle|^2$. The system will change from state $|\Psi\rangle$ to $|a_k\rangle$ as a result of the measurement. The quantity $\langle a_k|\Psi\rangle$ is the amplitude.

5.6.1 The Heisenberg Picture

In the Heisenberg approach, a system is quantized by letting q_k and p_k be Hermitian operators in a Hilbert space such that $q_k \to \widehat{q}_k$ and $p_k \to -i\hbar\partial/\partial q_k$, and replacing Poisson brackets with commutators, $\{A, B\} \to [\widehat{A}, \widehat{B}]/i\hbar$ where $[\widehat{A}, \widehat{B}] = \widehat{A}\widehat{B} - \widehat{B}\widehat{A}$. If $[\widehat{f}, \widehat{g}] = i\hbar$, the operators $\widehat{f}$ and $\widehat{g}$ are said to be canonically conjugate. The resulting Heisenberg equations of motion for a quantum and mechanical system are

$$i\hbar\dot{p}_k = [p_k, H] \quad \text{and} \quad i\hbar\dot{q}_k = [q_k, H]. \tag{180}$$

The equations in Eq. (180) are basic for Heisenberg (matrix) mechanics.

5.6.2 The Schrödinger Picture

From a classical mechanical point of view, the Hamiltonian of a particle subject to conservative forces equals the total energy of the particle, and one may write

$$H = E = \frac{\mathbf{p}^2}{2m} + V(x, y, z). \qquad (181)$$

The transition to quantum mechanics in the Schrödinger picture is achieved by use of the replacements $E \to i\hbar\partial/\partial t$ and $\mathbf{p} \to -i\hbar\nabla$; by use of these replacements, Eq. (182) is transformed into an operator equation. Operating on some function $\Psi(x, y, z, t)$ or $|\Psi\rangle$ in Hilbert space yields

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi \quad \text{or}$$
$$i\hbar\frac{\partial|\Psi\rangle}{\partial t} = \widehat{H}|\Psi\rangle. \qquad (182)$$

Schrödinger's equation, Eq. (182), is the basic equation of motion of a particle in quantum mechanics in the Schrödinger picture.

5.6.3 The Feynman Path-Integral Approach to Quantum Mechanics

The special case of one particle with one degree of freedom is considered here to simplify the notation and make the explanations clear. Feynman's formulation of quantum mechanics was stimulated by some work of Dirac (1933) and is based on the following two postulates:

1. The amplitude $\langle q(t'')|q(t')\rangle$ for a particle to be found at $q(t'')$ at time t'' if its initial position is $q(t')$ at time t' equals a sum of complex contributions (amplitudes) for each space-time path starting at $q(t')$ and ending at $q(t'')$.
2. All paths connecting $q(t')$ and $q(t'')$ contribute equally in magnitude, but the phase (weight) of their contribution is $\exp(iS/\hbar)$ where S is the classical action integral for the corresponding paths.

The measure on the functional space of paths $q(t)$ is denoted by $\mathscr{D}[q(t)]$, and appropriate normalization factors for the amplitude are contained in $\mathscr{D}[q(t)]$. Feynman's interpretation of the indicated functional integration is as follows: Divide the time interval $t'' - t'$ into N equal parts, each with duration $\varepsilon = t_{k+1} - t_k$; and in the limit $N \to \infty(\varepsilon \to 0)$, it is assumed that the sequence of points $q(t_0), \ldots, q(t_n)$ approximates the path $q(t)$. The action functional associated with the classical path joining $q(t_k) = q_k$ and $q(t_{k+1}) = q_{k+1}$ is

$$S[q_{k+1}, q_k] = \int_{t_k}^{t_{k+1}} L(q, \dot{q})\, dt.$$

Feynman's postulates thus assert that the amplitude $\langle q(t'')|q(t')\rangle$ is a sum of all amplitudes for all paths connecting $q(t'')$

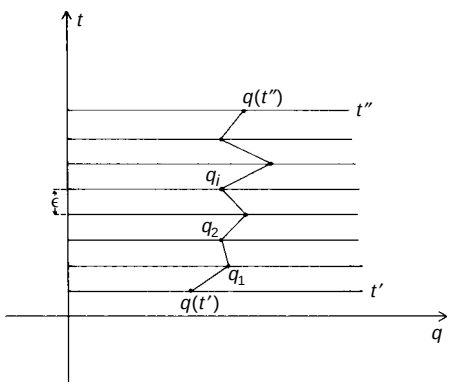


Fig. 18 A representative sequence of paths between $q(t')$ and $q(t'')$

and $q(t')$, and it may be written as (see Fig. 18)

$$\langle q(t'')|q(t')\rangle = \lim_{\substack{\varepsilon\to 0\\ N\to\infty}} \int\cdots\int \prod_{k=0}^{N} \exp\left(\frac{i}{\hbar} S(q_{k+1}, q_k)\right) \frac{dq_k}{A_k} = \int\cdots\int \exp\left(\frac{i}{\hbar}\int_{t'}^{t''} L(q, \dot{q})\, dt\right) \mathscr{D}[q(t)]. \tag{183}$$

The normalization factors A_k in Eq. (183) are independent of the path from q_k to q_{k+1} but depend on the mass of the particle and on the time interval ε. Equation (183) is a mathematical statement that the amplitude for a particle at $q(t')$ at time t' to move to $q(t'')$ at time t'' equals the sum of all possible paths between the two points times exp $(iS/\hbar)$; the probability is the absolute square of the amplitude.

The path integral approach to quantum mechanics can be extended to include formulations of quantum field theory (a combination of quantum mechanics and special relativity), the partition function in statistical mechanics, and systems obeying Bose–Einstein and Fermi–Dirac statistics. The path integral method is the foundation for Feynman diagrams.

Glossary

Complex Variable: An ordered pair of real variables ($z = x + iy$) with a real and an imaginary part.

Ordinary Differential Equation: An equation containing derivative(s) with respect to one independent variable.

Partial Differential Equation: An equation containing partial derivatives with respect to two or more independent variables.

Integral Equation: An equation where the unknown function appears in an integrand.

Functional: A function which depends on one or more functions.

Further Reading

Arfken, G. (1985), *Mathematical Methods for Physicists*, New York: Academic.

Bradbury, T. C. (1984), *Mathematical Methods with Applications to Problems in the Physical Sciences*, New York: Wiley.

Churchill, R. V., et al. (1974), *Complex Variables and Applications*, New York: McGraw-Hill.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill.

Glimm, J., Jaffe, A. (1987), *Quantum Physics: A Functional Integral Point of View*, New York: Springer.

Goldstein, H. (1950), *Classical Mechanics*, Reading, MA: Addison-Wesley.

Harper, C. (1976), *Introduction to Mathematical Physics*, Englewood Cliffs, NJ: Prentice-Hall.

Itzykson, C., Zuber, J. B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Jones, L. M. (1979), *An Introduction to Mathematical Methods of Physics*, Reading, MA: Benjamin/Cummings.

Kreyszig, E. (1962), *Advanced Engineering Mathematics*, New York: Wiley.

Kreyszig, E. (1978), *Introductory Functional Analysis With Applications*, New York: Wiley.

Lanczos, C. (1970), *The Variational Principles of Mechanics*, New York: Dover.

Lovitt, W. V. (1950), *Linear Integral Equations*, New York: Dover.

Mikhlin, S. G. (1964), *Integral Equations*, New York: MacMillan.

Moiseiwitsch, B. L. (1977), *Integral Equations*, New York: Longman.

Morse, P. M., Feshbach, H. (1953), Methods of Theoretical Physics, Vols. 1 and 2. New York: McGraw-Hill.

Pipes, L. A., Harvill, L. R. (1970), *Applied Mathematics for Engineers and Physicists*, New York: McGraw-Hill.

Potter, M. C., Goldberg, J. (1978), *Mathematical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Press, W. H., et al. (1987), *Numerical Recipes*, New York: Cambridge University.
Pryce, J. D. (1973), *Basic Methods of Linear Functional Analysis*, London: Hutchinson.
Schulman, L. S. (1981), *Techniques and Applications of Path Integration*, New York: Wiley.
Schutz, B. F. (1980), *Geometrical Methods of Mathematical Physics*, New York: Cambridge University.
Simmons, G. F. (1972), *Differential Equations*, New York: McGraw-Hill.
Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.

Fourier and Other Mathematical Transforms

Ronald N. Bracewell
Electrical Engineering Department, Stanford University, Stanford, California, USA

Introduction *84*
1 The Fourier Transform *84*
2 Continuous Versus Discrete Transforms *85*
3 Some Common Transforms *87*
4 The Laplace Transform *91*
5 Convergence Conditions *92*
6 Why Transforms Are Useful *93*
7 Fields of Application *94*
8 The Hartley Transform *95*
9 The Fast Fourier Transform *96*
10 The Fast Hartley Algorithm *97*
11 The Mellin Transform *98*
12 The Hilbert Transform *99*
13 Multidimensional Transforms *99*
14 The Hankel Transform *100*
15 The Abel Transform *100*
16 Tomography and the Radon Transform *102*
17 The Walsh Transform *103*
18 The z Transform *104*
19 Convolution *104*
20 Summary *106*
Glossary *106*
List of Works Cited *106*
Further Reading *107*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Introduction

Fourier analysis, which gained prominence from the work of J. B. J. Fourier in the early 1800s, led immediately to applications in mechanics and heat conduction but also contributed to the advance of pure mathematics as regards the basic notions of limit, convergence, and integrability; the impact on mathematics and applied physics has continued to this day. Applications of transform methods were developed in connection with differential and integral equations and became very powerful; more recently, numerical analysis, aided by electronic computing, has added an extra dimension to the applied relevance of mathematical transforms and especially of the Fourier transform. The analytic and computational aspects will be dealt with first; among applied examples, heat conduction, Fourier-transform spectroscopy, diffraction, sampled data, and tomography will be mentioned.

When one looks for antecedents from which Fourier analysis might have evolved they are not hard to find. Euler had published trigonometric series, and the sum to infinity, in such statements as

$$\sin x - \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x + \cdots = \tfrac{1}{2}x. \tag{1}$$

Gauss analyzed motion in astronomical orbits into harmonics and indeed utilized the fast algorithm now favored for computing. Much earlier in Roman times Claudius Ptolemy expressed motion in planetary orbits by the geometrical equivalent of trigonometric series and, according to Neugebauer (1983), the idea of epicycles has roots in Mesopotamian astronomy where the solar motion was matched by zigzag functions, rough approximations of the sinusoids to come.

1 The Fourier Transform

There are many transforms, each characterized by its own explicit operator, which we may call **T**. The operand, or entity operated on, is a function such as $f(x)$, where x is a real variable ranging from $-\infty$ to ∞. The notation $\mathbf{T}\{f(x)\}$ signifies the outcome of applying the operator **T** to the function $f(x)$. To illustrate, the operation that converts a given function $f(x)$ to its Fourier transform, which is a different function $F(s)$, is as follows: "Multiply the function $f(x)$ by $\exp(-i2\pi sx)$ and integrate with respect to x from $-\infty$ to ∞." Applying this operation to $f(x) = \exp(-|x|)$ we find that $\mathbf{T}\{f(x)\} = F(s) = 2/[1 + (2\pi s)^2]$, which is the Fourier transform of $\exp(-|x|)$. The symbolic expression of the Fourier transform operation is

$$F(s) = \int_{-\infty}^{\infty} f(x)e^{-i2\pi sx}\,dx. \tag{2}$$

It is apparent that any particular value of $F(s)$ [for example, $F(2)$, which equals 0.0126] takes into account the whole range of x; that is, the value depends on the shape of $f(\ )$ as a whole, not on any single point. Thus the Fourier operation is quite unlike the operation that converts $f(x) = \exp(-|x|)$ to $\sin[\exp(-|x|)]$; the outcome of this latter operation is referred to as a "function of a function," and the resulting values each depend on only a single value of x. When the result depends on the shape of $f(x)$ over part or all of the range of x, an entity such as $F(s)$ is called a functional of $f(\ )$. The variable s is called the transform variable and may have a physical meaning; if so, its units will be cycles per unit of

Much of this material was published earlier in *Science*, **248**, 697–704, 1990.

Tab. 1 Selected Fourier transforms. The quantity a is a constant

$f(x)$	$F(s)$
$e^{-\|x\|}$	$2/[1+(2\pi s)^2]$
$\delta(x)$	1
$\cos(2\pi x/a)$	$\frac{1}{2}\delta(s+a^{-1})+\frac{1}{2}\delta(s-a^{-1})$
rect x	sinc s
$e^{-\pi x^2}$	$e^{-\pi s^2}$
$e^{-\pi(x/a)^2}$	$\|a\|e^{-\pi(as)^2}$
$f(x/a)$	$\|a\|\,F(as)$
$f(x+a)$	$e^{i2\pi as}F(s)$
$f'(x)$	$i2\pi sF(s)$
Autocorrelation of $f(x)$	$\|F(s)\|^2$
$\int_{-\infty}^{\infty} f(x-u)g(u)\,du$	$F(s)G(s)$

x. A short list of Fourier transforms for illustration is shown in Table 1.

In this list rect x is the unit rectangle function (equal to unity where $|x| < 0.5$, else-where zero) and sinc $x = (\sin \pi s)/\pi s$. The last five lines are representative theorems of the form, "If $f(x)$ has Fourier transform $F(s)$, then [modification of $f(x)$] has transform [modification of $F(s)$]." Extensive lists of such transform pairs and theorems are available from the reference texts; the short list given would cover a sizable fraction of the analytic forms encountered in the literature.

With some transforms – the Abel transform is an example – each transform value depends on only a part of, not all of, $f(\ )$; and with other transforms the transform variable does not necessarily have a different identity (as s is different from x) but may have the same identity (Hilbert transform). The integral from $-\infty$ to x is a transform with both of the above restrictive properties.

All the transforms dealt with here are linear transforms, which are the commonest type; they all obey the superposition rule that $\mathbf{T}\{f_1(x)+f_2(x)\} = \mathbf{T}\{f_1(x)\}+\mathbf{T}\{f_2(x)\}$ for any choice of the given functions $f_1(x)$ and $f_2(x)$. An example of a nonlinear transformation is provided by $\mathbf{T}\{f(x)\} = a + bf(x)$, as may be tested by reference to the superposition definition; clearly the term linear in "linear transform" does not have the same meaning as in Cartesian geometry.

2 Continuous Versus Discrete Transforms

Before defining the main transforms succinctly by their operations $\mathbf{T}$, all of which involve integration over some range, it is worth commenting on a numerical aspect. One could take the point of view, as is customary with numerical integration, that the desired integral is an entity in its own right; that the integral may on occasion be subject to precise evaluation in analytic terms, as with $F(s) = 2/[1+(2\pi s)^2]$; and that if numerical methods are required a sum will be evaluated that is an approximation to the desired integral. One would then discuss the desired degree of approximation and how to reach it. Now this is quite unlike the customary way of thinking about the discrete Fourier transform. What we evaluate is indeed a sum, but we regard the sum as precise and not as an approximation to an integral. There are excellent reasons for this. Meanwhile, the important thing to realize is that there are both a Fourier transform and a discrete Fourier transform, each with its own definition. The discrete Fourier transform operation is

$$F(\nu) = \frac{1}{N}\sum_{\tau=0}^{N-1} f(\tau)e^{-i2\pi\nu\tau/N}. \qquad (3)$$

The word "discrete" is used in antithesis to "continuous," and in the cases

discussed here means that an independent variable assumes integer values. In order to understand the discrete Fourier transform, which is exclusively what we compute when in numerical mode, it is best to forget the Fourier integral and to start afresh. Instead of starting with a complex function $f(x)$ that depends on the continuous real variable x, we start with N data (complex in general, but often real) indexed by an integer serial number τ (like time) that runs from 0 to $N-1$. In the days when FORTRAN did not accept zero as a subscript, summation from $\tau = 0$ caused much schizophrenia, but the mathematical tradition of counting from zero prevailed and is now unanimous. In cases where $f(\,)$ is a wave form, as it often is, the quantity τ can be thought of as time that is counted in units starting from time zero. Clearly, N samples can never fully represent $\exp(-|x|)$, for two reasons: the samples take no account of the function where x exceeds some finite value, and no account is taken of fine detail between the samples. Nevertheless, one may judge that, for a given particular purpose, 100 samples will suffice, and the confidence to judge may be bolstered by trying whether acquisition of 200 samples significantly affects the purpose in hand. Numerical intuition as developed by hand calculation has always been a feature of mathematical work but was regarded as weak compared with physical intuition. Nowadays, however, numerical intuition is so readily acquired that it has become a matter of choice whether to attack questions about the size of N by traditional analytic approaches. A new mix of tools from analysis, finite mathematics, and numerical analysis is evolving.

The discrete transform variable ν reminds us of frequency. If τ is thought of as time measured in integral numbers of seconds, then ν is measured in cycles per second, and is indeed like frequency (c/s or Hz), but not exactly. It is ν/N that gives correct frequencies in Hz, and then only for $\nu \leq N/2$. Where ν exceeds $N/2$ we encounter a domain where the discrete approach conflicts with the continuous. When the Fourier transform is evaluated as an integral, it is quite ordinary to contemplate negative values of s, and a graph of $F(s)$ will ordinarily have the vertical $s = 0$ axis in the middle, giving equal weight to positive and negative "frequencies." (The unit of s is always cycles per unit of x; if x is in meters, s will be a spatial frequency in cycles per meter; if x is in seconds, s will be a temporal frequency in cycles per second, or Hz.) However, the discrete Fourier transform, as conventionally defined, explicitly requires the transform variable ν to range from 0 to $N-1$, not exhibiting negative values at all. There is nothing wrong with that, but persons coming from continuous mathematics or from physics may like to know that, when ν is in the range from $N/2$ to $N-1$, the quantities $N-\nu$ correspond to the negative frequencies familiar to them as residing to the left of the origin on the frequency axis. This is because the discrete transform is periodic in ν, with period N.

In the familiar Fourier series

$$p(x) = a_0 + \sum_1^\infty (a_\nu \cos 2\pi\nu x + b_\nu \sin 2\pi\nu x), \qquad (4)$$

for a periodic function $p(x)$ of period 2π, the first term a_0 represents the direct-current, zero-frequency, or mean value over one period as calculated from

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} p(x)\,dx.$$

So the first term $F(0)$ of the discrete Fourier transform is the average of the

N data values. This is the reason for the factor $1/N$ in front of the summation sign in Eq. (3), a factor that must be remembered when checking. In practical computing it is efficient to combine the factor $1/N$ with other factors such as calibration factors and graphical scale factors that are applied later at the display stage. The remaining Fourier coefficients, given by

$$a_\nu = \frac{1}{\pi}\int_0^{2\pi} p(x)\cos 2\pi\nu x\,dx,$$

$$b_\nu = \frac{1}{\pi}\int_0^{2\pi} p(x)\sin 2\pi\nu x\,dx,$$

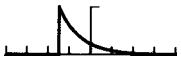
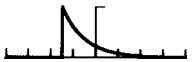
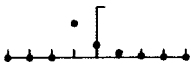
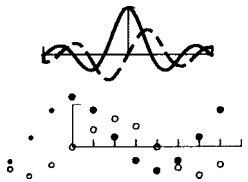
are related to the discrete Fourier transform by $a_\nu - ib_\nu = F(\nu)$. The minus sign arises from the negative exponent in the Fourier kernel $e^{-i2\pi sx}$. The reason for the choice of the negative exponent is to preserve the convention that d/dt be replaceable by $+i\omega$ in the solution of linear differential equations, as when the impedance of an inductance L to alternating voltage of angular frequency ω is written $+i\omega L$ (more usually $j\omega L$).

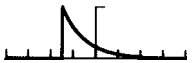
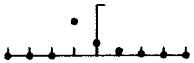
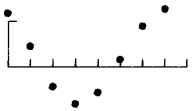
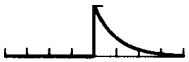
How to decide whether the discrete Fourier transform is an adequate approximation to the Fourier transform is a very interesting question. But the question itself is open to challenge. If I am studying cyclicity in animal populations, perhaps seasonal influence on bird migration, I may start with 365 reports of how many birds were seen each day of the year. In such a case, and in many other cases, discrete data mean that the integrals, even though convenient, are themselves the approximations; the discrete Fourier transform, given N equispaced data, is a valid entity in its own right. Unexpected discrepancies may arise, however, over the choice of N, which may be taken too large or too small. Among the bad consequences are slow computing (N too large), unwanted sensitivity to measurement error (N too small), and aliasing. Aliasing is the word for the following phenomenon. Measurements are made of some time-varying phenomenon at regularly spaced time intervals – perhaps temperature is recorded twice a day or perhaps speech samples are taken at a 10-kHz rate. Such data can represent harmonic components with period longer than one day or longer than 2×10^{-4}s, but cannot faithfully follow faster harmonic variation. The samples will not ignore the presence of such high frequencies, because the high-frequency variations will indeed be sampled, but the samples will be consistent with, and indistinguishable from, a long-period sinusoidal component that is not actually present. The imperfectly sampled component emerges under the alias of a lower, counterfeit frequency.

3 Some Common Transforms

As a convenient reference source, definitions of several transforms (Laplace, Fourier, Hartley, Mellin, Hilbert, Abel, Hankel, Radon) are presented in Table 2. When one has the transform, there is a way of returning to the original function in all the cases chosen. In some cases the inverse operation $\mathbf{T}^{-1}$ is the same as the defining operation $\mathbf{T}$ (e.g., Hartley and Hilbert, which are reciprocal transforms), but the majority differ, as shown. In addition, examples of each transform are presented. These will be found to convey various general properties at a glance and may be helpful for numerical checking.

Tab. 2 Transform definitions, inverses, and examples. (Ticks are at unit spacing.)

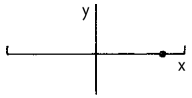
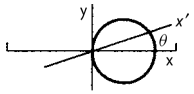
Name of transform	*Nature of function domain variable*	*Example of function*	*Nature of transform variable*	*Nature of transform*	*Example of transform*	*Defining formula and the inverse*
Laplace	Cont., real	$f(x) = e^{-x-1.5}H(x+1.5)$ 	Cont., complex	Complex	$\frac{e^{1.5s}}{1+s}$, $-1 < \mathrm{Re}\, s$	$F_L(s) = \int_{-\infty}^{\infty} f(x)e^{-sx}\,dx$ $\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} F_L(s)e^{xs}\,ds$
Fourier	Cont., real	$f(x) = e^{-x-1.5}H(x+1.5)$ 	Cont., real	Complex	$F(s) = \frac{e^{3\pi s}}{1+i2\pi s}$	$F(s) = \int_{-\infty}^{\infty} f(x)e^{-i2\pi sx}\,dx$ $f(x) = \int_{-\infty}^{\infty} F(s)e^{i2\pi xs}\,ds$
Discrete Fourier	Discrete, real	$f(\tau) = e^{-\tau-1.5}H(\tau+1.5)$ 	Discrete, real	Complex		$F(\nu) = N^{-1}\sum_{\tau=0}^{N-1} f(\tau)e^{-i2\pi\nu\tau}$ $f(\tau) = \sum_{\nu=0}^{N-1} F(\nu)e^{i2\pi\tau\nu}$

Hartley[a]	Cont., real	$f(x) = e^{-x-1.5}H(x+1.5)$ 	Cont., real	Real	$H(s) = \frac{\mathrm{cas}(-3\pi s)+2\pi s\,\mathrm{cas}\,3\pi s}{1+4\pi^2 s^2}$	$H(s) = \int_{-\infty}^{\infty} f(x)\mathrm{cas}2\pi sx\,dx$ $f(x) = \int_{-\infty}^{\infty} H(s)\mathrm{cas}2\pi xs\,ds$
Discrete Hartley	Discrete, real	$f(\tau) = e^{-\tau-1.5}H(\tau+1.5)$ 	Discrete, real	Real		$H(\nu) = N^{-1}\sum_{0}^{N-1} f(\tau)\mathrm{cas}2\pi\nu\tau$ $f(\tau) = \sum_{\nu=0}^{N-1} H(\nu)\mathrm{cas}2\pi\nu\tau$
Mellin	Cont., real	$f(x) = e^{-x}H(x)$ 	Discrete, real	Real	s: 0 1 2 3 $F_M(s)$: 1 1 2 6	$F_M(s) = \int_0^{\infty} f(x)x^{s-1}\,dx$ $f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+\infty} F_M(s)x^{-s}\,ds$
Hilbert	Cont., real	$f(x) = e^{-\pi x^2}\cos 4\pi x$	Cont., complex	Real	$F_{\mathrm{Hi}}(x) = -e^{-\pi x^2}\sin 4\pi x$	$F_{\mathrm{Hi}}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(x')\,dx'}{x'-x}$ $f(x) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{F_{\mathrm{Hi}}(x')\,dx'}{x'-x}$

continued overleaf

[a]The function $\mathrm{cas}x = \cos x + \sin x$.

Tab. 2 *Continued*

Name of transform	*Nature of function domain variable*	*Example of function*	*Nature of transform variable*	*Nature of transform*	*Example of transform*	*Defining formula and the inverse*
Abel	Cont., real	$f(r) = 1,\ r < 0.5$	Cont., complex	Real	$F_A(s) = \sqrt{1 - s^2},\ s < 1,$	$F_A(s) = 2\int_s^\infty \frac{f(r)rdr}{\sqrt{r^2 - s^2}}$ $f(r) = \frac{1}{\pi}\int_r^\infty \frac{f_A'(x)dx}{\sqrt{x^2 - r^2}}$
Hankel	Cont., real	$f(r) = 1,\ r < 0.5$	Cont., complex	Real	$F_{\mathrm{Ha}}(s) = jinc\ s\frac{J_1(\pi s)}{2s}$	$F_{\mathrm{Ha}}(s) = 2\pi\int_0^\infty f(r)J_0(2\pi sr)rdr$ $f(r) = 2\pi\int_0^\infty F_{\mathrm{Ha}}(s)J_0(2\pi rs)sds$
Radon	2-D cont., real	$f(x, y) = \delta(x - a)\ \delta(y)$ 	2-D, cont., real	Real	$F_{\mathrm{Ra}}(x', \theta) = \delta(x' - a\cos\theta)$ 	$F_{\mathrm{Ra}}(x', \theta) = \int_{-\infty}^{\infty} f(x, y)\,dy'$ Complicated inverse

4
The Laplace Transform

A long and diverse history (Deakin, 1985) characterizes the Laplace transform, which was in use long before Laplace, but became known to current generations mainly through its pertinence to the linear differential equations of transient behavior in electricity and heat conduction. Many tough technological problems of electric circuits that arose in connection with telegraphy, submarine cables, and wireless, and related industrial-process problems of thermal diffusion, were cracked around the turn of the century, sometimes by novel methods such as those of Heaviside (1970), which were to be justified subsequently (Nahin, 1987) to the satisfaction of academic mathematics by systematic application of the Laplace transform. Heaviside is remembered for stimulating the application of the Laplace transform to convergence of series and for Maxwell's equations, the delta function, the Heaviside layer, impedance, non-convergent series that are useful for computing, fractional order derivatives and integrals, and operational calculus.

Table 2 gives, as an example, the Laplace transform of $f(x) = \exp(-x - 1.5)\mathrm{H}(x + 1.5)$. The Heaviside unit step function $\mathrm{H}(x)$ jumps, as x increases, from 0 to 1, the jump being where $x = 0$; one of its uses is as a multiplying factor to allow algebraic expression of functions that switch on. The transform of $f(x)$, which is easy to verify by integration, is $(\exp 1.5s)/(1 + s)$; the transform variable s may be complex but must lie among those numbers whose real parts are greater than -1 (otherwise the integral does not exist). It is rather cumbersome to exhibit the complex transform graphically on the complex plane, and so an illustration is omitted. To invert the transform requires integration on the complex plane along a semicircular contour with indentations if necessary to circumvent points where the integrand goes to infinity (poles). The constant c in the inversion formula is to be chosen to the right of all poles.

To some extent Laplace transforms were computed numerically, but more typically, development led to compilations of analytic transforms resembling the tables of integrals (Erdélyi et al., 1954; Campbell and Foster, 1948). Programs for deriving the Laplace transform of the impulse response from electrical networks given diagrammatically are also available. Consequently it is hardly ever necessary to derive Laplace transforms analytically today. The analytic solution of transients in electric circuits, a subject traditionally used for sharpening the minds of electrical engineers, is obsolescent because impulse responses and transfer functions have been concisely published (McCollum and Brown, 1965). Furthermore, the advent of integrated circuits has meant that inductance is seldom included in new designs, and that circuits containing more than two or three elements have become less common. Mature programs are also available for step-by-step integration of circuit differential equations.

On the numerical side the Laplace transform has also been largely eroded by use of the Fourier transform. This is because angular frequency ω is a real quantity, mathematically, and it ought to be possible to compute the behavior of an electrical, acoustical, or mechanical system without reference to a complex frequency $\omega - i\sigma$. Certainly the Laplace transform is computable over its strip of convergence from any single slice therein. Nevertheless practitioners of control theory find it convenient to think on the complex plane

of s in terms of poles and zeros that are off the real frequency axis, and theirs is one tradition that keeps the complex plane alive; the convenience stems from the fact that the Laplace transform is analytic, and thus specifiable by its poles and isolated zeroes. There are problems that used to be handled by the Laplace transform, paying strict attention to the strip of convergence, because the Fourier integral did not converge; but these situations are now universally handled by Fourier methodology with the aid of delta-function notation for impulses and their derivatives, and no longer call for special treatment. When it comes to discrete computing, the impulse, and its associated spectrum reaching to indefinitely large frequencies, may in any case be forgotten. Thus, it has been wondered (Körner, 1988) "whether the Laplace transform will keep its place in the standard mathematical methods course for very much longer," but it will never die out; a new balance between curricular segments will be struck.

5
Convergence Conditions

Much attention used to be given to the existence of the Fourier integral because of paradoxes with such wanted entities as $f(x) = 1, f(x) = \cos x$, or $f(x) = \delta(x)$, where $\delta(x)$ is the unit impulse at $x = 0$, none of which possessed a Fourier integral. Today we reason as a physicist would, recognizing that a voltage waveform cannot have a value of 1 V forever, but must have turned on at some time in the past and will turn off at some time in the future. The finite-duration function does have a Fourier transform. We then consider a sequence of waveforms of longer and longer duration and the corresponding sequence of transforms, arriving at the concept of "transforms in the limit." This attitude has received mathematical respectability under the rubric of generalized functions (Lighthill, 1958) and is the basis for saying that the Fourier transform of $\delta(x)$ is 1 [while conversely the Fourier transform of 1 is $\delta(s)$]. The elaborate conditions for the existence of a transform when generalized functions were excluded have thus lost interest. Even $\delta'(x)$ now has the indispensable transform $i2\pi s$; under the rules of analysis $\delta'(x)$ was an unthinkable entity – certainly not qualifying as a function of x; to physicists it was a commonplace dipole, and in mechanics a local load such as a moment applied at a point on a beam.

The fact that the Laplace integral converged when the Fourier transform did not gave the Laplace transform a certain prestige, even though convergence was achieved at the cost of tapering the given function by a real, exponentially decaying factor. In addition, the strip of convergence had to be specified for the complex transform variable s. The convenience of dealing with the real and physically intuitive frequency as the transform variable has shifted preference in favor of the Fourier and Hartley transforms. The only effective condition for the existence of a Fourier or Hartley transform today is that the given function should have a physical interpretation, or be representable by a sequence of physically interpretable functions whose individual transforms approach a limit. Consequently it is no longer necessary to require that $f(x)$ be absolutely integrable ($\int_{-\infty}^{\infty} |f(x)|\, dx$ exists) or that any discontinuities be finite; on the contrary, the "shah function" $\mathrm{III}(x) = \Sigma_{n=-\infty}^{n=\infty} \delta(x - n)$, which could be said to possess an infinite number of infinite discontinuities, now has a Fourier transform thanks to the theory of generalized functions (Bracewell,

1956). Interestingly, the Fourier transform of III(x) is III(s).

The function $\sin(x^{-1})$ raises a convergence question as a result of possessing an infinite number of maxima in any interval containing $x = 0$; this sort of behavior is without interest in the world of numerical computing but of considerable interest to the theory of integration. Possession of an infinite number of maxima does not in itself define the convergence condition because the Fourier integral may converge if the amplitude of the oscillation dies down so that the function exhibits bounded variation. Nor does bounded variation define the convergence condition because Lipschitz has demonstrated functions of unbounded variation whose Fourier integrals converge. However, the Lipschitz condition is not the ultimate convergence condition, as has been shown by Dini (Bracewell, 1986a). This style of analysis has lost practitioners as activity has moved in the direction of finite, or discrete, mathematics.

6
Why Transforms Are Useful

Many problems can be posed in the form of a differential equation (or a difference equation, or an integral equation, or an integro-differential equation) that has to be solved for some wanted function subject to stated boundary conditions or initial conditions. Laplace's equation in three dimensions describes the potential distribution set up by an array of electric charges, and the diffusion equation describes the heat flow distribution set up by a given distribution of heat. By applying a transformation such as the Laplace or Fourier to each term of such an equation, we arrive at a new equation that describes the transform rather than the original wanted function. The interesting thing about this is that the new equation may be simpler, sometimes solvable just by algebra. We solve that equation for the transform of the solution, and then invert. Not all differential equations simplify in this way; those that do are characterized by linearity and coordinate invariance (such as time invariance), and the presence of these characteristics in nature is responsible for a good deal of the numerical activity with transforms. Transfer functions, such as the frequency response curves of amplifiers, are corresponding manifestations of these same characteristics. The passage of a speech waveform through an amplifier is described by a differential equation that may be hard to solve; but having used a Fourier transform to go to the frequency domain, we apply the transfer function, frequency by frequency, by complex multiplication to get the transform of the output. Then retransforming gives the output waveform.

There is also a differential equation, describing the bending of a beam under the influence of a load distribution, that may be thought of as a spatial input analogous to an input waveform, while the curve of deflection is analogous to the output waveform. Although Hooke's law, the first of the linear laws, may apply, we do not use transform methods. If we analyze the load distribution into spatially sinusoidal components and find the bending response to each component, and linearly sum the responses, we will get the desired shape of the bent beam, but there is no transfer function to facilitate getting the individual responses by simple algebra. The reason is that we have linearity but not space invariance – if we shift the load, the response does not shift correspondingly without change

of shape; a sinusoidal load does not produce sinusoidal deflection. If, on the contrary, we delay the input to an amplifier or a vibratory mechanical system, the response is correspondingly delayed but is unchanged as to shape; furthermore, a sinusoidal input produces a sinusoidal output.

7
Fields of Application

Fourier (Grattan-Guinness, 1972) originally thought of representing the temperature on a heat-conducting bar as a sum of sinusoids. To avoid a problem of integration he considered the bar to be bent around on itself in a large circle, a distortion that is not harmful to the discussion of any given finite straight bar because the arc of interest can be made as straight as you wish by taking the circle large enough. Since the temperature distribution on the ring is of necessity now periodic in space, only a fundamental and harmonics need be considered, plus the constant temperature a_0 representing the mean temperature. As time elapses, the temperature distribution varies as the heat flows under the influence of the temperature gradients, ultimately approaching the uniform value a_0 in the limit. Fourier found that the component sinusoids decay exponentially with a time constant proportional to the spatial period, or wavelength, the nodes of each sinusoid remaining fixed. By attenuating each component in accordance with the elapsed time, and summing, one gets the same result as if the spatially variable heat flow were followed in real time. This is an example of the duality of the function domain (space domain in this instance) and the transform domain (spatial frequency domain) that permeates Fourier applications.

Music can be thought of in terms of the wave form of the wave that conveys the sound through the air (function domain), or in terms of the harmonic constituents (spectral domain) that are separately discernible by the ear and are treated separately by an amplifier. In crystallography there is the arrangement of the atoms in space (crystal lattice domain) and the spatial Fourier components (reciprocal lattice domain) which, under illumination by x rays or neutron beams, evidence themselves by diffraction at defined angles. Image formation with cameras and radio telescopes can be conceived as operating on the object domain, or "sky plane," or we can think in terms of complex coherence measurements in the transform domain. All these dual modes of thought, under their respective terminologies, are fully equivalent; it helps to be familiar with both and to be able to translate from one domain to the other. In addition, it is most helpful to be able to translate between fields, converting a problem in one subject into the analogous problem in another subject where the solution may be intuitively obvious. As an example, persons who know very well that the diffraction from a pair of side-by-side pinholes is sinusoidal in space may not know that the spectrum of a pair of audible clicks in succession is sinusoidal in frequency. How much richer this knowledge becomes when they are able to translate from acoustics to optics and vice versa!

As a formal illustration of the methodology let us calculate the response of an electric circuit consisting of an inductance L in series with a resistance R to which a voltage impulse of strength A is applied at $t = -1.5$. Equating the sum of the voltages in the circuit to zero, as taught by

Kirchhoff, gives the differential equation

$$A\delta(t) = L\frac{di}{dt} + Ri,$$

where $i(t)$ is the current flow in response to the applied voltage. Taking the Fourier transforms term by term (Table 1) we find that

$$Ae^{i2\pi\times 1.5s} = i2\pi sLI(s) + RI(s),$$

where $I(s)$ is the transform of the wanted current. Solving this algebraic equation gives

$$I(s) = \frac{Ae^{i2\pi\times 1.5s}}{R + i2\pi Ls},$$

and taking the inverse Fourier transforms of both sides gives

$$i(t) = \frac{A}{L}e^{-R(t+1.5)/L}\mathrm{H}(t+1.5).$$

The transform involved is illustrated in Table 2 for the Fourier transform. The method for solving the same problem by the Laplace transform is similar but involves reference to convergence of the integral, a complication that is circumvented when generalized function theory is combined with the Fourier integral.

Newton showed how to split sunlight into its constituent colors with a prism, where we think in the spatial domain, but there is another way that we learned from Michelson that is explicable in the time domain. We split a beam of light, and then recombine the two beams on a photodetector, but not before a controlled delay is introduced into one of the beams, for example, by retroreflection from a movable plane mirror. The detector output reveals the autocorrelation of the light beam from which, by using the autocorrelation theorem (Table 1) and numerical Fourier transformation, we get the spectral distribution of power.

8
The Hartley Transform

Table 2 illustrates by example that the Fourier transform in general is a complex function of the real transform variable s; consequently two transform curves must be drawn, one for the real part and one (broken) for the imaginary part. The example $f(\tau)$ for the discrete Fourier transform is based on samples of the previous $f(x)$. Imaginary values of the discrete transform $F(\nu)$ are shown as hollow circles. Three features may be noted: no matter how closely samples are spaced, some detail can be missed; no outlying parts beyond a finite range are represented; the indexing convention 0 to $N-1$ has the effect of cutting off the left side of $F(s)$, translating it to the right, and reconnecting it. To convey the nature of this third comment, the points for $\tau > N/2$ have been copied back on the left.

The Hartley transform differs from the Fourier transform in that the kernel is the real function cas $2\pi sx$ instead of $\exp(-i2\pi sx)$. The cas function, which was introduced by Hartley (1942), is defined by $\mathrm{cas}x = \cos x + \sin x$ and is simply a sinusoid of amplitude $\sqrt{2}$ shifted one-eight of a period to the left. The consequences of the change are that the Hartley transform is real rather than complex and that the transformation is identical to the inverse transformation. As may be obvious from the graphical example, the Hartley transform contains all the information that is in the Fourier transform and one may move freely from one to the other using the relations

$$H(s) = \mathrm{Re}\ F(s) - \mathrm{Im}\ F(s)$$

and

$$2F(s) = H(s) + H(N-s) - iH(s) + iH(N-s).$$

The convenience that arises from familiarity with complex algebra when one is thinking about transforms loses its value in computing. What one thinks of compactly as one complex product still means four real multiplications to computer hardware, which must be instructed accordingly.

The Hartley transform is fully equivalent to the Fourier transform and can be used for any purpose for which the Fourier transform is used, such as spectral analysis. To get the power spectrum from the complex-valued Fourier transform one forms $[\mathrm{Re} f(s)]^2 + [\mathrm{Im} F(s)]^2$; starting from the real-valued Hartley transform one forms $[H(s)]^2 + [H(-s)]^2$. The phase is obtained from

$$\tan\phi(s) = \frac{\mathrm{Im}(s)}{\mathrm{Re}(s)} = \left[\frac{H(-s)}{H(s)}\right] - \frac{\pi}{4}.$$

We see that for purposes of spectral analysis by the Hartley transform it is not necessary to work with complex quantities, since power spectrum is an intrinsic property independent of choice of kernel; the phase depends on the x origin which is locked to the peak of the cosine function in one case and the peak of the cas function in the other, hence the term $\pi/4$.

9
The Fast Fourier Transform

Around 1805 C.F. Gauss, who was then 28, was computing orbits by a technique of trigonometric sums equivalent to today's discrete Fourier synthesis. To get the coefficients from a set of a dozen regularly spaced data he could if he wished explicitly implement the formula that we recognize as the discrete Fourier transform. To do this he would multiply the N data values $f(\tau)$ by the weighting factors $\exp(-i2\pi\nu\tau)$, sum the products, and repeat these N multiplications N times, once for each value of ν, making a total of N^2 multiplications. But he found that, in the case where N is a composite number with factors such that $N = n_1 n_2$, the number of multiplications was reduced when the data were partitioned into n_2 sets of n_1 terms. Where N was composed of three or more factors a further advantage could be obtained. Gauss (1876) wrote, "illam vero methodum calculi mechanici taedium magis minuere, praxis tentatem docebit." He refers to diminishing the tedium of mechanical calculation, as practice will teach him who tries. This factoring procedure, usually into factors of 2, is the basis of the fast Fourier transform (FFT) algorithm, which is explained in many textbooks (Bracewell, 1986a; Elliott and Rao, 1982; IEEE, 1979; Nussbaumer, 1982; Press et al., 1986; Rabiner and Gold, 1975) and is available in software packages. The fast method (Cooley and Tukey, 1965) burst on the world of signal analysis in 1965 and was for a time known as the Cooley-Tukey algorithm (IEEE, 1967), but as the interesting history (Heideman et al., 1985) of prior usage in computing circles became known the term FFT became universal.

Most FFT programs in use take advantage of factors by adopting a choice of N that is some power P of 2, i.e., $N = 2^P$. The user may then design the data collection to gather, for example, $256 = 2^8$ readings. Alternatively, when such a choice does not offer, a user with 365 data points can simply append sufficient zeros to reach $512 = 2^9$ values. This might seem wasteful, but an attendant feature is the closer spacing of the resulting transform samples, which is advantageous for visual

presentation. Perhaps one could do the job faster, say by factoring into 5×73. There are fast algorithms for 5 points and for many other small primes, but not for 73, as far as I know; it is simply not practical to store and select from lots of special programs for peculiar values of N. On the other hand, a significant speed advantage is gained if one elects more rigidity rather than more flexibility, tailors one's data collection to a total of 4^P values, and uses what is referred to as a radix-4 program. Since $1024 = 4^5$, the radix-4 approach is applicable to $N = 1024$ data samples (or to 256 for example), but not to 512 unless one appends 512 zeros. Packing with just as many zeros as there are data is commonly practised because twice as many transform values result from the computation, and when the power spectrum is presented graphically as a polygon connecting the computed values the appearance to the eye is much smoother.

Much practical technique is involved. If the sound level of an aircraft passing over a residential area is to be recorded as a set of measurements equispaced in time, the quantity under study begins and ends at zero value. But in other cases, such as a record of freeway noise, the noise is present when measurements begin and is still there when they cease; if the N values recorded are then packed with zeros, a discontinuity is introduced whose effects on the transform, such as overshoot and negative-going oscillation, may be undesirable. Packing with plausible (but unobserved) data can eliminate the undesired artifacts and is probably practised in more cases than are admitted to. Authors often mitigate the effects of implied discontinuities in the data by multiplying by a tapering function, such as a set of binomial coefficients, that approaches zero at both the beginning and end of the data taken; they should then explain that they value freedom from negatives more than accuracy of amplitude values of spectral peaks or than resolution of adjacent peaks.

The FFT is carried out in P successive stages, each entailing N multiplications, for a total of NP. When NP is compared with N^2 (as for direct implementation of the defining formula) the savings are substantial for large N and make operations feasible, especially on large digital images, that would otherwise be unreasonably time consuming.

10
The Fast Hartley Algorithm

When data values are real, which is very commonly the case, the Fourier transform is nevertheless complex. The N transform values are also redundant (if you have the results for $0 \leq \nu \leq N/2$ you can deduce the rest). This inefficiency was originally dealt with by the introduction of a variety of efficient but unilateral algorithms that transformed in half of the time of the FFT, albeit in one direction only; now we have the Hartley transform, which for real data is itself real, is not redundant, and is bidirectional. The Hartley transform is elegant and simple and takes you to the other domain, regardless of which one you are in currently (Bracewell, 1986b; Buneman, 1989).

When a Hartley transform is obtained, there may be a further step required to get to the more familiar complex Fourier transform. The time taken is always negligible, but even so the step is usually unnecessary. The reason is that although we are accustomed to thinking in terms of complex quantities for

convenience, it is never obligatory to do so. As a common example, suppose we want the power spectrum, which is defined in terms of the real and imaginary parts of the Fourier transform by $P(\nu) = [\mathrm{Re}F(\nu)]^2 + [\mathrm{Im}F(\nu)]^2$. If we already have the Hartley transform $H(\nu)$, then it is not necessary to move first to the complex plane and then to get the power spectrum; the desired result is obtained directly as $\{[H(\nu)]^2 + [H(N-\nu)]^2\}/2$. Likewise phase $\phi(\nu)$, which is required much less often than $P(\nu)$, is defined by $\tan[\phi(\nu)] = \mathrm{Im}F(\nu)/\mathrm{Re}F(\nu)$; alternatively, one can get phase directly from $\tan[\phi(\nu) + \pi/4] = H(N-\nu)/H(\nu)$, thus circumventing the further step that would be necessary to go via the well-beaten path of real and imaginary parts.

To illustrate the application to power spectra take as a short example the data set {1 2 3 4 5 6 7 8}, whose discrete Hartley transform is

$$H(\nu) = \{4.5 \quad -1.707 \quad -1 \quad -0.707 \quad -0.5 \quad -0.293 \quad 0 \quad 0.707\}.$$

The first term, 4.5, is the mean value of the data set. The power spectrum for zero frequency is 4.5^2, for frequency $1/8(\nu = 1)$, $P(1) = (-1.707)^2 + (0.707)^2$, for frequency $2/8(\nu = 2)$, $P(2) = (-1)^2 + 0^2$. Similarly $P(3) = (-0.707)^2 + (-0.293)^2$ and $P(4) = (-0.5)^2 + (0.5)^2$. The highest frequency reached is 4/8, corresponding to a period of 2, which is the shortest period countenanced by data at unit interval.

The encoding of phase by a real transform has added a physical dimension to the interest of the Hartley transform, which has been constructed in the laboratory with light and microwaves (Villasenor and Bracewell, 1987, 1988, 1990; Bracewell, 1989; Bracewell and Villasenor, 1990) and has suggested a new sort of hologram.

11
The Mellin Transform

The vast majority of transform calculations that are done every day fall into categories that have already been dealt with and much of what has been said is applicable to the special transforms that remain to be mentioned. The Mellin transform has the property that $F_{\mathrm{M}}(n+1)$ is the nth moment of $f(x)$ when n assumes a finite number of integer values 1, 2, 3, ... The special value $F_{\mathrm{M}}(1)$ is the zeroth moment of, or area under, $f(x)$. But the transform variable does not have to be integral, or even real, so one can think of the Mellin transform as a sort of interpolate passing through the moment values. When the scale of x is stretched or compressed, for example, when $f(x)$ is changed to $f(ax)$, the Mellin transform becomes $a^{-2}F_{\mathrm{M}}(s)$, a modification that leaves the position of features on the s axis unchanged and is useful in some pattern-recognition problems.

If we plot $f(x)$ on a logarithmic scale of x, a familiar type of distortion results, and we have a new function $f(e^{-x})$ whose Laplace transform is exactly the same as the Mellin transform of $f(x)$. An equally intimate relation exists with the Fourier transform. Consequently the FFT may be applicable in numerical situations. Because of the intimate relationship with moments and with spectral analysis, Mellin transforms have very wide application. A specific example is given by the solution of the two-dimensional Laplace equation expressed in polar coordinates, namely $\partial^2 V/\partial r^2 + r^{-1}\partial V/\partial r + r^{-2}\partial^2 V/\partial\theta^2 = 0$. Multiply each term by

r^{s-1} and integrate with respect to r from 0 to ∞. We get $d^2F_M/d\theta^2 + s^2F_M = 0$. Solve this for $F_M()$ and invert the transform to get the solution. In this example, a partial differential equation is converted to a simple differential equation by the transform technique.

12
The Hilbert Transform

As the example in Table 2 shows, the Hilbert transform, or quadrature function, of a cosinusoidal wave packet is a similar, but odd, waveform sharing the same envelope. But what do we mean by the envelope of an oscillation that only touches the intuitively conceived envelope at discrete points? The Hilbert transform provides an answer in the form $\sqrt{[f(x)]^2 + [f_{Hi}(x)]^2}$. Likewise, the original wave packet reveals its phase at its zero crossings. But what is the phase at intermediate points? The Hilbert transform supplies an instantaneous phase ϕ in the form $\tan\phi = f_{Hi}(x)/f(x)$. The operation **T** for the Hilbert transform is simply convolution with $-1/\pi x$. It is known that the Fourier transform of $-1/\pi x$ is i sgn s, where sgn s is 1 for $s > 0$ and -1 for $s < 0$. Therefore, by the convolution theorem (last line of Table 1), according to which the Fourier transform of a convolution is the product of the separate Fourier transforms, it would seem that a fast Hilbert transform of $f(x)$ could be calculated as follows. Take the FFT of $f(x)$, multiply by i for $0 < \nu < N/2$ and by $-i$ for $N/2 < \nu < N$, set $F(0)$ and $F(N/2)$ equal to zero, and invert the FFT to obtain the Hilbert transform. This sounds straightforward, but the procedure is fraught with peril, for two reasons. We are proposing to multiply a given function $f(x)$ by $-1/\pi[(x + \text{const})]$ and to integrate from $-\infty$ to ∞, but we are only given N samples. The extremities of $-1/\pi x$ approach zero and have opposite signs, but there is infinite area under these tails no matter how far out we start. Consequently we are asking two oppositely signed large numbers to cancel acceptably. How can we expect satisfaction when the convolving function $-1/\pi x$ is not symmetrically situated about the extremes of the data range? The second reason is that we are asking for similar cancellation in the vicinity of the pole of $1/x$. Experience shows that satisfactory envelopes and phases only result when $f(x)$ is a rather narrow-band function. Under other circumstances an N-point discrete Hilbert transform can be defined and will give valid results free from worries about the infinities of analysis, but the outcome may not suit expectation.

An optical wave packet $\exp(-\pi t^2/T^2)$ $\sin 2\pi\nu t$ of equivalent duration T easily meets the narrow-band condition when the duration T is much greater than the wave period $1/\nu$; it has a Hilbert transform $\exp(-\pi t^2/T^2)\cos 2\pi\nu t$. The square root of the sum of the squares yields $\exp(-\pi t^2/T^2)$ for the envelope, in full accord with expectation.

13
Multidimensional Transforms

The two-dimensional Fourier and Hartley transforms are defined respectively by

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \times e^{-i2\pi(ux+vy)}\,dx\,dy,$$

$$F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \times \text{cas}[2\pi(ux+vy)]\,dx\,dy,$$

where the transform variables u and v mean spatial frequency components in the x and y directions. Work with images involves two dimensions, electrostatics and x-ray crystallography involve three, and fluid dynamics involves four. Multidimensional transforms can be handled numerically with a one-dimensional FFT subprogram, or a fast Hartley, as follows. Consider an $N \times N$ data array. Take the 1-D (one-dimensional) transform of each row and write the N transform values in over the data values. Now take the 1-D transform of each resulting column (Bracewell, 1984; Bracewell et al., 1986). In three and four dimensions the procedure is analogous (Hao and Bracewell, 1987; Buneman, 1989). Further simple steps lead to the Hartley transform and to the real and imaginary parts of the Fourier transform if they are wanted, but usually they are not; more often the quadratic content (power spectrum) suffices.

When a 2-D function has circular symmetry, as commonly arises with the response functions of optical instruments, not so much work is required, as explained below in connection with the Hankel transform. Cylindrical symmetry in 3-D is essentially the same, while spherical symmetry in 3-D is also referred to below.

14
The Hankel Transform

In two dimensions, where there is circular symmetry as expressed by a given function $f(r)$, the two-dimensional Fourier transform is also circularly symmetrical; call it $F_{\mathrm{Ha}}(s)$. It can be arrived at by taking the full 2-D transform as described earlier, or it can be obtained from a single 1-D Hankel transform as defined in Table 2. The inverse transform is identical. There is apparently no opening for the Hartley transform because in the presence of circular symmetry the 2-D Fourier transform of real data contains no imaginary part. The kernel for the Hankel transform is a zero-order Bessel function, which is a complication that hampers the FFT factoring approach, but there is an elegant sidestep around this that is explained below in connection with the Abel transform. Under spherical symmetry, the 3-D Fourier transform reduces to a different one-dimensional transform

$$4\pi \int_0^\infty f(r)\mathrm{sinc}(2sr)r^2\,dr. \qquad (5)$$

The inverse transform is identical.

To illustrate by a well-known result from optical diffraction we consider a telescope aperture $f(r)$ representable as $\mathrm{rect}(r/D)$, a two-dimensional function that is equal to unity over a circle of diameter D. The Hankel transform is $D^2\mathrm{jinc}\,Ds$, the familiar Fraunhofer diffraction field of a circular aperture. The jinc function [$\mathrm{jinc}\,x = J_1(\pi x)/2x$], which is the Hankel transform of the unit rectangle function of unit height within a radius of 0.5, has the property that jinc $1.22 = 0$; this is the source of the constant in the expression $1.22\lambda/D$ for the angular resolution of a telescope.

15
The Abel Transform

Most commonly, although not always, the Abel transform arises when a 2-D function $g(x, y)$ has circular symmetry, as given by $f(r)$. The Abel transform (Table 2) then simplifies to $F_{\mathrm{A}}(x) = \int_{-\infty}^{\infty} g(x, y)\,dy$. In other words, if the given $f(r)$ is represented by a square matrix of suitably spaced samples, then the Abel transform results

when the columns are summed. There might not seem to be any future in trying to speed up such a basic operation, apart from the obvious step of summing only half-way and doubling. However, when it is remembered that for each of $N^2/8$ matrix elements we have to calculate $\sqrt{x^2+y^2}$ to find r, and thence $f(r)$, it gives pause. The alternative is to proceed by equal steps in r rather than in y; then the oversampling near the x axis is mitigated. But the variable radial spacing of elements stacked in a column needs correction by a factor $r/\sqrt{r^2-s^2}$, which takes more time to compute than $\sqrt{x^2+y^2}$. This is an excellent case for decision by using the millisecond timer found on personal computers. Of course, if many runs are to be made, the factors $r/\sqrt{r^2-s^2}$ can be precomputed and the preparation time can be amortized over the successive runs.

Figure 1 shows a given function $g(x, y)$ and its one-dimensional projection (labeled P) which is derived by integrating along the y axis in the (x, y) plane. Integrating along the y' axis of a rotated coordinate system gives the projection P'. Now if $g(x, y)$ were circularly symmetrical, being a function $f(r)$ of r only, then the projections P and P' would be identical and equal to the Abel transform of $f(r)$. This is the graphical interpretation of the Abel transform.

Applications of the Abel transform arise wherever circular or spherical symmetry exists. As an example of the latter consider a photograph of a globular cluster of stars in the outer reaches of the galaxy. The number of stars per unit area can be counted as a function of distance from the center of the cluster; this is the projected density. To find the true volume density

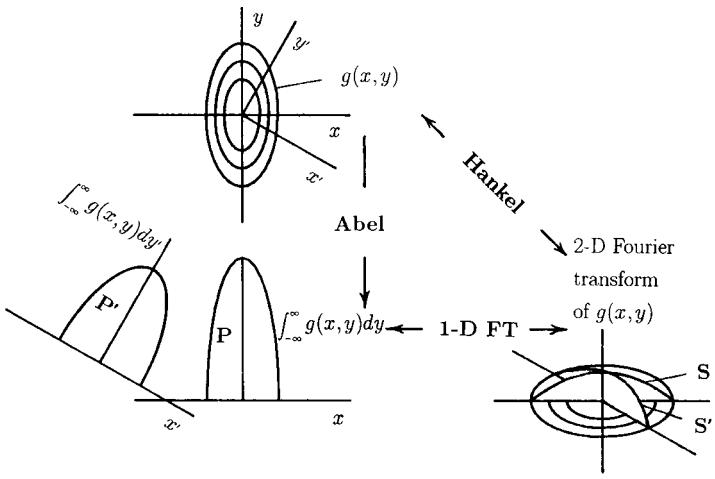


Fig. 1 Illustrating the projection-slice theorem, which states that if a distribution $g(x, y)$ has a projection P', in the y' direction, its 1-D Fourier transform is the slice S' through the 2-D Fourier transform of $g(x, y)$. The set of projections P' for all inclination angles of the (x', y') coordinates constitutes the Radon transform. In the presence of circular symmetry where $g(x, y) = f(r)$, the projection P in any direction is the Abel transform of $f(r)$. The 1-D Fourier transform of P is the slice S in any direction; this slice S is then the Hankel transform of $f(r)$. Thus the Abel, Fourier, and Hankel transforms form a cycle of transforms.

as a function of radius requires taking the inverse Abel transform (Table 2) of the projected density.

With the Abel transform under control we can now see a way of doing the Hankel transform without having to call up Bessel functions. The Abel, Fourier, and Hankel transforms form a cycle known as the FHA cycle (Bracewell, 1956), so that if we take the Abel transform and then take the FFT we get the Hankel transform; the theorem is

$$\int_0^\infty dr J_0(2\pi\xi r) r \int_{-\infty}^\infty ds e^{i2\pi rs} \times \int_s^\infty \frac{dx 2xf(x)}{\sqrt{x^2 - s^2}} = f(\xi).$$

The FFT required will not be complex, except in the extraordinary case of complex 2-D data; consequently it will in fact be appropriate to use the fast Hartley to get the Hankel transform. Because of symmetry the result will also be exactly the same as obtained with the FFT, if after taking the FFT we pay no attention to the imaginary parts that have been computed, which should all be zero or close to zero.

The FHA cycle of transforms is a special case of the projection slice theorem, a theorem which refers to the more general situation where $g(x, y)$ is not circularly symmetrical. Circular symmetry characterizes instruments, especially optical instruments, which are artifacts. Lack of symmetry characterizes data; tomographic data will be taken as the illustration for the projection-slice theorem.

16 Tomography and the Radon Transform

Consider a set of rotated coordinates (x', y') centered on the (x, y) plane, but rotated through θ. The expression $\int_{-\infty}^{\infty} g(x, y)\, dy$ given for the Abel transform, representing a line integral in the y direction at a given value of x, would equal the line integral $\int_{-\infty}^{\infty} g(x, y)\, dy'$ in the rotated direction y' provided $g(x, y)$ had circular symmetry as specified for the Abel transform. But when $g(x, y)$ does not have symmetry, then the line-integral values depend both on x' and on the angle θ (Fig. 1). The set of integrals with respect to dy' is the Radon transform of $g(x, y)$, named after Johann Radon (1917). Such integrals arise in computed x-ray tomography, where a needle-beam of x rays scans within a thin plane section of an organ such as the brain with a view to determining the distribution of absorption coefficient in that plane. If there are N^2 pixels for which values have to be determined, and since one scan will give N data, at least N different directions of scan spaced $180°/N$ apart will be needed to acquire enough data to solve for the N^2 unknowns. In practice more than $2N$ directions are helpful in order to compensate for diminished sample density at the periphery. To compute a Radon transform is easy; the only tricky part is summing a given matrix along inclined directions. One approach is to rotate all the matrix and interpolate onto a rotated grid, for each direction of scan; but this may be too costly. At the other extreme one sums, without weighting, the matrix values lying within inclined strips that, independently of inclination, preserve unit width in the direction parallel to the nearer coordinate direction. How coarse the increment inclination angle may be depends on acceptability as judged by the user in the presence of actual data.

The harder problem is to invert the line-integral data to retrieve the wanted absorption coefficient distribution. A solution was given by Radon (1917). Later Cormack (1963, 1964, 1980), working in the context

of x-ray scanning of a solid object, gave a solution in terms of sums of transcendental functions. Other solutions include the modified back-projection algorithm (Bracewell, 1956; Bracewell and Riddle, 1967) used in CAT scanners (Deans, 1983; Brooks and Di Chiro, 1976; Rosenfeld and Kac, 1982). The algorithm depends on the projection-slice theorem (see Fig. 1). According to this theorem (Bracewell, 1956) the 1-D Fourier transform of the projection P' (or scan) of $g(x, y)$ in any one direction is the corresponding central cross section or slice S' through the 2-D Fourier transform of the wanted distribution $g(x, y)$. The proof is as follows. Let the 2-D Fourier transform of $g(x, y)$ be $G(u, v)$ as defined by

$$G(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) \times e^{-i2\pi(ux+vy)}\,dxdy.$$

Setting $v = 0$, so as to have the representation $G(u, 0)$ for the slice S, we get

$$\begin{aligned} G(u, 0) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) \times e^{-i2\pi ux}dx\,dy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x, y)\,dy\right] \times e^{-i2\pi ux}\,dx \\ &= \int_{-\infty}^{\infty} P(x)e^{-i2\pi ux}\,dx, \end{aligned}$$

where $P(x)$ is the projection of $g(x, y)$ onto the x axis. Thus the 1-D transform of the projection $P(x)$ is the slice $G(u, 0)$ through the 2-D transform of $g(x, y)$. If we rotate the coordinate axes to any other orientation (x', y') we see that the same proof applies.

Because the density of polar coordinate samples is inversely proportional to radius in the Fourier transform plane, a simple correction factor followed by an inverse 2-D Fourier transform will yield the solution. But a way was found (Bracewell and Riddle, 1967; Brooks and Di Chiro, 1976), based on this theoretical reasoning, to avoid numerical Fourier transforms entirely. An equivalent correction term, arrived at by convolving each projection P' with a few coefficients, can be directly applied to each P', after which the modified projections are accumulated on the (x, y) plane by back projection to reconstitute $g(x, y)$. Back projection means assigning the projected value at x' to all points of the (x, y) plane which, in the rotated coordinate system, have the abscissa x'. Accumulation means summing the back-projected distributions for all inclination angles.

17 The Walsh Transform

A function defined on the interval (0,1) can be expressed as a sum of sines and cosines of frequency $1, 2, 3, \ldots$, but can also be expressed as a sum of many other sets of basis functions. Among the alternatives, Walsh functions (Elliott and Rao, 1982; Walsh, 1923; Hsu and Wu, 1987) are particularly interesting because they oscillate between values of $+1$, 0, and -1, a property that is most appropriate to digital circuits, telecommunications, and radar. Furthermore, multiplication by a Walsh function value takes much less time than multiplication by a trigonometric function. Walsh functions, not being periodic, are not to be confused with the periodic square cosine and sine functions $\mathrm{C}(x) = \mathrm{sgn}(\cos x)$ and $\mathrm{S}(x) = \mathrm{sgn}(\sin x)$; but on a finite support they do form a complete set from which any given function can be composed. They are also orthonormal (mutually orthogonal and with fixed quadratic content, as with Fourier components), which leads to simple relations for

both analysis and synthesis. The Walsh (or Walsh–Hadamard) transform has found use in digital signal and image processing and for fast spectral analysis. Fast algorithms are available that use only addition and subtraction and have been implemented in hardware. A vast, enthusiastic literature sprang into existence in the 1970s, a guide to which can be found in the text by Elliott and Rao (1982).

18
The *z* Transform

In control theory, in dealing with signals of the form

$$f(t) = \sum_{-\infty}^{\infty} a_n \delta(t-n) \qquad (6)$$

and systems whose response to $\delta(t)$ is

$$h(t) = \sum_{0}^{\infty} h_n \delta(t-n), \qquad (7)$$

the response $g(t)$ is the convolution integral

$$g(t) = \int_{-\infty}^{\infty} f(t')h(t-t')\,dt'. \qquad (8)$$

This response is a series of equispaced impulses whose strengths are given by $\Sigma_i a_i h_{n-i}$, an expression representable in asterisk notation for convolution by $\{g_n\} = \{a_n\} * \{h_n\}$ [in this notation the sequence $\{a_n\}$ sufficiently represents $f(t)$]. For example, a signal $\{1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \ldots\}$ applied to a system whose impulse response is $\{8 \quad 4 \quad 2 \quad 1\}$ produces a response

$$\{1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \ldots\} * \{8 \quad 4 \quad 2 \quad 1\}$$
$$= \{8 \quad 12 \quad 14 \quad 15 \quad 15 \quad 15 \ldots\}.$$

This is the same rule as that which produces the coefficients of the polynomial that is the product of the two polynomials $\Sigma a_n z^n$ and $\Sigma h_n z^n$, as may be verified by multiplying $1 + z + z^2 + z^3 + z^4 + z^5 + \cdots$ by $8 + 4z + 2z^2 + z^3$. The z transform of the sequence $\{8 \quad 4 \quad 2 \quad 1\}$ is, by one definition, just the polynomial $8 + 4z + 2z^2 + z^3$; more often one sees $8 + 4z^{-1} + 2z^{-2} + z^{-3}$. If, conversely, we ask what applied signal would produce the response $\{8 \quad 12 \quad 14 \quad 15 \quad 15 \quad 15 \ldots\}$ we get the answer by long division:

$$\frac{(8 + 12z + 14z^2 + 15z^3 + 15z^4 + 15z^5 + \cdots)}{(8 + 4z + 2z^2 + z^3)}.$$

Occasionally, one of the polynomials may factor, or simplify, allowing cancellation of factors in the numerator and denominator. For example, the z transform of the infinite impulse response $\{8 \quad 4 \quad 2 \quad 1 \quad 0.5 \ldots\}$, where successive elements are halved, simplifies to $8/(1 - z/2)$. But with measured data, or measured system responses, or both, this never happens and the z notation for a polynomial quotient is then just a waste of ink compared with straightforward sequence notation such as $\{8 \quad 12 \quad 14 \quad 15 \quad 15 \quad 15 \ldots\} * \{8 \quad 4 \quad 2 \quad 1 \ldots\}^{-1}$. Whenever sampled data are operated on by a convolution operator (examples would be finite differences, finite sums, weighted running means, finite-impulse-response filters) the z transform of the outcome is expressible as a product of z transforms. Thus to take the finite difference of a data sequence one could multiply its z transform by $1 - z$ and the resulting polynomial would be the z transform of the desired answer; in a numerical environment one would simply convolve the data with $\{1 - 1\}$. In control theory and filter design, the complex plane of z is valued as a tool for thinking about the topology of the poles and zeroes of transfer functions.

19
Convolution

Sequences to be convolved may be handled directly with available subprograms for

convolution and inverse convolution that operate by complex multiplication in the Fourier transform domain. When two real sequences are to be convolved you can do it conveniently by calling the two Hartley transforms, multiplying term by term, and calling the same Hartley transform again to get the answer. Some subtleties are involved when the sequences are of unequal length or in the unusual event that neither of the factors has symmetry (even or odd) (Bracewell, 1986b). If one of the sequences is short, having less than about 32 elements, depending on the machine, then slow convolution by direct evaluation of the convolution sum may be faster, and a shorter program will suffice. When the Fourier transform is used, the multiplications are complex but half of them may be avoided because of Hermitian symmetry. Software packages such as CNVLV (Press et al., 1986) are available that handle these technicalities by calling two unilateral transforms, each faster than the FFT, or two equivalent subprograms; one fast Hartley transform, which is bilateral and, conveniently for the computer, real valued, now replaces such packages. Fast convolution using prime-factor algorithms is also available if general-purpose use is not a requisite.

As an example, suppose that $\{1 \quad 2 \quad 1\}$ is to be convolved with $\{1 \quad 4 \quad 6 \quad 4 \quad 1\}$, a simple situation where we know that the answer is the binomial sequence

$$\{1 \quad 6 \quad 15 \quad 20 \quad 15 \quad 6 \quad 1\}.$$

If we select $N = 8$ for the discrete transform calculation, the given factors become in effect

$$f_1(\tau) = \{1 \quad 2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\}$$

and

$$f_2(\tau) = \{1 \quad 4 \quad 6 \quad 4 \quad 1 \quad 0 \quad 0 \quad 0\},$$

respectively, where the boldface emphasizes the zeroth elements $f_1(0)$ and $f_2(0)$. The sequence $\{1 \quad 2 \quad 1\}$ is commonly used to apply some smoothing to a data sequence, but since the center of symmetry at the element 2 is offset from the origin at $\tau = 0$, a shift will be introduced in addition to the smoothing. Therefore it makes sense to permute the sequence cyclically and use

$$f_1(\tau) = \{2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1\}.$$

To compute the convolution

$$\begin{aligned} f_3 &= f_1 * f_2 \\ &= \{2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1\} \\ &\qquad * \{1 \quad 4 \quad 6 \quad 4 \quad 1 \quad 0 \quad 0 \quad 0\}, \end{aligned}$$

we take the two 8-element Hartley transforms to get the values H_1 and H_2 tabulated in Table 3. Multiply the corresponding values as shown under $H_1 H_2$ and take the Hartley transform again. The result is as expected; notice that the peak value 20 occurs where the peak value 6 of $f_2(\tau)$ occurs; this is a result of the precaution of centering $\{1 \quad 2 \quad 1\}$ appropriately. The noninteger results are a consequence of rounding to three decimals for demonstration purposes, and these errors will be

Tab. 3 Performing convolution by multiplying Hartley transforms

f_1	f_2	H_1	H_2	$H_1 H_2$	f_3
2	1	0.5	2	1	6
1	4	0.427	1.457	0.622	15.008
0	*6*	0.25	−0.5	−0.125	*20*
0	4	0.073	−0.043	−0.003	15.008
0	1	0	0	0	6
0	0	0.073	0.043	0.003	0.992
0	0	0.25	−0.5	−0.125	0
1	0	0.427	−1.457	−0.622	0.992

present, though smaller, if more decimals are retained.

20 Summary

A great analytic tradition of mathematical transform theory has gained far-ranging everyday importance by virtue of new numerical possibilities opened up by automatic computing machines.

Glossary

Alias: A sinusoid of low frequency spuriously introduced by insufficient sampling in the presence of a sinusoidal component of semiperiod shorter than the sampling interval.

Convolution of Two Functions: A third function composed by spreading each element of one given function out into the form of the second function and superimposing the spread components.

Discrete Transform: One suited to functions, such as those constituted by equispaced data samples, where the function values occur at discrete intervals and are usually finite in number.

Fast Fourier Transform (FFT): An algorithm for computing the discrete Fourier transform in less time than would be required to evaluate the sum of the products indicated in the defining formula.

Frequency, Negative: A convenient fiction arising from the representation of real sinusoids by complex quantities. The representation of the real function $\cos 2\pi ft$ in the form $\frac{1}{2}\exp[i2\pi ft] + \frac{1}{2}\exp[i2\pi(-f)t]$ involves clockwise rotation at frequency f and counter-clockwise rotation at frequency $-f$.

Frequency, Spatial: The reciprocal of the period of a periodic function of space. Values of spatial frequency are expressed in cycles per meter, or in cycles per radian, according as the spatial variable is distance or angle.

Frequency, Temporal: The reciprocal of the period of a periodic function of time. Values are expressed in cycles per second, or hertz.

Heaviside Unit Step Function: A function $H(x)$ that is equal to zero to the left of the origin and equal to unity to the right. The value $H(0)$ at the origin has no effect on the value of integrals but may conventionally be taken as 0.5.

Inverse Transformation: An operation that, when applied to the transform of a function, effectively recovers the function.

Linear Transformation: A transformation with the property that the transform of the sum of any two functions is the sum of the separate transforms.

Tomography: Originally a photographic technique for obtaining an x-ray image of a slice of tissue within the body; now applied in many fields to a technique of combining projections in many orientations to reconstruct an image.

Transform: A mathematical function, each value of which is derived from a set of values of a given function by an explicit operation.

List of Works Cited

Bracewell, R. N. (1956), *Aust. J. Phys.* **9**, 198–217.

Bracewell, R. N. (1984), *Proc. IEEE* **72**, 1010–1018.

Bracewell, R. N. (1986a), *The Fourier Transform and Its Applications*, 2nd ed. rev., New York: McGraw-Hill.

Bracewell, R. N. (1986b), *The Hartley Transform*, New York: Oxford University Press.
Bracewell, R. N. (1987), *Electron. Lett.* **23**, 1148–1149.
Bracewell, R. N. (1989), *J. Atmos. Terrest. Phys.* **51**, 791–795.
Bracewell, R. N. (1990), *IEEE Trans. ASSP* **38**, 2174–2176.
Bracewell, R. N., Riddle, A. C. (1967), *Astrophys. J.* **50**, 427–434.
Bracewell, R. N., Villasenor, J. D. (1990), *J. Opt. Soc. Am.* **7**, 21–26.
Bracewell, R. N., Buneman, O., Hao, H. Villasenor, J. (1986), *Proc. IEEE* **74**, 1283–1284.
Brooks, R. A., Di Chiro, G. (1976), *Phys. Med. Biol.* **21**, 689–732.
Buneman, O. (1989), *IEEE Trans. ASSP* **37**, 577–580. (Copyright held by the Board of Trustees of Leland Stanford Junior University.)
Campbell, G. A., Foster, R. M. (1948), *Fourier Integrals for Practical Applications*, New York: Van Nostrand.
Cooley, J. W., Tukey, J. W. (1965), *Math. Comput.* **19**, 297–301.
Cormack, A. M. (1963), *J. Appl. Phys.* **34**, 2722–2727.
Cormack, A. M. (1964), *J. Appl. Phys.* **35**, 2908–2913.
Cormack, A. M. (1980), *Med. Phys.* **7**, 277–282.
Deakin, M. A. B. (1985), *Math. Ed.* **1**, 24–28.
Deans, S. R. (1983), *The Radon Transform and Some of Its Applications*, New York: Wiley.
Elliott, D. F., Rao, K. R. (1982), *Fast Transforms: Algorithms, Analyses, Applications*, New York: Academic.
Erdélyi, A., Oberhettinger, F., Magnus, W., Tricomi, F. G. (1954), *Tables of Integral Transforms*, New York: McGraw-Hill.
Gauss, C. F. (1876), Werke, vol. 3, Göttingen: Königliche Gesellschaft der Wissenschaffen.
Grattan-Guinness, I. (1972), *Joseph Fourier, 1768–1830*, Cambridge, MA: The MIT Press.
Hao, H., Bracewell, R. N. (1987), *Proc. IEEE* **75**, 264–266.
Hartley, R. V. L. (1942), *Proc. IRE* **30**, 144–150.
Heaviside, O. (1970), Electromagnetic Theory, vols. 1–3, New York: Chelsea.
Heideman, M. T., Johnson, D. H., Burrus, C. S. (1985), *Arch. Hist. Exact Sci.* **34**, 265–277.
Hsu, C.-Y., Wu, J.-L. (1987), *Electron. Lett.* **23**, 466–468.
IEEE (1967), Special issue on Fast Fourier Transform, *IEEE Trans. Audio Electroacoustics* **AU-2**, 43–98.
IEEE (1979), Digital Signal Processing Committee, IEEE ASSP Soc. (Eds.), *Programs of Digital Signal Processing*, New York: IEEE Press.
Körner, T. W. (1988), *Fourier Analysis*, Cambridge U.K.: Cambridge Univ. Press.
Lighthill, M. J. (1958), *An Introduction to Fourier Analysis and Generalized Functions*, Cambridge U.K.: Cambridge Univ. Press.
McCollum, P. A., Brown, B. F. (1965), *Laplace Transforms Tables and Theorems*, New York: Holt, Rinehart and Winston.
Nahin, P. J. (1987), *Oliver Heaviside, Sage in Solitude*, New York: IEEE Press.
Neugebauer, O. (1983), *Astronomy and History, Selected Essays*, New York: Springer.
Nussbaumer, H. J. (1982), *Fast Fourier Transform and Convolution Algorithms*, New York: Springer.
Press, W. H., Flannery, B. P. Teukolsky, S. A., Vetterling, W. T. (1986), *Numerical Recipes*, Cambridge, U.K.: Cambridge Univ. Press.
Rabiner, L. R., Gold, B. A. (1975), *Theory and Application of Digital Signal Processing*, Englewood Cliffs, NJ: Prentice Hall.
Radon, J. (1917), *Ber. Sächs. Akad. Wiss. Leipzig, Math.-Phys. Kl.* **69**, 262–277.
Rosenfeld, A., Kac, A. C. (1982), Digital Picture Processing, Vol. 1, New York: Academic.
Villasenor, J. D., Bracewell, R. N. (1987), *Nature* **330**, 735–737.
Villasenor, J. D., Bracewell, R. N. (1988), *Nature* **335**, 617–619.
Villasenor, J. D., Bracewell, R. N. (1989), *Electron. Lett.*, **25**, 1110–1111.
Walsh, J. L. (1923), *Am. J. Math.* **45**, 5–24.

Further Reading

Texts treating the various transforms and computational methods are identifiable from their titles in the list of works cited. An indispensable source for locating recent material on any of the special branches mentioned in this article is *Mathematical Abstracts*.

Fractal Geometry

Paul Meakin
Department of Physics, University of Oslo, Oslo, Norway

Introduction ***110***
1 Self-Similar Fractals ***110***
1.1 The Cantor Set, A Simple Example *110*
1.2 Statistically Self-Similar Fractals *111*
1.3 The Characterization of Self-Similar Fractals *112*
1.4 Simple Rules for Self-Similar Fractals *114*
2 Self-Affine Fractals ***115***
2.1 The Brownian Process, A Simple Example *115*
2.2 The Characterization of Self-Affine Fractals *115*
3 Fractal Surfaces and Interfaces ***116***
3.1 Some Applications *116*
3.2 The Growth of Rough Surfaces *118*
3.2.1 Self-Similar Rough Surfaces *118*
3.2.2 Self-Affine Rough Surfaces *121*
4 Practical Considerations ***123***
Glossary ***123***
List of Works Cited ***124***
Further Reading ***125***

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Introduction

A quantitative description of the structure of physical objects plays an important role in our understanding of a wide range of phenomena. In many areas such as spectroscopy, solid-state physics, and engineering physics, the symmetry properties associated with this geometric description lead to important insights that would be difficult to obtain in other ways. Until recently structures have been described in terms of Euclidean geometry (straight lines, planar surfaces, spherical particles, etc.) and the associated symmetries of invariance to rotation, reflection, translation, and inversion. However, many systems of practical importance (colloids, rough surfaces and interfaces, polymer molecules, etc.) cannot be described satisfactorily in such terms. In the decade or so following the development and popularization of fractal geometry by Mandelbrot (1977, 1982) it has been shown that fractal geometry and the associated symmetry of scale invariance can be used to describe a wide variety of disorderly structures.

1 Self-Similar Fractals

1.1 The Cantor Set, A Simple Example

The first and perhaps the most simple example of a fractal (the Cantor set) is illustrated in Fig. 1. The Cantor set can be constructed by first taking a line and removing the middle third. In the next stage of this process the middle third from each of the remaining line segments is removed, etc. After n generations the number of line segments has grown to 2^n but their total length has decreased to $\left(\frac{2}{3}\right)^n$. In the limit $n \to \infty$ a self-similar fractal has been constructed. If this fractal is dilated by a factor of 3, it can be covered by two replicas of itself. Such self-similar fractals can be characterized in terms of their fractal dimensionality D given by $D = \log n / \log \lambda$, where n is the number of replicas required to cover the fractal after dilation by a factor of λ. For the Cantor set illustrated in Fig. 1 the fractal dimensionality is log 2/log 3, or about 0.6309. This is intuitively reasonable; the Cantor set is clearly more than a point ($D = 0$), since it contains an infinite number of them, but less than a line ($D = 1$), since its total length is zero. In many applications the fractal dimension can be thought of as the exponent that relates mass M to length L,

$$M \sim L^D. \tag{1}$$

Here, L is a characteristic length, such as the radius of gyration R_g or maximum diameter, that describes the overall spatial extent. Equation (1) is also appropriate for Euclidean shapes where D is now the ordinary, Euclidean, dimensionality d.

After they were first introduced, fractals such as the Cantor set were considered to be very unusual objects with no possible applications in the physical sciences. In

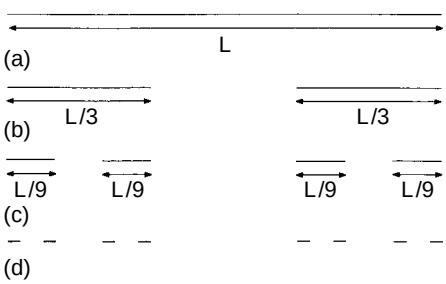


Fig. 1 Three stages in the construction of a Cantor set with a fractal dimensionality of log 2/log 3. The bottom line shows the third-generation prefractal

some areas fractals are still referred to by terms such as "strange sets"; this no longer seems to be appropriate. There is no precise definition of a fractal, but in general terms a fractal is an object that has the same degree of complexity on different length scales.

In real systems the geometric scaling relationships that characterize fractal objects do not extend over an infinite range of length scales. There are in general both lower (ε) and upper (L) cutoff lengths that limit the range of fractal scaling. For example, in the case of a flexible polymer molecule ε might correspond to the monomer size and L to the radius of gyration. If the ratio L/ε is large, then fractal geometry can be a very important asset in our attempts to understand complex, disorderly systems. If L/ε is small (say less than one order of magnitude), then fractal geometry is not likely to be of much practical importance. However, it may still be of considerable conceptual value if the structure was assembled by a mechanism that would lead to a fractal structure if it were not perturbed by other processes.

1.2 Statistically Self-Similar Fractals

Highly organized regular fractals such as the Cantor set (Fig. 1) that can be mapped exactly onto themselves after a change of length scales do not provide realistic models for describing most natural structures. Such natural fractals have a more complex disorderly structure that is self-similar only in a statistical sense. Statistically self-similar fractals can be described in terms of the scaling relationships such as Eq. (1) that describe regular, hierarchical fractals, but these equations must now be interpreted statistically (for example L might be the average radius of gyration for a large ensemble of structures of mass M).

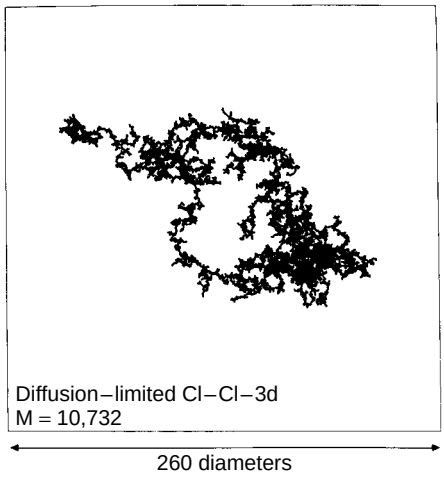


Fig. 2 A cluster of 10,732 particles generated using a three-dimensional off-lattice diffusion-limited cluster-cluster aggregation model. In this model the cluster size distribution is allowed to evolve in a natural way

Figure 2 shows an example of a statistically self-similar fractal structure. It is a projection, onto a plane, of a cluster of spherical particles generated by a three-dimensional model for the diffusion-limited cluster-cluster aggregation (colloidal flocculation: Meakin, 1988; Jullien and Botet, 1986). The fractal dimensionality of this structure is about 1.8. Since $D < 2$, the projection also has a fractal dimensionality of 1.8 (see Sec. 1.4).

The use of correlation functions has, for a long time, been a valuable approach toward the quantitative characterization of disorderly systems. For example, density correlation functions such as $C^n(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ defined as

$$C^n(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \langle \rho(\mathbf{r}_0)\rho(\mathbf{r}_0 + \mathbf{r}_1) \cdots \cdots \rho(\mathbf{r}_0 + \mathbf{r}_n) \rangle \qquad (2)$$

can be used to describe both fractal and nonfractal structure. Here $\rho(\mathbf{r})$ is the density at position $\mathbf{r}$ and the averaging is over all possible origins ($\mathbf{r}_0$). For self-similar fractals these correlation functions have a homogeneous power-law form,

$$C^n(\lambda\mathbf{r}_1, \lambda\mathbf{r}_2, \ldots, \lambda\mathbf{r}_n) = \lambda^{-n\alpha} C^n(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n). \quad (3)$$

The exponent α (called the co-dimensionality) in Eq. (3) is $d - D$, where d is the Euclidean dimensionality of the embedding space.

By far the most frequently used correlation function is the two-point density-density correlation function $C(r)$ given by

$$C(r) = \langle\langle \rho(\mathbf{r}_0)\rho(\mathbf{r}_0 + \mathbf{r})\rangle\rangle_{|\mathbf{r}|=r}. \quad (4)$$

Here $\langle\langle\rangle\rangle$ implies averaging over all origins ($\mathbf{r}_0$) and orientations. In addition $C(r)$ may be averaged over an ensemble of samples. For a simple self-similar fractal $C(r)$ has the powerlaw form

$$C(r) \sim r^{-\alpha}, \quad (5)$$

and the fractal dimensionality D_α is equal to $d - \alpha$.

1.3 The Characterization of Self-Similar Fractals

Correlation functions such as those described above can be used to measure the fractal dimensionality. In practice only the two-point density-density correlation function has been used extensively for this purpose. Figure 3(a) shows the two-point density-density correlation functions for clusters of different sizes generated using the three-dimensional diffusion-limited cluster-cluster aggregation model illustrated in Fig. 2. These correlation functions have the form

$$C(r) = r^{-\alpha} f\left(\frac{r}{L}\right), \quad (6)$$

where L is the cutoff length. The cutoff function $f(x)$ has the form $f(x) = \text{const}$ for $x \ll 1$ and $f(x)$ decays faster than any power of x with increasing x for $x \gg 1$. Figure 3(b) shows that the density-density correlation function can be represented by the scaling form

$$C(r) = N^{(D-d)/D} g\left(\frac{r}{N^{1/D}}\right), \quad (7)$$

where N is the number of particles in the cluster. Since $L \sim N^{1/D}$ and $\alpha = d - D$, the scaling forms in Eqs. (6) and (7) are equivalent. [The functions $f(x)$ and $g(x)$ are related by $g(x) \sim x^{-\alpha} f(x)$]. The results shown in Fig. 3 demonstrate that the internal structure of the clusters and their global mass-length scaling can be described in terms of the same fractal dimensionality ($D \simeq 1.8$).

Most approaches to the characterization of self-similar fractals are based on Eq. (1). For example, if we are concerned with structures formed by growth processes or systems in which a large number of objects of different sizes are present, then the fractal dimensionality can be obtained from the dependence of the radius of gyration on the mass. For an ensemble of statistically self-similar fractals we have

$$\langle R_g \rangle \sim M^\beta, \quad (8)$$

where $\langle R_g \rangle$ is the mean radius of gyration for structures of mass M. The corresponding fractal dimensionality D_β is then given by $D_\beta = 1/\beta$. In practice D_β is obtained by fitting the dependence of log R_g on log M by a straight line and taking the slope of the straight line as the exponent β. If data are available from many realizations over a broad range of length scales, the exponent

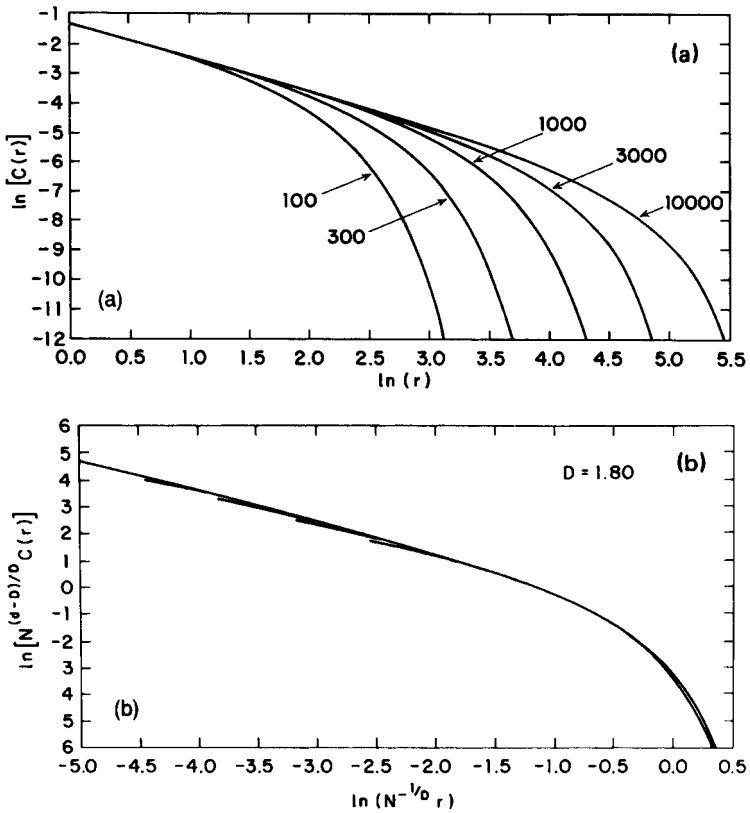


Fig. 3 Scaling of the two-point density-density correlation function for clusters generated using an off-lattice diffusion-limited cluster-cluster aggregation model illustrated in Fig. 2. (a) The correlation functions obtained from 10,000 100-particle clusters, 1000 300-particle clusters, 100 1000-particle clusters, 39 3000-particle clusters, and 13 10,000-particle clusters. (b) How these correlation functions can be scaled onto a common curve using the scaling form given in Eq. (7)

β can be measured over a number of mass intervals to assess how accurately Eq. (8) represents the dependence of R_g on M.

Another popular approach to the measurement of the fractal dimension of self-similar fractals is to cover the fractal by a series of grids with elements having sides of length ε. The number of elements completely or partially occupied by the fractal is then given by

$$N(\varepsilon) \sim \varepsilon^{-D_\varepsilon}, \qquad (9)$$

so that the fractal dimensionality (D_ε) can be obtained by plotting log $N(\varepsilon)$ against log ε. In practice this method appears to be quite sensitive to corrections associated with a limited range of accessible length scales, and it is difficult to obtain reliable results.

In many cases structures grow from a unique "seed" or nucleation site. In this case the fractal dimensionality can be obtained by measuring the mass $M(l)$ contained within a distance l measured

from this unique point. For a self-similar fractal $M(l)$ is given by

$$M(l) \sim l^{D_\gamma}, \qquad (10)$$

so that D_γ can be obtained by fitting a straight line to the dependence of $\log M(l)$ on $\log l$ over a suitable range of length scales.

In principle all of the methods described above (and many other methods) should lead to the same value for the fractal dimensionality ($D_\alpha = D_\beta = D_\gamma = D_\varepsilon = D$, where D is the "all purpose" fractal dimensionality) for simple self-similar fractals. In practice it is a good idea to use several approaches to measure D. This provides an assessment of the uncertainty in D and some indication of whether or not the structure is indeed a self-similar fractal. These methods can be applied equally well to fractal structures generated by physical phenomena or computer simulations.

1.4 Simple Rules for Self-Similar Fractals

The ability to describe complex, disorderly structures in quantitative terms (via fractal geometry) has stimulated scientific interest in problems that only a decade or so ago were thought to lie outside of the realm of quantitative scientific investigation. For a variety of model systems we now have quantitative (frequently exact but rarely rigorous) results and at least the beginnings of a sound theoretical understanding. In attempting to apply the concepts of fractal geometry to self-similar systems the following simple rules or ideas have been found to be useful (Vicsek, 1989; Meakin, 1990).

1. Two fractals with dimensionalities D_1 and D_2 can be placed together in the same region of a d-dimensional embedding space or lattice without contacting each other if $d > D_1 + D_2$. If the two fractals are following a relative trajectory with a fractal dimensionality of D_t, then they will not contact each other (except by accident) if $d > D_1 + D_2 + D_t$. An important implication of this rule is that fractals with $D < 2$ will be asymptotically transparent in three-dimensional space since they will not be contacted by "point" objects ($D = 0$) such as photons or electrons following linear ($D = 1$) trajectories. For such fractals one part is (in general) not hidden by another, and the fractal dimensionality of a projection onto a plane is the same as that of the fractal itself. If $D > 2$, then the structure is asymptotically opaque and the fractal dimensionality cannot be determined by analyzing a projection onto a plane. It follows from this that the projection of a D-dimensional fractal onto a d-dimensional space will have a fractal dimension of D if $D < d$. In this event the area (measure) of the projection will be proportional to the mass (measure of the fractal in the d-dimensional space).
2. A d_1-dimensional cross section of a D-dimensional fractal in a d_2-dimensional space will have a fractal dimensionality of $D + d_1 - d_2$.
3. The (set theoretical) intersection of two fractals with dimensionalities D_1 and D_2 in a d-dimensional space is given by $D_1 + D_2 - d$. This rule can be applied repeatedly to obtain the dimensionality of the intersection of three ($D_1 + D_2 + D_3 - 2d$) or more fractals.
4. The union of two fractals with dimensionalities D_1 and D_2 has a fractal dimensionality of $\max(D_1, D_2)$.

5. The product of two fractals with dimensionalities D_1 and D_2 has a dimensionality of $D_1 + D_2$. For example, the region swept out by a fractal of dimensionality D following a trajectory of dimensionality D_t is $D + D_t$ (if $D + D_t < d$).
6. Many random fractals can be described in terms of a power-law distribution of unoccupied "holes,"

$$N_s \sim s^{-\tau}, \tag{11}$$

where N_s is the number of holes of size s (s would be the number of sites contained in the hole for a lattice model). For such fractals the size distribution exponent τ is given by $\tau = (d + D)/d$.

2 Self-Affine Fractals

2.1 The Brownian Process, A Simple Example

Fractals that have different scaling structures in different directions are said to be self-affine. Perhaps the most simple and most important example of a self-affine fractal is the Brownian process $B(t)$ that describes the distance moved by a Brownian particle in time t. It is well known that (on average) the distance moved by a Brownian particle in time t is proportional to $t^{1/2}$ so that the Brownian process can be rescaled by simultaneously changing the time scale by a factor of b and the distance scale by a factor of $b^{1/2}$. More formally this symmetry property of the Brownian process can be written as

$$B(t) \equiv b^{-1/2} B(bt). \tag{12}$$

In this equation the symbol "$\equiv$" should be interpreted as meaning "statistically equivalent to."

Figure 4 shows different "lengths" of the same discretized Brownian process in which the distance is increased randomly by $+1$ or -1 each time the time is incremented by 1. In each part of the figure the horizontal (time) scale is proportional to the total time T and the vertical (distance) scale is proportional to $T^{1/2}$. The observation that the four rescaled curves in Fig. 4 look "similar" illustrates the self-affine scaling of the Brownian process.

The Brownian process can be generalized to give the "fractal" Brownian process $B_H(t)$, for which the self-affine scaling properties can be represented as

$$B_H(t) \equiv b^{-H} B(bt), \tag{13}$$

where the exponent H is referred to as the Hurst exponent (roughness exponent or wandering exponent). Values of H larger than $\frac{1}{2}$ correspond to persistent processes and $H < \frac{1}{2}$ implies antipersistent fluctuations. This is illustrated in Fig. 5 where fractal Brownian curves with Hurst exponents of 0.1, 0.5, and 0.9 are shown.

2.2 The Characterization of Self-Affine Fractals

In many cases (such as the Brownian process described above) self-affine fractals can be represented as single-valued functions $z(\mathbf{x})$ of the coordinates $x_1, x_2, \ldots, x_n$. For this class of self-affine fractals the scaling properties can be described in terms of the correlation functions $C_q(x)$ defined as

$$(C_q(x))^q = \langle |z(\mathbf{x}_0) - z(\mathbf{x}_0 + \mathbf{x})|^q \rangle_{|\mathbf{x}|=x}. \tag{14}$$

In this case it has been assumed that all of the coordinates $(x_1, \ldots, x_n)$ are

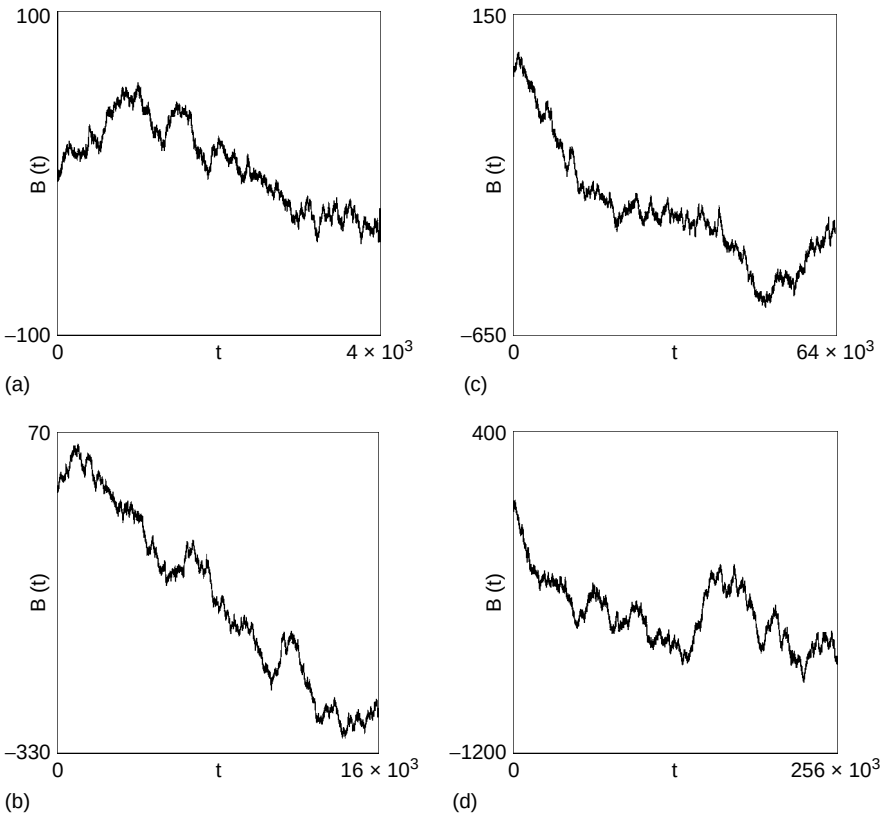


Fig. 4 Four sections of the same (discretized) Brownian process starting at the point $[t = 0, B(t) = 0]$. In each successive stage the time scale is increased by a factor of 4 and the vertical scale by 2 ($4^{1/2}$)

equivalent. In general the self-affine fractal $z(\mathbf{x})$ can be characterized by taking cross sections through the function $z(\mathbf{x})$ in the direction $x_m, y(x_m)$, and measuring the correlation function

$$(C_q(x_m))^q = \langle |y(x_m^0) - y(x_m^0 - x_m)|^q \rangle. \tag{15}$$

For self-affine fractals the correlation functions $C_q(x)$ or $C_q(x_m)$ have the form

$$C_q(x) \sim x^H \tag{16a}$$

or

$$C_q(x_m) \sim x_m^{H_m}. \tag{16b}$$

3 Fractal Surfaces and Interfaces

3.1 Some Applications

One of the most important applications of fractal geometry is the quantitative description of irregular surfaces and the development of a better understanding of their behavior. During the past few decades the technological importance of rough surfaces has motivated the development of a large number of procedures for characterizing their structure. Many of

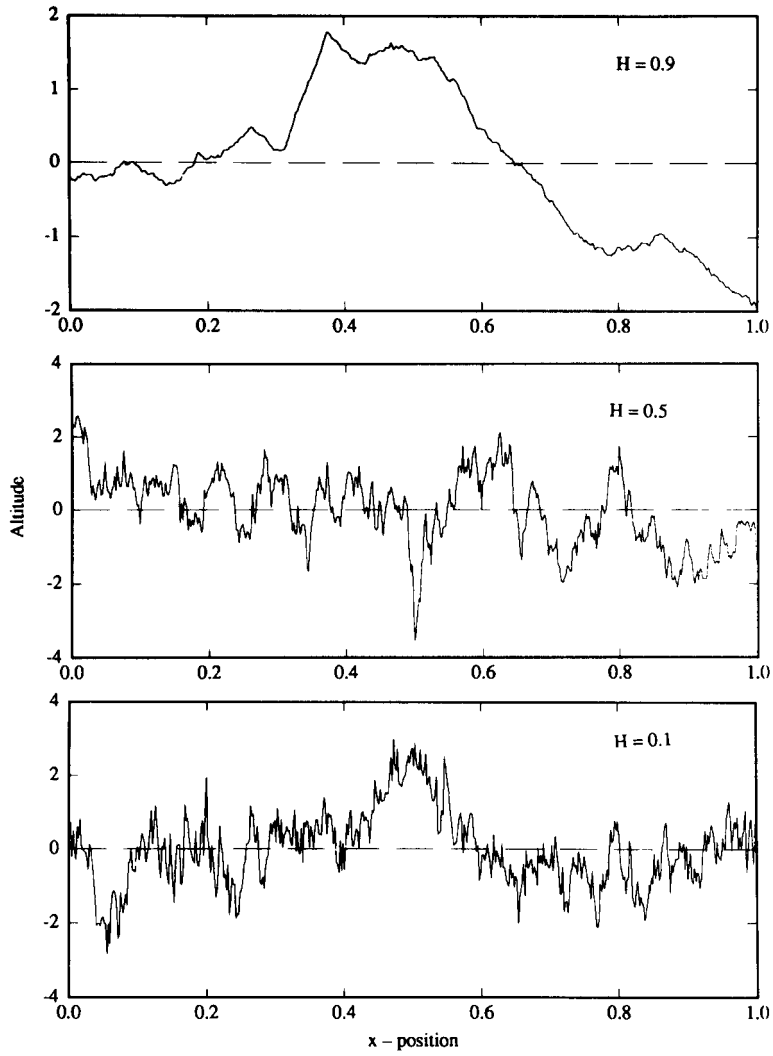


Fig. 5 Fractal Brownian curves generated using three different values for the Hurst exponent [$H = 0.9$ (top), $H = 0.5$, and $H = 0.1$]. This figure is taken from Feder (1988) and was provided by J. Feder

these approaches involve a large number of parameters that are of little fundamental importance. In a recent review (Nowicki, 1985) 32 parameters and/or functions that have been used to characterize rough surfaces are described. More recently it has been shown that a wide variety of rough surfaces generated by processes such as fracture (Mandelbrot et al., 1984; Charmet et al., 1989; Herrmann and Roux, 1990), corrosion (Oppenheim et al., 1991; Holland-Moritz et al., 1991), deposition (Meakin, 1987; Family and Vicsek, 1991; Krug and Spohn, 1991),

or fluid-fluid displacement (Feder, 1988) can be described in terms of fractal geometry. Both self-similar and self-affine fractal surfaces are common but most surfaces appear to be self-affine.

3.2 The Growth of Rough Surfaces

In many processes of practical importance rough surfaces are generated from a more or less smooth surface. Under these circumstances the surface roughness can often be described in terms of the correlation lengths $\xi_\perp$ and $\xi_\parallel$ and the manner in which they grow. The correlation length $\xi_\perp$ describes the amplitude of the surface roughness in a direction perpendicular to the general direction of the surface and $\xi_\parallel$ is the distance over which fluctuations in the surface height persist in a direction parallel to the coarse-grained surface. The length $\xi_\perp$ can be defined as

$$\xi_\perp^2 = \langle h^2 \rangle - \langle h \rangle^2, \tag{17}$$

where $h(\mathbf{x})$ is the height of the surface above (or below) position $\mathbf{x}$ on the initially smooth surface.

For many simple processes the correlation length $\xi_\perp$ grows algebraically with increasing time

$$\xi_\perp \sim t^\omega. \tag{18}$$

For self-affine surfaces the correlation lengths $\xi_\perp$ and $\xi_\parallel$ are related by

$$\xi_\perp \sim \xi_\parallel^H, \tag{19}$$

where H is the Hurst exponent. In some cases surface properties in the x and y directions parallel to the surface may be quite distinct so that Eq. (17) may be replaced by

$$\xi_\perp \sim \xi_x^{H_x} \sim \xi_y^{H_y}. \tag{20}$$

In some simple cases the amplitude of the surface roughness ($\xi_\perp$) may grow indefinitely according to Eq. (18); but in many cases $\xi_\perp$ is limited by other physical processes, and this limits the range of length scales over which fractal scaling can be observed.

Simple model systems that are quite well understood are used to illustrate how rough surfaces can be characterized using fractal geometry and scaling ideas in Secs. 3.3 and 3.4.

3.2.1 Self-Similar Rough Surfaces

The invasion percolation model (Lenormand and Bories, 1980; Wilkinson and Willemsen, 1983; Feder, 1988) provides a simple description of the slow displacement of a wetting fluid by a nonwetting fluid in a porous medium. In the site-invasion percolation model the sites on a lattice are assigned random "threshold" values at the start of a simulation. At each step in the simulation the unoccupied perimeter site with the lowest threshold value is filled to represent the fluid-fluid displacement process (unoccupied perimeter sites are unoccupied sites with one or more occupied nearest neighbors). In the two-dimensional version of this model the displacement pattern is self-similar with a fractal dimensionality of $\frac{91}{48}$ (about 1.89), or about 1.82 if growth is not allowed to take place in regions that have been surrounded by the growing cluster. The invasion front (outer perimeter) has a fractal dimensionality of $\frac{4}{3}$ (Grossman and Aharony, 1986; Saleur and Duplantier, 1987; Coniglio et al., 1987) for both versions of the model (with and without "trapping").

If the fluid-fluid displacement processes take place in a vertical or inclined cell and the two fluids have different densities, the process may be either stabilized (Birovljev

et al., 1991) or destabilized by gravity. In this situation the invasion percolation process can be simulated using thresholds given by

$$t_i = x_i + gh_i, \tag{21}$$

where t_i is the threshold associated with the ith site, x_i is a random number (in the most simple model, used here, the random numbers are uniformly distributed over the range $0 < x_i < 1$), and h_i is the height of the ith site. Figure 6 shows the displacement fronts generated during simulations carried out using the gravity-stabilized invasion percolation model. In this model the invasion front evolves toward an asymptotic regime in which the statistical properties become stationary. The fronts shown in Fig. 6 were recorded in this asymptotic regime.

Self-similar fractal surfaces can be described in terms of the correlation functions defined in Eqs. (2) and (4). For structures generated by physical processes, Eqs. (3) and (5) are accurate over a limited range of length scales ($\varepsilon \ll r \ll \xi$). For the invasion percolation simulations described in this section, the inner cutoff length ε is one lattice unit and the outer cutoff length ξ is related to the stabilizing gradient g by

$$\xi \sim g^{-\gamma}, \tag{22}$$

where the exponent γ is given by (Wilkinson, 1984)

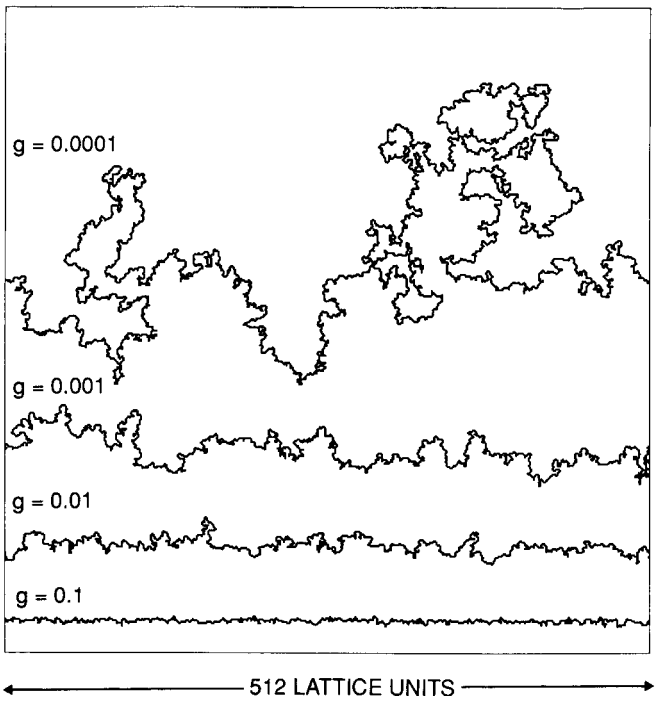


Fig. 6 Invasion fronts (unoccupied external perimeter sites) obtained from two-dimensional gradient-stabilized invasion percolation simulations. Fronts generated using four values of the stabilizing gradient g (10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}) are shown

$$\gamma = \frac{\nu}{(\nu + 1)}, \tag{23}$$

where ν is the correlation length exponent for ordinary percolation. The exponent ν has a value of exactly $\frac{4}{3}$ for two-dimensional percolation (Stauffer, 1985).

Figure 7(a) shows the density-density correlation functions $C(r)$ for the invasion fronts obtained from two-dimensional invasion percolation simulations with five values for the stabilizing gradient (g): $g = 0.1,\ 0.01,\ 0.001,\ 0.0001,\ 0.00001$. In these plots a crossover from a slope of $-\frac{2}{3}$ on short length scales (corresponding to a fractal dimensionality of $\frac{4}{3}$) to a slope of -1 on longer length scales ($D = 1$) can be seen. The results shown in Figs. 6 and 7 were obtained by starting with a flat surface and allowing the simulations to proceed until the vertical correlation length $\xi_\perp$ has grown to a stationary value

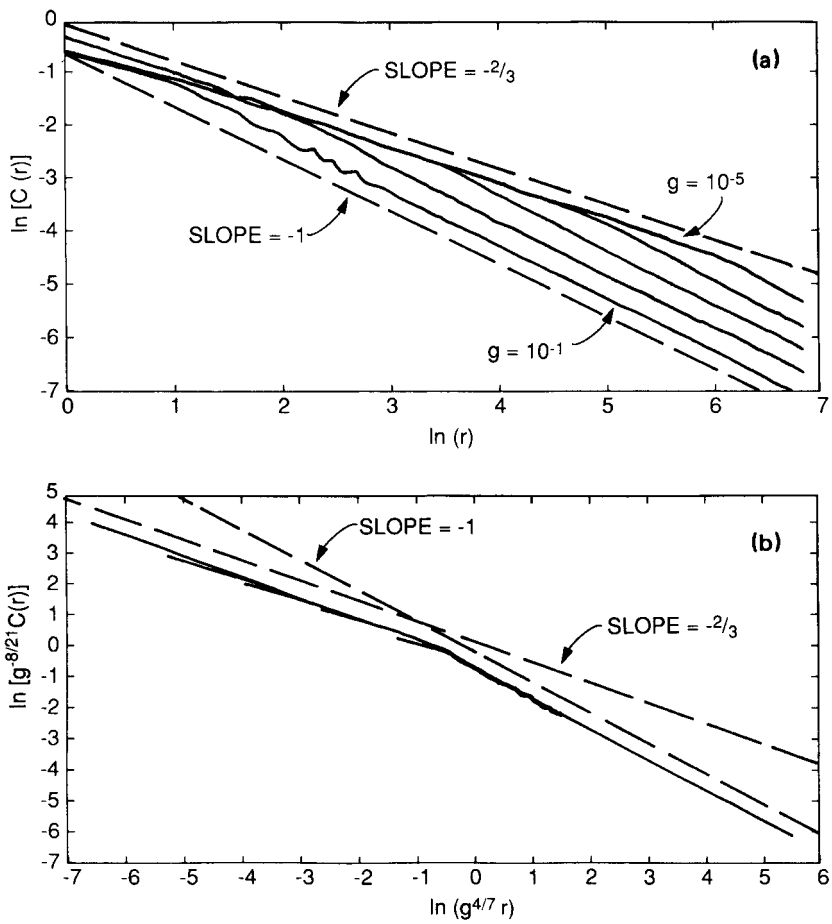


Fig. 7 Two-point density-density correlation functions for invasion fronts similar to those shown in Fig. 6. Each correlation function was obtained from 20 simulations carried out using a strip width L of 2048 lattice units. (b) How these correlation functions can be scaled using the scaling form given in Eq. (24)

$\xi_\perp(\infty)$ given by Eq. (22). In this stationary regime the growth of the invasion front exhibits large fluctuations, but the mean statistical properties such as $\xi_\perp$ and $C(r)$ do not change.

Figure 7(b) shows how the correlation function $C(r)$ shown in Fig. 7(a) can be collapsed onto a single curve. This figure illustrates that $C(r)$ (in the stationary regime) can be represented by the scaling form

$$C(r) = g^{(8/21)} f(g^{4/7} r) \tag{24}$$

or

$$C(r) = g^{\alpha\gamma} f(g^{\gamma} r), \tag{25}$$

where the scaling function $f(x)$ has the form $f(x) \sim x^{-2/3}$ for $x \ll 1$ and $f(x) \sim x^{-1}$ for $x \gg 1$. This means that the surface appears to be self-similar $\left(D = \frac{4}{3}\right)$ for $x \ll 1$ and flat ($D = 1$) for $x \gg 1$.

3.2.2 Self-Affine Rough Surfaces

The Brownian process described above provides a valuable paradigm for the geometry of rough surfaces. The correlation functions $C_q(x)$ [Eq. (14)] can be used to measure the Hurst exponent (by least-squares fitting of straight lines to the dependence of log $C_q(x)$ on log x. For $q = 1$ and 2 the value for the Hurst exponent is well within the 1% of the expected value $\left(H = \frac{1}{2}\right)$.

In many real systems the correlation length $\xi_\perp$ may be finite because $\xi_\perp$ has not had enough time to grow or because it is limited by physical processes or finite size effects. Figure 8 shows "surfaces" generated using a simple modification of the discrete Brownian process model in which the probability of moving toward the origin is $0.5 + k|x|$ and the probability of moving away from the origin is $0.5 - k|x|$, where x is the displacement from the origin. This may be regarded as a model

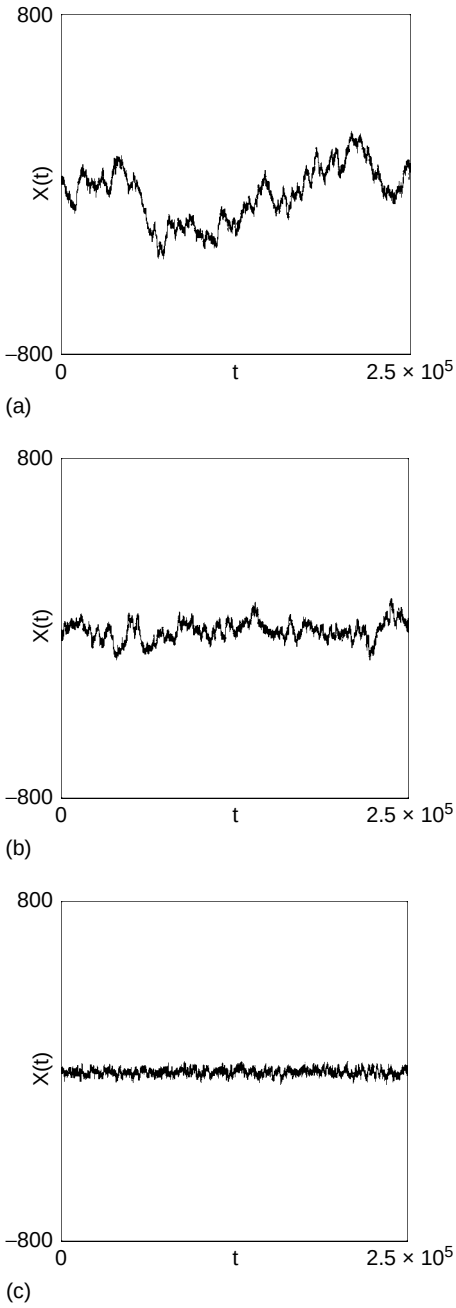


Fig. 8 Displacement curves $x(t)$ obtained from a simple model for Brownian motion in a harmonic potential. (a)–(c) Results for $k = 10^{-2}$, 10^{-3}, and 10^{-4}, respectively

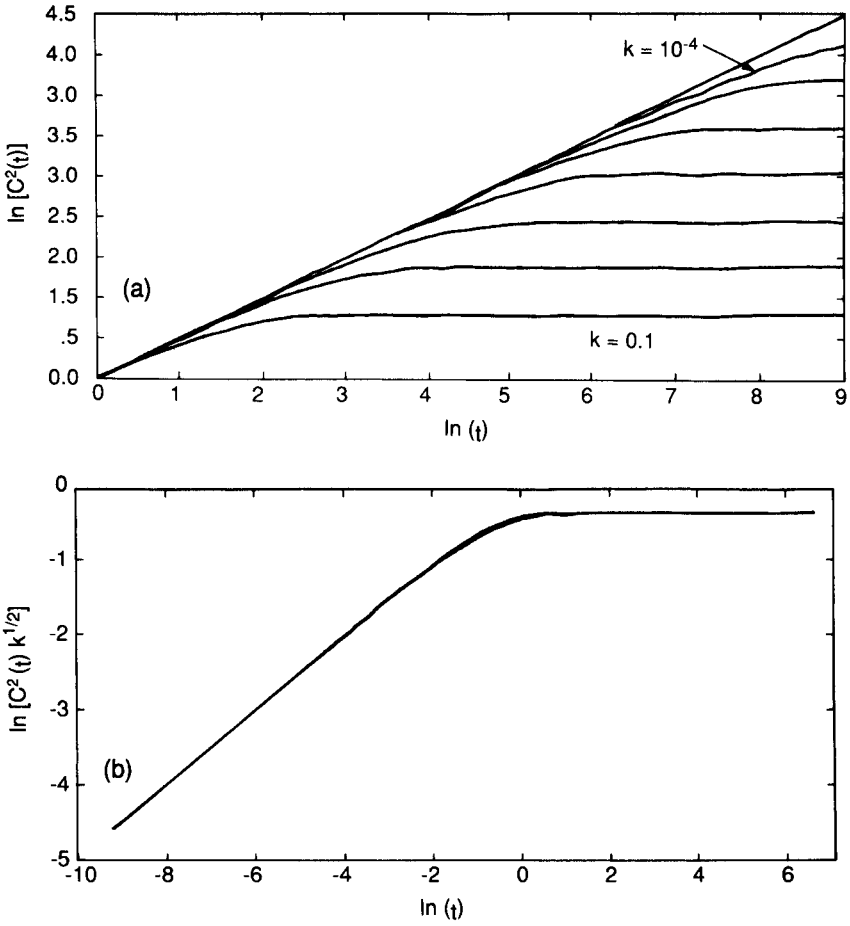


Fig. 9 (a) Correlation functions $C_2(t)$, obtained from simulations carried out using the model illustrated in Fig. 8 for eight values of k(0, 10^{-4}, 3×10^{-4}, 10^{-3}, 3×10^{-3}, 10^{-2}, 3×10^{-2}, and 10^{-1}). (*b*) How these curves can be collapsed onto a common curve using the scaling form given in Eq. (24)

for the motion of a Brownian particle in a harmonic potential. Figure 9(a) shows the height-difference correlation functions $C_2(t)$ [Eq. (13)] obtained from simulations carried out using this model for six values of the parameter k (and $k = 0$). Figure 9(b) shows that these correlation functions can be represented quite well by the scaling form

$$C_2(t) = k^{-1/2} h(kt), \qquad (26)$$

where the scaling function $h(x)$ has the form $h(x) \sim x^{-1/2}$ for $x \ll 1$ and $h(x) =$ const for $x \gg 1$. This means that the surface appears to be self-affine $\left(H = \frac{1}{2}\right)$ for $x \ll 1$ and flat for $x \gg 1$.

If a cross section is taken through a self-affine curve or surface parallel to the coarse-grained direction of the surface, then the intersection between the self-affine surface and the d-dimensional intersecting plane is a self-similar fractal

with a dimensionality D_s given by

$$D_s = d - H. \tag{27}$$

Consequently, the relatively reliable methods that have been used to analyze self-similar fractals can be used [via Eq. (27)] to measure the Hurst exponent of self-affine fractal surfaces. This is the basis of the slit island method (Mandelbrot et al., 1984) that has been applied successfully to a variety of rough surfaces.

4 Practical Considerations

In Sec. 3 it was shown that quite accurate values for the fractal dimensionality can be obtained for self-similar and self-affine surfaces using simple procedures. Very similar approaches can be used for other types of fractal structures. However, large quantities of data spanning a large range of length scales were available from the simple models used in Sec. 3. In practice the statistical uncertainties may be larger and the range of length scales smaller for experiments or more realistic models. In most cases statistical uncertainties can be reduced to quite tolerable levels by averaging, but symmetric uncertainties due to correlations to the asymptotic scaling behavior and other finite-size effects may be much more difficult to detect and control. In addition, real systems may have a much more complex scaling structure than that of the examples used here.

An account of the geometric properties of systems that must be described in terms of multifractal geometry and related concepts such as multiscaling and multiaffinity is beyond the scope of this survey. Information on those topics may be found in recent books (Feder, 1988), reviews (Paladin and Vulpiani, 1987; Stanley and Meakin, 1988; Meakin, 1990), and conference proceedings (Pietronero, 1988). In general, there is no well-established general procedure for characterizing the scaling structure of these more complex systems. In some cases it appears that the corrections to the asymptotic scaling behavior are large and convergence is slow.

One of the main difficulties in the past has been a failure to distinguish carefully between self-similar and self-affine fractals. This is particularly true in the case of surfaces and interfaces. As a consequence much of the literature in this area is subject to reinterpretation and/or is ambiguous.

Finally, there are no simple standards or criteria for determining whether a structure is fractal or not. To some extent this depends on the application and on theoretical considerations. However, geometric scaling (power-law behavior) over at least a decade of length scales combined with some understanding of the deviations from scaling outside of this range is probably a minimum requirement for fractal analysis to be a useful practical tool. In practice, scaling over more than two orders of magnitude of the length scale is rare.

Glossary

Dimensionality: Dimensionality can be defined in many ways, but in practical terms it is the exponent relating mass (or measure) to length.

Embedding Space: The Euclidean space in which a structure resides. In most practical cases this is ordinary three-dimensional

space ($\mathbb{R}^3$), but two-dimensional embedding spaces are common in computer simulations and for processes occurring at smooth interfaces.

Fractal: A structure that exhibits geometric scaling. In general terms a fractal is a structure that has similar complexity on all length scales; it "looks the same" on different length scales or at different magnifications.

Percolation: The transition associated with the formation of a continuous path spanning an arbitrarily large ("infinite") range. Site percolation on a lattice is a simple model for percolation. In this model the sites on a lattice are selected randomly and filled. For an infinitely large lattice an infinite cluster will be formed when a fraction ρ_c (the site percolation threshold probability) of the lattice sites have been filled. This cluster is a self-similar fractal. However, the entire system (including all the smaller clusters) has a finite density (ρ_c) and is uniform on all but small length scales.

Prefractal: An intermediate (nonasymptotic) stage in the construction of a regular fractal.

Radius of Gyration: The root mean square (rms) radius measured from the center of mass.

Self-Affine Fractal: A fractal that can be rescaled by a transformation that requires different changes of length scale (with different exponents) in different directions.

Self-Similar Fractal: A fractal that can be rescaled by an isotropic change of length scales (by the same amount in all directions).

List of Works Cited

Birovljev, A., Furuberg, L., Feder, J., Jøssang, T., Måløy, K. J., Aharony, A. (1991), *Phys. Rev. Lett.* **67**, 584–587.

Charmet, J. C., Roux, S., Guyon, E. (Eds.) (1989), *Disorder and Fracture*, NATO ASI Series B, Vol. 235, New York: Plenum.

Coniglio, A., Jan, N., Majid, I., Stanley, H. E. (1987), *Phys. Rev. B* **35**, 3617–3620.

Family, F., Vicsek, T. (Eds.) (1991), *Dynamics of Fractal Surfaces*, Singapore: World Scientific.

Feder, J. (1988), *Fractals*, New York: Plenum.

Grossman, T., Aharony, A. (1986), *J. Phys. A* **19**, L745–L751.

Herrmann, H. J., Roux, S. (Eds.) (1990), *Statistical Models for the Fracture of Disordered Materials*, Amsterdam: North Holland.

Holland-Moritz, E., Gordon, J., Kanazawa, K., Sonnenfeld, R. (1991), *Langmuir*, **1**, 1981–1987.

Jullien, R., Botet, R. (1986), *Aggregation and Fractal Aggregates*, Singapore: World Scientific.

Krug, J., Spohn, H. (1991) in: C. Godreche (Ed.), *Solids Far from Equilibrium: Growth Morphology and Defects*, Cambridge, U.K.: Cambridge Univ. Press.

Lenormand, R., Bories, S. (1980), *C. R. Acad. Sci. (Paris)* **B291**, 279–281.

Mandelbrot, B. B. (1977), *Fractals: Form, Chance and Dimension*, San Francisco: Freeman.

Mandelbrot, B. B. (1982), *Fractal Geometry of Nature*, New York: Freeman.

Mandelbrot, B. B., Passoja, D. E., Paullay, A. J. (1984), *Nature* **308**, 721–722.

Meakin, P. (1987), *CRC Crit. Rev. Solid State Mater. Sci.* **13**, 143.

Meakin, P. (1988), C. Domb and J. L. Lebowitz (Eds.), in: *Phase Transitions and Critical Phenomena*, New York: Academic, Vol. 12, p. 353–498.

Meakin, P. (1990), *Prog. Solid State Chem.* **20**, 135–233.

Nowicki, B. (1985), *Wear* **102**, 161–176.

Oppenheim, C. I., Trevor, D. J., Chidsey, C. E. D., Trevor, P. L., Sieradski, K. (1991), *Science* **254**, 687–689.

Paladin, G. and Vulpiani, A. (1987), *Phys. Rep.* **156**, 147–225.

Pietronero, L. (1988) (Ed.), *Fractals: Physical Origin and Properties*, New York: Plenum.

Saleur, H., Duplantier, B. (1987), *Phys. Rev. Lett.* **58**, 2325–2328.

Stanley, H. E., Meakin, P. (1988) *Nature* **335**, 405–409.
Stauffer, D. (1985), *Introduction to Percolation Theory*, London: Taylor and Francis.
Vicsek, T. (1989), *Fractal Growth Phenomena*, Singapore: World Scientific.
Wilkinson, D. (1984), *Phys. Rev. A* **30**, 520–531.
Wilkinson, D., Willemsen, J. (1983), *J. Phys. A* **16**, 3365–3376.

Further Reading

At the present time several dozen books concerned with fractal geometry and its applications in the physical sciences have appeared. The Mandelbrot (1977, 1982) classics (particularly *The Fractal Geometry of Nature*, Mandelbrot, 1982) are still a primary source of information. For those interested in the applications of fractal geometry to physical processes the books by Feder (1988) and Vicsek (1989) are highly recommended. A reprint collection (with useful commentaries) assembled by Family and Vicsek (1991) provides an up-to-date account of the rapidly developing surface growth area. Surveys of the applications to growth phenomena may be found in the books by Feder and Vicsek and recent reviews (Meakin, 1988, 1990). A collection of reviews concerned with application in chemistry has been edited by Avnir [D. Avnir (Ed.) (1989), *The Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers*, Chichester: Wiley]. Many conference proceedings have appeared: A selection of those most relevant to applied physics include A. Aharony and J. Feder (Eds.) (1989), *Fractals in Physics, Essays in Honour of Benoit B. Mandelbrot*, Amsterdam: North Holland; M. Fleischmann, D. J. Tildesley, and R. C. Ball (Eds.), (1990), *Fractals in the Natural Sciences*, Princeton: Princeton Univ. Press; and the proceedings edited by Pietronero (1988).

Geometrical Methods

V. Alan Kostelecký
Physics Department, Indiana University, Bloomington, Indiana, USA

Introduction ***128***
1 Analytic Geometry ***128***
1.1 Plane Analytic Geometry *129*
1.2 Conic Sections *130*
1.3 Plane Trigonometry *131*
1.4 Curvilinear Coordinates *132*
1.5 Solid Analytic Geometry *133*
1.6 Example: The Kepler Problem *135*
2 Differential Geometry ***136***
2.1 Manifolds *137*
2.2 Vectors and One-Forms *137*
2.3 Tensors *139*
2.4 Differential Forms *141*
2.5 Fiber Bundles *142*
2.6 Connection and Curvature *143*
2.7 Example: Electromagnetism *144*
2.8 Complex Manifolds *147*
2.9 Global Considerations *149*
2.10 Further Examples *150*
3 Projective Geometry ***151***
3.1 Some Theorems *152*
3.2 Homogeneous Coordinates *153*
3.3 Group of Projective Transformations *153*
4 Algebraic Geometry ***154***
4.1 Affine Varieties *154*
4.2 Projective Varieties *155*
4.3 Classification *155*
Glossary ***156***
Further Reading ***157***

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Introduction

The word "geometry" derives from Greek, meaning "earth measurement." Geometry was originally the mathematics describing the shapes of objects and their spatial relationships. Simple geometrical notions and ideas were known to ancient Babylonians and Egyptians 4000 years ago. Starting approximately 2500 years ago, the ancient Greeks developed fundamental geometrical ideas, including some relatively rigorous proofs based on logical reasoning. Dating from this era is Euclid's *Elements*, which introduced the basis for the axiomatic method and summarizes the knowledge at that time.

Prior to the sixteenth century, geometry and algebra were treated as independent subjects. The notion of combining the two was introduced in 1631 by René Descartes (1596–1650). This led to the field of analytic geometry, which permits the investigation of geometric questions using analytical methods. This area was extensively investigated in the eighteenth century, in particular by Leonhard Euler (1707–1783) and Gaspard Monge (1746–1818). Toward the end of the eighteenth century the use of calculus resulted in the beginnings of differential geometry, studied by Christian Gauss (1777–1855) and others. The introduction by Bernhard Riemann (1826–1866) of the theory of algebraic functions initiated the field of algebraic geometry. In parallel with these developments, the synthetic approach to geometry was extended by Victor Poncelet (1788–1867), who formulated postulates for projective geometry. In the past century and a half, the work of David Hilbert (1862–1943) and others has led to an extension of the scope of geometry to include the study of geometrical relationships between abstract quantities.

This article presents material concerning analytical, differential, projective, and algebraic geometry. The choice of topics and their depth of coverage were dictated primarily by consideration of their importance in applied physics and by limitations of space. In particular, the reader is warned that the weighting assigned to the topics discussed is uncorrelated with their present importance as mathematical fields of research. The treatment is not mathematically rigorous, but introduces sufficient mathematical terminology to make basic textbooks in the subject accessible. Some of these are listed in the references at the end of the article.

1 Analytic Geometry

The underlying concepts of analytic geometry are the simple geometric elements: points, lines and curves, planes and surfaces, and extensions to higher dimensions. The fundamental method is the use of coordinates to convert geometrical questions into algebraic ones. This is called the "method of coordinates."

To illustrate the basic notion, consider a straight line l. Following the method of coordinates, select one point O on l as the origin. This separates l into two halves. Call one half positive, the other negative. Any point P on the line can then be labeled by a real number, given by the distance OP for the positive half and by the negative of the distance OP for the negative half. There is thus a unique real number x assigned to every point P on l, called the Cartesian coordinate of P. Geometrical questions about the line can now be transcribed into analytical ones involving x.

1.1 Plane Analytic Geometry

In two dimensions, basic geometric entities include points, lines, and planes. For a plane π the method of coordinates provides to each point P an assignment of two real numbers, obtained as follows. Take two straight lines in the plane, and attribute Cartesian coordinates to each line as described above. For simplicity, the lines will be assumed perpendicular and intersecting at their origins. These lines are called rectangular coordinate axes, and the Cartesian coordinates of the first are called abscissae while those of the second are called ordinates. The lines themselves are also referred to as the abscissa and the ordinate. The location of a point P on π is then specified uniquely by two real numbers, written (x, y). The number x is defined as the perpendicular distance to the first coordinate axis, while y is the distance to the second. Using these Cartesian coordinates, geometrical questions about points can be expressed in analytical terms. For example, a formula for the distance d between two points P and Q specified by the coordinates (x_1, y_1) and (x_2, y_2) is

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2)}. \qquad (1)$$

Given an assignment of Cartesian coordinates on a plane π, a curve segment s in the plane may be analytically specified by providing a set of paired real numbers (x, y) assigned to all points on the curve. In many useful cases, s can be specified by an equation $f(x, y) = 0$ between x and y that is satisfied by all points P on s but not by any other points on π. For example, the equation $x = 0$ describes the straight line consisting of all points having coordinates of the form $(0, y)$, i.e., the ordinate. The method of coordinates thus permits geometrical questions about s to be transcribed into analytical ones concerning $f(x, y)$. For example, the set of points lying both on a curve $f(x, y) = 0$ and on another curve $g(x, y) = 0$ is specified by values (x, y) satisfying both equations, which can in principle be found by analytical methods.

A simple example of a curve in the plane is a straight line l. The slope m of l can be defined in terms of the coordinates (x_1, y_1) and (x_2, y_2) of any two distinct points on l. Provided $x_1 \neq x_2$, the slope is given by

$$m = \frac{y_1 - y_2}{x_1 - x_2}. \qquad (2)$$

The slope is zero for lines parallel to the abscissa, and is undefined (infinite) for lines parallel to the ordinate. A line l with given finite slope m is uniquely specified by its intersection point $(x, y) = (0, c)$ with the ordinate. The equation for l is

$$y = mx + c. \qquad (3)$$

If l is parallel to the ordinate instead, it is determined by its intersection point $(x, y) = (a, 0)$ with the abscissa, and its equation is simply $x = a$.

The equation of a straight line l is also determined entirely by the coordinates (x_1, y_1) and (x_2, y_2) of any two distinct points on l. It can be written

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}. \qquad (4)$$

Alternatively, a straight line can be viewed as the curve given by the most general equation linear in the coordinates x and y:

$$Ax + By + C = 0, \qquad (5)$$

where at least one of A and B is nonzero.

Analytical solutions to geometrical problems involving straight lines and points can be obtained using the above results.

For example, the equation of the line l_P that is perpendicular to a given line l with equation $y = mx + c$ and that passes through the point P on l with abscissa x_1 is

$$y = -\frac{1}{m}x + \frac{m^2+1}{m}x_1 + c. \qquad (6)$$

Another example is the expression for the perpendicular distance d_P between a line l with equation $y = mx + c$ and a point P at (a, b), which is

$$d_P = \frac{|b - ma - c|}{\sqrt{m^2+1}}. \qquad (7)$$

1.2 Conic Sections

An important curve is the circle, denoted by S^1, which can be viewed as the set of points in the plane that are equidistant from a specified fixed point. The fixed point C is called the center of the circle, and the distance r between the center and the points on the circle is called the radius. If the Cartesian coordinates of C are (h, k), then the equation of the circle is

$$(x-h)^2 + (y-k)^2 = r^2. \qquad (8)$$

The circle is a special case of a set of curves called conic sections or conics. These curves include ellipses, parabolas, and hyperbolas. Geometrically, the conics can be introduced as the curves obtained by slicing a right circular cone with a plane. Analytically, they can be viewed as the curves given by the most general expression quadratic in the coordinates x and y:

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0, \qquad (9)$$

where at least one of the coefficients A, B, C is nonzero. These are called second-order curves. From this equation it follows that any five points lying on the conic specify it completely. The quantity $B^2 - 4AC$ is called the discriminant of the conic. If the discriminant is positive, the conic is a hyperbola; if negative, an ellipse; and if zero, a parabola.

A third definition, combining geometrical and analytical notions, is often useful. Consider a straight line l, a fixed point F not on l, and another point P. Denote the distance between P and F by d_F and the perpendicular distance between P and l by d_l. Then the conic sections are given by the set of points P that obey the equation

$$d_F = ed_l, \qquad (10)$$

where $e \geq 0$ is a constant real number called the eccentricity. The line l is called the directrix and the point F is called the focus. If $e > 1$, the conic is a hyperbola. If $e = 1$, the conic is a parabola. If $0 < e < 1$, the conic is an ellipse. The degenerate case $e = 0$ gives a circle; in this case, the directrix is at infinity.

The equation determining a conic has a particularly simple form, called the canonical form, if the focus F is chosen to lie on the abscissa and the directrix l is chosen parallel to the ordinate. The canonical form depends on at most two real positive parameters a and b, where $a \geq b$ is taken for convenience.

For a hyperbola, the canonical form is

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) = 1. \qquad (11)$$

The eccentricity of the hyperbola is $e = \sqrt{a^2+b^2}/a$. One focus is the point $(ae, 0)$, and the corresponding directrix is the line $x = a/e$. There is a second focus at $(-ae, 0)$ and a second directrix at $x = -a/e$. The hyperbola has two branches, each of which asymptotically approaches

the lines $y = \pm bx/a$ as $|x|$ becomes large. The distance between the points where the hyperbola intersects the abscissa is $2a$. This is also the difference between the distances from the two foci to any given point on the hyperbola. If $a = b$, the hyperbola is called rectangular.

For a parabola, the canonical form is

$$y^2 = 4ax. \tag{12}$$

The eccentricity is $e = 1$, the focus is at $(a, 0)$, and the directrix is the line $x = -a$.

For an ellipse, the canonical form is

$$\left(\frac{x^2}{a^2}\right) + \left(\frac{y^2}{b^2}\right) = 1. \tag{13}$$

The ellipse has eccentricity $e = \sqrt{a^2 - b^2}/a$. There are again two foci, at $(\pm ae, 0)$, and two directrices $x = \pm a/e$. The sum of the distances from the two foci to any given point on the ellipse is a constant, $2a$. The line between the points of intersection of the ellipse with the abscissa is called the major axis of the ellipse, and it has length $2a$. Similarly, the minor axis of the ellipse is given by the intersection points with the ordinate and has length $2b$. If $a = b$ the equation reduces to that of a circle of radius a centered at the origin.

1.3 Plane Trigonometry

Consider a point P with coordinates (x, y) lying on a circle of radius r centered at the origin O. Denote by X the point $(x, 0)$. Call θ the angle XOP between the line segments OX and OP. The choice of a unit of measure for angles permits the assignment of a numerical value to θ. One widely used unit is the degree, defined by the statement that there are 360 degrees in a circle. The SI unit is the radian, of which there are 2π in a circle.

Certain functions of the angle θ, called trigonometric or circular functions, are of particular use in plane analytic geometry. The ratio $\sin\theta = y/r$ is called the sine of θ while $\cos\theta = x/r$ is called the cosine of θ. The sine is odd in θ while the cosine is even, and both functions have period π radians. They obey the relations

$$\sin^2\theta + \cos^2\theta = 1 \tag{14}$$

following from the Pythagorean theorem, and

$$\sin(\theta \pm \phi) = \sin\theta\cos\phi \pm \sin\phi\cos\theta, \tag{15}$$

$$\cos(\theta \pm \phi) = \cos\theta\cos\phi \mp \sin\theta\sin\phi. \tag{16}$$

The latter two equations are called addition formulas. Other, related functions of θ include the tangent $\tan\theta = y/x = \sin\theta/\cos\theta$, the cosecant $\csc\theta = r/y = 1/\sin\theta$, the secant $\sec\theta = r/x = 1/\cos\theta$, and the cotangent $\cot\theta = x/y = \cos\theta/\sin\theta$. From these definitions and the above equations many identities can be obtained. Inverse trigonometric functions can also be introduced; for example, if $x = \sin\theta$ then $\sin^{-1}x = \theta$.

Consider a triangle with angles A, B, C and sides of length a, b, c, where by convention the side labeled by a is opposite the vertex with angle A and there are similar conventions for the other sides. A basic problem in plane trigonometry is to determine one of a, b, c, A, B, C in terms of the others. This is called solving a triangle. The following relations hold: the law of sines,

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}; \tag{17}$$

the first law of cosines,

$$a = b\ \cos C + c\ \cos B; \tag{18}$$

and the second law of cosines,

$$a^2 = b^2 + c^2 - 2bc\ \cos A. \tag{19}$$

1.4 Curvilinear Coordinates

For certain geometrical problems, the analytical details of a calculation may be simplified if a non-Cartesian coordinate system is used. Consider two functions $u = u(x, y)$ and $v = v(x, y)$ of the Cartesian coordinates x and y on a plane π. Take the functions to be continuous and invertible, except perhaps at certain special points that require separate treatment. Any curve $u = c$ for some constant c is called a coordinate curve, as is any curve $v = c$. A point P on π is uniquely specified by two real numbers (u_1, v_1) that are the values of the constants determining the two coordinate curves passing through P. This construction generalizes the method of coordinates, and the functions u and v are called curvilinear coordinates. If the coordinate curves meet at right angles, the curvilinear coordinates are called orthogonal. All the analytical geometry described above using Cartesian coordinates can be rephrased using orthogonal curvilinear coordinates.

An important set of orthogonal curvilinear coordinates is generated by the equations

$$x = r\cos\theta, \quad y = r\sin\theta;$$
$$r = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}\left(\frac{y}{x}\right), \tag{20}$$

where $r \geq 0$ and $0 \leq \theta < 2\pi$. In this system, the coordinate curves consist of circles of varying radii centered at the origin and straight lines through the origin at varying angles with respect to the abscissa. The coordinates (r, θ) of a point P are called plane polar coordinates. As an illustration of their use, consider the conic sections expressed in polar coordinates. In canonical form, with the origin of the polar coordinates placed at the focus at $(ae, 0)$, the equation for a conic section is

$$r = \frac{l}{(1 + e\cos\theta)}, \tag{21}$$

where l is called the latus rectum. It is given by $l = b^2/a$ for hyperbolas and ellipses and by $l = 2a$ for parabolas, and it represents the distance from the focus to the curve as measured along a straight line parallel to the ordinate. The quantity l/e is the distance from the focus to the associated directrix.

The conic sections themselves can be used to generate systems of orthogonal curvilinear coordinates. For example, parabolic coordinates can be defined by

$$x = \tfrac{1}{2}(u^2 - v^2), \quad y = uv, \tag{22}$$

where $v \geq 0$. The coordinate curves are parabolas. Similarly, elliptic coordinates can be defined by

$$x = a\cosh u\cos v, \quad y = a\sinh u\sin v, \tag{23}$$

where $u \geq 0$ and $0 \leq v < 2\pi$. Here, the so-called hyperbolic functions $\sinh u$ and $\cosh u$ are defined by

$$\sinh u = \tfrac{1}{2}(e^u - e^{-u}),$$
$$\cosh u = \tfrac{1}{2}(e^u + e^{-u}). \tag{24}$$

The coordinate curves are ellipses and hyperbolas. Another common set is the system of bipolar coordinates, defined by

$$x = \frac{a\sinh v}{\cosh v - \cos u'}, \quad y = \frac{a\sin u}{\cosh v - \cos u'} \tag{25}$$

with $0 \leq u < 2\pi$. The coordinate curves are sets of intersecting circles.

1.5 Solid Analytic Geometry

Solid analytic geometry involves the study of geometry in three dimensions rather than two. Many of the ideas of plane analytic geometry extend to three dimensions. For instance, the method of coordinates now provides an assignment of three real numbers (x,y,z) to each point P. A three-dimensional rectangular coordinate system can be introduced by taking three mutually perpendicular straight lines, each given Cartesian coordinates, to form the coordinate axes. The axes are called the abscissa, the ordinate, and the applicate. Each of the values (x,y,z) is defined as the perpendicular distance to the corresponding axis.

A two-dimensional surface σ can now be specified by providing an equation $f(x, y, z) = 0$ that is satisfied only by points on the surface. The method of coordinates thus converts geometrical questions about σ to analytical questions about $f(x,y,z)$. Similarly, a curve s can be viewed as the intersection set of two surfaces. If the surfaces are specified by the equations $f(x, y, z) = 0$ and $g(x, y, z) = 0$, s is given analytically by the set of points (x,y,z) obeying both equations simultaneously.

By definition, a surface of the first order is given by the most general equation linear in x, y, z:

$$Ax + By + Cz + D = 0. \tag{26}$$

If at least one of A, B, C is nonzero, this equation describes a plane. A straight line can be viewed as the intersection of two nonparallel planes and is therefore given analytically by two equations of this form. Just as in the two-dimensional case, the analytical formulation allows solutions to geometrical questions involving planes, lines, and points to be obtained. For example, the perpendicular distance d_P between a plane given by the above equation and a point P located at (a,b,c) can be shown to be

$$d_P = \frac{|Aa + Bb + Cc + D|}{\sqrt{A^2 + B^2 + C^2}}. \tag{27}$$

As another example, two planes given by

$$\begin{aligned} A_1x + B_1y + C_1z + D_1 &= 0, \\ A_2x + B_2y + C_2z + D_2 &= 0 \end{aligned} \tag{28}$$

are parallel if and only if

$$(A_1, B_1, C_1) = (cA_2, cB_2, cC_2) \tag{29}$$

for some constant c.

In analogy to the two-dimensional introduction of conics as curves obeying a quadratic expression in x and y, a surface of the second order is defined to consist of points satisfying a quadratic expression in x, y, and z:

$$\begin{aligned} Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz \\ + Gx + Hy + Iz + J = 0. \end{aligned} \tag{30}$$

Such surfaces are also called quadrics. An important example is the sphere, denoted by S^2, which can be viewed as the set of points equidistant from a fixed point called the center. The distance from the center to any point on the sphere is called the radius. If the Cartesian coordinates of the center are (h,k,l), the equation of a sphere of radius r is

$$(x - h)^2 + (y - k)^2 + (z - l)^2 = r^2. \tag{31}$$

The quadrics can be classified. Among the surfaces described are ellipsoids,

hyperboloids, paraboloids, cylinders, and cones. Canonical forms of these surfaces are

$$\left(\frac{x^2}{a^2}\right) + \left(\frac{y^2}{b^2}\right) + \left(\frac{z^2}{c^2}\right) = 1 \qquad (32)$$

for an ellipsoid;

$$\left(\frac{x^2}{a^2}\right) + \left(\frac{y^2}{b^2}\right) - \left(\frac{z^2}{c^2}\right) = 1 \qquad (33)$$

for a hyperboloid of one sheet;

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) - \left(\frac{z^2}{c^2}\right) = 1 \qquad (34)$$

for a hyperboloid of two sheets;

$$\left(\frac{x^2}{a^2}\right) + \left(\frac{y^2}{b^2}\right) = 2z \qquad (35)$$

for an elliptic paraboloid;

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) = 2z \qquad (36)$$

for a hyperbolic paraboloid;

$$\left(\frac{x^2}{a^2}\right) + \left(\frac{y^2}{b^2}\right) = 1 \qquad (37)$$

for an elliptic cylinder;

$$\left(\frac{x^2}{a^2}\right) - \left(\frac{y^2}{b^2}\right) = 1 \qquad (38)$$

for a hyperbolic cylinder;

$$\left(\frac{x^2}{a^2}\right) = 2z \qquad (39)$$

for a parabolic cylinder; and

$$\left(\frac{x^2}{a^2}\right) \pm \left(\frac{y^2}{b^2}\right) - \left(\frac{z^2}{c^2}\right) = 0 \qquad (40)$$

for a cone. The parameters a, b, c are called the lengths of the principal axes of the quadric.

The notions of plane trigonometry also extend to three dimensions. A spherical triangle is defined as a portion of a spherical surface that is bounded by three arcs of great circles. Denote by A, B, C the angles generated by straight lines tangent to the great circles intersecting at the vertices, and call the lengths of the opposite sides a, b, c as for the planar case. The angles now add up to more than π radians, by an amount called the spherical excess E:

$$A + B + C = \pi + E. \qquad (41)$$

The following relations hold for a spherical triangle:
the law of sines,

$$\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}; \qquad (42)$$

the first law of cosines,

$$\cos a = \cos b \cos c + \sin b \sin c \cos A; \qquad (43)$$

and the second law of cosines,

$$\cos A = -\cos B \cos C + \sin B \sin C \cos a. \qquad (44)$$

Curvilinear coordinates can be introduced via three locally continuous invertible functions $u(x, y, z)$, $v(x, y, z)$, $w(x, y, z)$, following the two-dimensional case. A coordinate surface is specified by setting any curvilinear coordinate u, v, or w equal to a constant. The coordinate curves are generated by the intersection of the coordinate surfaces, and the system is said to be orthogonal if the surfaces intersect at right angles. Many useful three-dimensional orthogonal curvilinear coordinate systems can be generated from families of quadrics. One particularly useful set is the system of spherical polar

coordinates, defined by

$$x = r\sin\theta\cos\phi,$$
$$y = r\sin\theta\sin\phi,$$
$$z = r\cos\theta, \tag{45}$$

where $r \geq 0$, $0 \leq \theta \leq \pi$, and $0 \leq \phi < 2\pi$. The coordinate surfaces are spheres centered at the origin, right circular cones with axes along the applicate and vertices at the origin, and half-planes with the applicate as one edge. Other common coordinates are the cylindrical coordinates, given by

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z, \tag{46}$$

where $r \geq 0$ and $0 \leq \theta < 2\pi$. This system is generated from plane polar coordinates by translation along the applicate. The coordinate surfaces are right circular cylinders centered at the origin, half-planes with the applicate as one edge, and planes parallel to the plane of the abscissa and ordinate.

The notions of plane and solid analytic geometry can be extended to higher dimensions, too. A space can be defined in which the method of coordinates specifies a point by n real numbers $(x^1, x^2, \ldots, x^n)$. This n-dimensional space, called Euclidean space, is denoted by the symbol R^n. Using coordinates, geometrical questions in n dimensions can be converted to analytical ones involving functions of n variables. Surfaces of the first order are $(n-1)$-dimensional hyperplanes, and surfaces of the second order, or quadric hypersurfaces, can be introduced. An example is the hypersphere of radius r in n dimensions, denoted by S^{n-1}, which when centered at the origin satisfies the equation

$$(x^1)^2 + (x^2)^2 + \cdots + (x^n)^2 = r^2. \tag{47}$$

The notion of curvilinear coordinates also extends to higher dimensions.

A one-dimensional curve s in n dimensions can be specified by $n-1$ equations among the n coordinates $(x^1, \ldots, x^n)$. If s is continuous, its points can be labeled by a parameter t that is a real number. Any particular point can be specified by giving the values of the n coordinates $(x^1, \ldots, x^n)$. As t varies, so do the coordinates. This means that an alternative specification of s can be given in terms of the n expressions

$$x^j = x^j(t), \quad j = 1, \ldots, n, \tag{48}$$

determining the n coordinates $(x^1, \ldots, x^n)$ as functions of t. This is called the parametric representation of a curve. Similarly, the points of a continuous two-dimensional surface can be labeled by two real numbers (t^1, t^2). The surface can be specified either in terms of $n-2$ equations among the n coordinates $(x^1, \ldots, x^n)$ or in parametric form by the n equations

$$x^j = x^j(t^1, t^2), \quad j = 1, \ldots, n. \tag{49}$$

A parametric representation can also be given for continuous surfaces of more than two dimensions.

1.6 Example: The Kepler Problem

An example of the appearance of analytic geometry in a physical problem occurs in the study of the classical motion of two bodies under a mutual inverse-square attractive force. Consider for definiteness two bodies of masses m_1 and m_2, each acted on by the gravitational field of the other and free to move in three dimensions. This is called the Kepler problem.

The first step is to introduce a convenient coordinate system. For simplicity, the origin can be placed on one mass. The problem can then be reduced to determining the relative position of the

second mass and the uniform motion of the center of mass. The latter is neglected here for simplicity. It is natural to select a system of spherical polar coordinates with the applicate along the direction of the angular momentum. Since angular momentum is conserved, the motion of the second mass about the origin must lie in a plane. This means that plane polar coordinates $(r,\ \theta)$ suffice to describe the position of the second mass relative to the first.

It can be shown that the resulting equations governing the motion of the second mass are precisely those obtained for the behavior of a reduced mass $m = m_1 m_2/(m_1 + m_2)$ orbiting a fixed center of force. In polar coordinates, the kinetic energy T of the reduced mass is

$$T = \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2), \tag{50}$$

where a dot over a letter signifies a derivative with respect to time. The potential energy is

$$V = \frac{-k}{r} \tag{51}$$

with $k = Gm_1m_2$, where G is Newton's gravitational constant.

The equations of motion are

$$\frac{d}{dt}(mr^2\dot{\theta}) = 0 \tag{52}$$

and

$$m\ddot{r} - mr\dot{\theta}^2 + \left(\frac{k}{r^2}\right) = 0. \tag{53}$$

The first of these integrates immediately in terms of the constant magnitude L of the angular momentum:

$$mr^2\dot{\theta} = L. \tag{54}$$

This equation can be used to eliminate $\dot{\theta}$ from Eq. (53) by direct substitution. Also, since

$$\frac{d}{dt} = \dot{\theta}\frac{d}{d\theta}, \tag{55}$$

the independent variable in Eq. (53) can be converted from time t to angle θ. An additional change of dependent variable from r to

$$s = \left(\frac{1}{r}\right) - \left(\frac{mk}{L^2}\right) \tag{56}$$

converts Eq. (53) into the simple form

$$\frac{d^2s}{d\theta^2} = -s. \tag{57}$$

The solution is readily found. Reconverting s to the variable r yields the equation for the orbit as

$$r = \frac{l}{(1 + e\cos\theta)}, \tag{58}$$

where a particular choice for the location of the abscissa relative to the orbit has been made for simplicity. In this equation,

$$e = \sqrt{\frac{1 + 2EL^2}{mk^2}}, \quad l = \frac{L^2}{mk}, \tag{59}$$

and E can be identified with the energy of the two bodies in the orbit. This demonstrates that the motion of two masses under gravity is described by a conic section; cf. Eq. (21). The energy E determines the shape of the orbit. If $E > 0, e > 1$ and the orbit is a hyperbola. If $E = 0,\ e = 1$ and the orbit is a parabola. If $E < 0, e < 1$ and the orbit is an ellipse. Finally, if $E = -mk^2/2L^2,\ e = 0$ and the orbit is a circle.

2 Differential Geometry

The requirement of differentiability provides a restriction on geometrical objects that is sufficiently tight for new and useful results to be obtained and sufficiently loose to include plenty of interesting cases. Differential geometry is of vital importance in physics because many physical problems

involve variables that are both continuous and differentiable throughout their range.

2.1 Manifolds

A manifold is an extension of the usual notions of curves and surfaces to arbitrary dimensions. The basic idea is to introduce an n-dimensional manifold as a space that is like Euclidean space R^n locally, i.e., near each point. Globally, i.e., taken as a whole, a manifold may be very different from R^n. An example of a one-dimensional manifold is a straight line. This is both locally and globally like R^1. Another one-dimensional example is a circle S^1. The neighborhood of each point on a circle looks like the neighborhood of a point in R^1, but globally the two are different. The circle can be constructed by taking two pieces of R^1, bending them, and attaching them smoothly at each end. Generalized to n dimensions, this notion of taking pieces of R^n and attaching them smoothly forms the basis for the definition of a manifold.

To define a manifold more rigorously, first introduce the concept of a topological space T. This is a set S and a collection t of (open) subsets of S satisfying the following criteria:

1. Both the null set and S itself are in t.
2. The intersection of any two subsets of t is in t.
3. The union of any collection of subsets of t is in t.

Suppose in addition there is a criterion of separability: For any two elements of S there exist two disjoint subsets of S, each containing one of the elements. Then T is called a Hausdorff space. The elements of S for a manifold are its points.

Next, define a chart C of the set S as a subset U of S, called a neighborhood, together with a continuous invertible map $f : U \to R^n$ called the coordinate function. For a manifold, the subset U plays the role of a region locally like R^n, and the function f represents the introduction of local coordinates in that region. Two charts C_1, C_2 with overlapping neighborhoods and coordinate functions f_1, f_2 are called compatible if the composition map $f_1 \circ f_2^{-1}$ is differentiable. The requirement of compatibility basically ensures that the transition from one coordinate patch to another is smooth. A set of compatible charts covering S is called an atlas.

A differentiable manifold M can now be defined as a Hausdorff topological space with an atlas. Given that the range of the coordinate functions is R^n, the dimension of M is defined as n and M is sometimes denoted by M^n. An example of an n-dimensional manifold is the hypersphere S^n. An example of an object that is *not* a manifold is a figure-eight curve, since the neighborhood of the intersection point is not locally like R^n for any n.

2.2 Vectors and One-Forms

The usual definition of a vector in a Euclidean space as a directed straight-line segment does not immediately extend to a general manifold. For instance, the circle S^1 does not contain any straight-line segments. Instead, vectors at a point of a manifold can be introduced using the notion of the tangents to curves passing through the point.

Consider a curve s through a point P. In a neighborhood of P, local coordinates $(x^1, \ldots, x^n)$ can be used to specify s in the parametric representation $x^j = x^j(t)$, $j = 1, \ldots, n$. A vector tangent to s at P can be specified by the n quantities dx^j/dt forming its components. A familiar

example in mechanics is the velocity vector of a moving particle, obtained by differentiation with respect to time of the particle's position vector. If the tangent vectors to all possible curves in the manifold through P are considered, an n-dimensional vector space (see ALGEBRAIC METHODS, See. 3) is generated. This is called the tangent space T_PM to M at P.

In differential geometry, it is desirable to introduce basic concepts in a manner that is independent of any coordinate choice. For this reason, the differential-geometric definition of a tangent vector is different from the more intuitive one above. Given a curve s, introduce an arbitrary differentiable function f assigning a real number to every point t on s. The derivative df/dt of f along s is called the directional derivative. In a local coordinate patch,

$$\frac{df}{dt} = \sum_{j=1}^{n} \frac{dx^j}{dt} \partial_j f, \tag{60}$$

where $\partial_j f = \partial f/\partial x^j$. This shows that the operator d/dt acting on the space of real functions on M contains all components of the tangent vector, each associated with the corresponding partial derivative ∂_j. A tangent vector at P can therefore be defined as the directional-derivative operator d/dt, with a natural coordinate basis of vectors for the vector space being the set of partial-derivative operators $\{\partial_j\}$. However, this definition has the disadvantage that it still explicitly includes the parameter t.

The formal definition of a tangent vector is therefore slightly more abstract. Given the space $F(M)$ of all differentiable real functions on a manifold M, a tangent vector at P is defined as an object v acting on elements of $F(M)$ to produce real numbers,

$$v : F(M) \to R, \tag{61}$$

that satisfies two criteria:

$$v(af + bg) = av(f) + bf(g),$$
$$v(f \circ g) = g(P)v(f) + f(P)v(g), \tag{62}$$

where $f, g \in F(M)$ and $a, b \in R$. This definition extracts the important properties of the tangent vector without explicit reference to a coordinate system or parametrization. Note that the coordinate realization of a tangent vector at P along x_j as ∂_j acting at P satisfies this definition. The set of all linearly independent tangent vectors at P spans the tangent space T_PM to M at P, and the set $\{\partial_j\}$ forms a basis for T_PM called the coordinate basis. An arbitrary vector v can be expanded in this basis as $v = \Sigma_j v^j \partial_j$. Physicists sometimes say that the components v^j are the contravariant components of a vector. Although in a coordinate basis the intuitive physics notion of a vector and the differential-geometric one contain the same information about components, the latter also contains information about the coordinate basis itself. In the remainder of this article except where noted, the word vector refers to the differential-geometric object.

Since T_PM is a vector space, there exists a dual vector space $\mathrm{Hom}(T_PM, R)$ consisting of linear maps

$$\omega : T_PM \to R \tag{63}$$

(see ALGEBRAIC METHODS, Sec. 3.7). This space is called the cotangent space at P and is denoted by T_P^*M. Notice that duality also implies $T_PM = \mathrm{Hom}(T_P^*M, R)$. Elements of T_P^*M are called one-forms. An important example of a one-form is the total differential df of a function $f \in F(M)$, defined as the element of T_P^*M satisfying

$$df(v) = v(f) \tag{64}$$

for any $v \in T_PM$.

In a chart around P, the set $\{dx^j\}$ of total differentials of the coordinates forms a natural coordinate basis for the cotangent space T_P^*M. It is a dual basis to $\{\partial_j\}$, since

$$dx^j(\partial_k) = \partial_k x^j = \delta^j_k. \tag{65}$$

An arbitrary one-form ω can be expanded in the dual basis as $\omega = \Sigma_j \omega_j \, dx^j$. Note that for an arbitrary vector $v = \Sigma_j v^j \partial_j$ the action of ω on v is then

$$\omega(v) = \omega_j v^k \, dx^j(\partial_k) = \omega_j v^j. \tag{66}$$

In this equation and subsequent ones, the Einstein summation convention is introduced to simplify notation: Repeated indices in the same term are understood to be summed. The vector v is said to be contracted with the one-form ω. Physicists sometimes say the components ω^j form the covariant components of a vector. As an example, the definitions above can be used to show that

$$df = \partial_j f \, dx^j, \tag{67}$$

a standard result.

2.3 Tensors

The generalization of vectors and one-forms to tensors is straightforward. A tensor T of type (a, b) can be defined at a point P of a manifold M as a multilinear mapping of a one-forms and b vectors giving a real number:

$$T : T_P^*M \otimes \cdots \otimes T_P^*M \otimes T_PM \otimes \cdots \otimes T_PM \to R, \tag{68}$$

where there are a factors of T_P^*M and b factors of T_PM. The space of tensors of type (a, b) at P is denoted $T^a_b(P)$. Examples introduced above include $T^1_0(P) = T_PM$ and $T^0_1(P) = T_P^*M$.

A tensor T of type (a, b) can be expanded using a coordinate basis. In the natural basis introduced above,

$$T = T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b} \, \partial_{j_1} \cdots \partial_{j_a} \, dx^{k_1} \cdots dx^{k_b}. \tag{69}$$

Almost all physicists and the older mathematics literature call the quantities $T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b}$ the components of an ath-rank contravariant and bth-rank covariant tensor. Most modern mathematicians by convention interchange the usage of contravariant and covariant. This article uses the physicists' convention.

A tensor is called symmetric with respect to two contravariant or two covariant indices if its components are unaltered when the indices are interchanged. A tensor with indices of only one type is said to be totally symmetric if it is symmetric with respect to all pairs of indices. Similarly, a tensor is antisymmetric with respect to two contravariant or two covariant indices if its components change sign when the indices are interchanged, and a totally antisymmetric tensor is one with pairwise-antisymmetric indices of only one type. The sum and difference of two tensors of the same type is another tensor of the same type. The tensor product $T_1 \otimes T_2$ of two tensors T_1 and T_2 of types $(a_1, \ b_1)$ and $(a_2, \ b_2)$, respectively, is a tensor of type $(a_1 + a_2, \ b_1 + b_1)$ with components given by the product of components of T_1 and T_2 (see ALGEBRAIC METHODS, Sec. 3.8). Various contractions of two tensors can be introduced that generalize the contraction of a vector with a one-form.

A useful concept in physical applications is that of a tensor field of type (a, b) defined as a particular choice of tensor of type (a, b) at each point of M. The field is called smooth if the components $T^{j_1 j_2 \cdots j_a}_{k_1 k_2 \cdots k_b}$ of a tensor field are differentiable. Special

cases are vector fields and one-form fields. By convention, if $a = b = 0$ the field is called a scalar field and is just an element of $F(M)$, the real-valued functions on M.

An example of a tensor that plays a crucial role in physics is the metric tensor g. On the manifold M, it is a symmetric tensor field of type (0,2) such that if $g(v_1, \ v_2) = 0$ for any $v_1 \in T_P M$, then $v_2 = 0$. In component form in a coordinate basis near a point P

$$g = g_{jk}\, dx^j\, dx^k, \tag{70}$$

where g_{jk} form the components of a symmetric, invertible matrix. The metric tensor g associates any two vectors with a real number. For instance, in the usual geometry in a Euclidean space R^n the matrix $g_{jk} = \delta_{jk}$ and the real number is the scalar or dot product of the two vectors. In other applications different metrics may be required. For example, in special relativity space-time is taken as a four-dimensional manifold with a Minkowskian metric. If the number $g(v_1, \ v_2)$ has the same sign for all $v_1, \ v_2$ at all P on M, i.e., if the eigenvalues of the matrix g_{jk} are all of the same sign, the metric is called Riemannian. Manifolds admitting such metrics are called Riemannian manifolds. Other metrics are called pseudo-Riemannian. The special case of a metric with one eigenvalue of different sign is called Lorentzian. By diagonalization and normalization, it is always possible to choose a basis at any given P such that $g_{jk}(P)$ is a diagonal matrix with entries that are ± 1. If the entries are all of the same sign, the metric in this form is called Euclidean. If one entry has a different sign, it is called Minkowskian.

Since g is a map $T_P M \otimes T_P M \to R$, any given vector v defines a linear map $g(v)$ from $T_P M$ to R. This map is evidently a one-form, by definition. The components v_j of this one-form are given by

$$v_j = g_{jk} v^k. \tag{71}$$

The map is said to lower the index of the vector, and the result is called the associated one-form. An inverse map can be defined that uses the matrix inverse g^{jk} of g_{jk} to raise the index of a form, yielding a vector.

A significant part of the classical literature on differential geometry is concerned with the relationships between different manifolds, in particular in manifolds endowed with metrics. Consider two manifolds M_1 and M_2 of dimensions n_1 and n_2. If there exists a smooth and regular map $f : M_1 \to M_2$, then M_1 is said to be a submanifold of M_2. The map f is called an embedding. The notion of a regular map is readily understood in coordinate patches $\{x^j\}$ on a chart U in M_1 and $\{y^k\}$ on a chart V in M_2: the matrix with components $\partial y^k / \partial x^j$ must have maximal rank n_1 at each point. Intuitively, the requirements for an embedding can be viewed as ensuring for the submanifold its differentiability, the absence of self-intersections, and that curves through a point in M_1 look locally like their images in M_2. The references at the end of this article provide details of the methods and results of this subject. A simple example of a question involving the notion of embedding is the determination of equations, called the Frenet-Serret formulas, for a curve in R^n. A more complicated example is the description of the embedding of a hypersurface M into R^n, which, according to Bonnet's theorem, is determined by the metric tensor g on M (which in this context is called the first fundamental form), by another symmetric tensor of type (0,2) called the second fundamental form, and by a set of partial differential

equations called the Gauss-Codazzi equations. General results on the possibility of embedding an m-dimensional manifold into R^n are also available. An example is Whitney's theorem, which may be viewed as the statement that for compact manifolds such an embedding is possible for $n = 2m + 1$.

2.4 Differential Forms

A particularly important class of tensors is the set of totally antisymmetric tensors of type (0,p) at a point of M^n. These span a vector space denoted by $\wedge^p T_P^* M$ or just $\wedge^p T^*$, and they are called p-forms. The number $p \leq n$ is called the degree of the form. For the case $p = 0$, $\wedge^0 T_P^* M$ is chosen as $F(M)$, the space of real smooth functions on M. The dimension of $\wedge^p T^*$ as a vector space is given by the binomial coefficient ${}^nC_p$. Note that this implies that $\wedge^p T^*$ and $\wedge^{(n-p)} T^*$ have the same dimension.

Introduce the wedge product $\omega_1 \wedge \omega_2$ of two one-forms by the definition

$$\omega_1 \wedge \omega_2 = \omega_1 \otimes \omega_2 - \omega_2 \otimes \omega_1. \qquad (72)$$

By construction, this is an antisymmetric tensor of type (0,2), i.e., a two-form. It can be shown that a coordinate basis for the two-forms is the set $\{dx^j \wedge dx^k\}$. In general, antisymmetric tensor products of one-forms can be used to generate p-forms, and an element $\omega \in \wedge^p T^*$ can be expanded in a coordinate basis as

$$\omega_p = \frac{1}{p!}\omega_{j_1 \cdots j_p}\, dx^{j_1} \wedge \cdots \wedge dx^{j_p}. \qquad (73)$$

A natural induced wedge product exists that combines a p-form ω_1 with a q-form ω_2 to give a $(p+q)$-form. This product obeys

$$\omega_1 \wedge \omega_2 = (-1)^{pq}\omega_2 \wedge \omega_1. \qquad (74)$$

A larger vector space $\wedge T^*$ consisting of the direct sum of all the spaces $\wedge^p T^*$ can also be considered. Its dimension is 2^n, and it is called the Cartan exterior algebra of $T_P^* M$.

Analogous constructions can be introduced for the case of antisymmetric tensors of type (p,0), called p-vectors. The totality of these spans a space denoted $\wedge^p T$. The p-forms, $(n-p)$-forms, p-vectors, and $(n-p)$-vectors thus all form vector spaces of dimension ${}^nC_p$ at a point P of M^n. Various relations can be constructed between these spaces. An important example is the Hodge star map $*$, defined for manifolds M that have a metric g. This is a linear map $* : \wedge^p T_P^* M \to \wedge^{(n-p)} T_P^* M$ that is most easily understood by its action on coordinate components. Define the totally antisymmetric symbol by

$$e_{j_1 \cdots j_n} = \begin{cases} +1 & \text{if } (j_1 \cdots j_n) \text{ is an even permutation of } (1, \ldots, n) \\ -1 & \text{if } (j_1 \cdots j_n) \text{ is an odd permutation of } (1, \ldots, n) \\ 0 & \text{otherwise.} \end{cases} \qquad (75)$$

If a p-form ω is given in a coordinate basis by Eq. (73), then

$$*\omega = \frac{\sqrt{|g|}}{p!(n-p)!} g^{j_1 k_1} \cdots g^{j_p k_p} \omega_{k_1 \cdots k_p} e_{j_1 \cdots j_n}\, dx^{j_{p+1}} \wedge \cdots \wedge dx^{j_n}, \qquad (76)$$

where g^{jk} is the inverse metric matrix introduced in Sec. 2.3 and g is the determinant of the matrix g_{jk}.

From the definition (64), the total differential of a zero-form is a one-form. An extension of the notion of differential can be introduced to obtain a $(p+1)$-form via a p-form. Formally, a map $d : \wedge^p T^* \to \wedge^{(p+1)} T^*$ called the exterior derivative can be defined by the following requirements:

1. $d(\omega_1 + \omega_2) = d\omega_1 + d\omega_2$ for $\omega_1, \omega_2 \in \wedge^p T$;
2. $d(\omega_1 \wedge \omega_2) = (d\omega_1 \wedge \omega_2) + (-1)^p(\omega_1 \wedge d\omega_2)$ for $\omega_1 \in \wedge^p T$ and $\omega_2 \in \wedge^q T$; and
3. $d(d\omega) = 0$ for $\omega \in \wedge^p T$.

It can be shown that the exterior derivative is unique. In a coordinate basis, the exterior derivative of a p-form given by Eq. (73) is

$$d\omega_p = \left(\frac{1}{p!}\right) \partial_k \omega_{j_1 \cdots j_p}\, dx^k \wedge dx^{j_1} \wedge \cdots \wedge dx^{j_p}. \tag{77}$$

A p-form field with vanishing exterior derivative is said to be closed, while one that is obtained as the exterior derivative of a $(p-1)$-form is called exact. The definition of d implies that an exact form is necessarily closed.

The exterior derivative combines in a single notation valid for manifolds M^n extensions of the gradient, divergence, and curl operations of usual three-dimensional vector calculus. For instance, the gradient of a function f is a covariant vector with components $\partial_j f$. These are precisely the components of the one-form in Eq. (67). The components of the curl make their appearance in the exterior derivative of a one-form $\omega = \omega_x dx + \omega_y dy + \omega_z dz$:

$$\begin{aligned} d\omega = {} & (\partial_x \omega_y - \partial_y \omega_x)\, dx \wedge dy \\ & + (\partial_y \omega_z - \partial_z \omega_y) dy \wedge dz \\ & + (\partial_z \omega_x - \partial_x \omega_z)\, dz \wedge dx. \end{aligned} \tag{78}$$

The divergence enters the expression for the exterior derivative of a two-form $\omega = \omega_{xy} dx \wedge dy + \omega_{yz} dy \wedge dz + \omega_{zx} dz \wedge dx$:

$$\begin{aligned} d\omega = {} & (\partial_x \omega_{yz} + \partial_y \omega_{zx} \\ & + \partial_z \omega_{xy})\, dx \wedge dy \wedge dz. \end{aligned} \tag{79}$$

The statement $dd = 0$ contains the usual identities $\operatorname{div}(\operatorname{curl} v) = \operatorname{curl}(\operatorname{grad} f) = 0$ for a vector v and a function f.

The existence of the Hodge star map $*$ makes it possible to define a map from p-forms to $(p-1)$-forms by applying first $*$ [producing an $(n-p)$-form], then d [giving an $(n-p+1)$-form], and finally $*$ again. This map is called the adjoint exterior derivative and denoted δ. For Riemannian metrics it is defined as

$$\delta = (-1)^{np+n+1} * d *, \tag{80}$$

while for Lorentzian metrics there is an additional factor of -1. The adjoint exterior derivative satisfies $\delta\delta\omega = 0$. A p-form field with vanishing adjoint exterior derivative is said to be coclosed, while one that is obtained as the adjoint exterior derivative of a $(p+1)$-form is called coexact.

It is possible to express the Laplacian Δ on a manifold M^n in terms of the maps d and δ:

$$\Delta = (d + \delta)^2 = d\delta + \delta d. \tag{81}$$

For example, acting on a function f in three dimensions, this definition reproduces the standard expression of vector calculus,

$$\Delta f = \left(\frac{1}{\sqrt{|g|}}\right) \partial_j (\sqrt{|g|} g^{jk} \partial_k f). \tag{82}$$

A p-form ω is said to be harmonic if $\Delta\omega = 0$. This generalizes the usual notion of harmonic functions.

2.5 Fiber Bundles

In addition to involving a manifold of variables, many physical situations also exhibit symmetry of some kind. The natural geometrical framework in which to formulate such problems is the

language of fiber bundles. Here, attention is restricted to a special type of bundle, appearing widely in physics, that involves continuous symmetries. The latter are described mathematically via the theory of Lie groups.

This paragraph presents a few essential definitions involving Lie groups. More details may be found in the articles GROUP THEORY and ALGEBRAIC METHODS. For the present purposes, a Lie group G may be viewed as a group that is also an r-dimensional manifold such that for two group elements $g, h \in G$ the map $gh^{-1} : G \times G \to G$ exists and is continuous. Denote coordinates in a chart near some point P of G by $\{a^A\}$, $A = 1, \ldots, r$. Then the group composition function $f : G \times G \to G$ defined for $g(a), h(b), k(c) \in G$ by $f(h, g) = k = hg$ can be written in terms of r functions ϕ^A acting on the coordinates as

$$c^A = \phi^A(b, a). \tag{83}$$

The generators D_A of infinitesimal group transformations on G span the tangent space T_0G at the group identity and are given by

$$D_A = U_A^B \partial_B, \quad U_A^B = \left.\frac{\partial \phi^B}{\partial b^A}\right|_{b=0}. \tag{84}$$

This space is called the Lie algebra associated with the group. The dual basis is spanned by the one-forms

$$\Omega^A = da^B (U^{-1})_B^A. \tag{85}$$

As a simple example, consider the group $U(1)$. The group manifold is a circle S^1; if the coordinate is denoted by θ, the group composition function is $\theta_3 = \theta_2 + \theta_1$. The generator D_θ is just ∂_θ and the dual basis is $d\theta$.

A fiber bundle is basically a manifold acted on by a symmetry. One important type of bundle, called a principal bundle, looks locally (but not necessarily globally) like a product of a continuous symmetry group with a manifold. The action of the symmetry provides a natural means of moving around in each local piece of bundle. The idea is to patch together these local pieces in a smooth way to get the whole principal bundle. Globally, the patching can introduce various twists into the overall structure, in which case the bundle is called nontrivial. A trivial bundle is one where no twists arise: the global and local structure are similar.

Here is a more formal definition. Given a manifold B and a Lie group G, a principal fiber bundle $E(B,G)$ is a manifold such that

1. G acts differentiably and without fixed points on E;
2. B is the quotient space of E by equivalence under G, and there exists a differentiable map $\pi : E \to B$; and
3. for each chart U_j in an atlas for B, there exists a differential and invertible mapping $\phi_j : \pi^{-1}(U_j) \to U_j \times G$ given by $E \to (\pi(P), f(P))$ for any point $P \in E$, where $f : \pi^{-1}(U_j) \to G$ satisfies $f(gP) = gf(P)$ for any $g \in G$.

The group G is called the structure group and the manifold B is called the base manifold. The map π is called the projection. The inverse image of π is the fiber; in effect, each fiber is like a copy of G. A (global) cross section or section s of a bundle is defined as a smooth map s: $B \to E$ such that $\pi^\circ s$ is the identity on B. Local sections, i.e., sections defined only on $\pi^{-1}(U_j)$, always exist. If the bundle admits a global section, it is called trivial.

2.6 Connection and Curvature

Since $\{\partial_j\}$ is a basis for the tangent space of the base manifold and $\{D_A\}$ is one for the tangent space of the group, a basis for the

tangent space to a point in the bundle is the set $\{\partial_j, D_A\}$. It has dual basis $\{dx^j, \Omega^A\}$. However, linear combinations could also be taken. The existence of this freedom permits the definition of a natural one-form called the connection that contains essential information about the structure of the bundle. The connection is basically a separation of the tangent space of E into two pieces, one along the group.

Formally, a connection is a choice of a subspace T_PH of T_PE at each point P of E such that

1. $T_PE = T_PG \otimes T_PH$, where T_PG is the space of vectors tangent to the fiber at P;
2. T_PH is invariant under action by G; and
3. the components in T_PG and T_PH of a smooth vector field in T_PE are also smooth. The spaces T_PG and T_PH are called the vertical and horizontal subspaces, respectively.

Some of the implications of this definition are most easily seen in a coordinate basis on the bundle. Let a basis for T_PH be defined as the linear combination

$$D_j = \partial_j - h_j^A D_A, \tag{86}$$

and require that D_j commute with D_A (among other consequences, this implies that h_j^A transforms under a particular representation of the Lie algebra of G, called the adjoint representation). Then the coefficients h_j^A are called connection coefficients and the basis elements $\{D_j\}$ are called the horizontal lifts or the covariant derivatives of the basis elements $\{\partial_j\}$. The dual to the basis $\{D_j, D_A\}$ for T_PE is the set $\{dx^j, \omega^A\}$, where the ω^A are given by

$$\omega^A = \Omega^A + h_j^A\, dx^j. \tag{87}$$

They form the components of a composite one-form $\omega = \omega^A D_A$ called the connection form.

The connection form ω encodes many of the interesting properties of the bundle in a concise notation. Its exterior derivative is also an important quantity in physical applications. Introduce a two-form R called the curvature form of the bundle by the definition

$$R = d\omega + \omega \wedge \omega. \tag{88}$$

The curvature is said to be a horizontal form because its action on any vertical vector vanishes. Its nonzero components are given by the expressions

$$R = R^A D_A, \quad R_{jk}^A = R^A(D_j, D_k), \tag{89}$$

and it follows that

$$[D_j, D_k] = R_{jk}^A D_A. \tag{90}$$

Applying another exterior derivative gives an identity

$$dR \equiv R \wedge \omega - \omega \wedge R = 0 \tag{91}$$

called the Bianchi identity, with components

$$\sum_{jkl} D_j R_{kl}^A = 0, \tag{92}$$

where the sum is over cyclic permutations of the indices j, k, l.

2.7 Example: Electromagnetism

An illustration of the role of some of these ideas in physics is provided by the formulation of the theory of electromagnetism in differential-geometric language. First, here is a summary of a few of the key equations of electromagnetism. In this section, a boldfaced symbol denotes a vector viewed as a collection of components. The symbol ∇ is the usual vector gradient operator, while $\cdot$ indicates the vector dot product and $\times$ represents the vector cross product.

The Maxwell equations in SI units include: Gauss's law,

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}; \tag{93}$$

Faraday's law,

$$\nabla \times \mathbf{E} + \partial_t \mathbf{B} = 0; \tag{94}$$

the equation expressing the absence of magnetic monopoles,

$$\nabla \cdot \mathbf{B} = 0; \tag{95}$$

and the Ampère-Maxwell law,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + c^{-2} \partial_t \mathbf{E}, \tag{96}$$

where ε_0 is the absolute permittivity, μ_0 is the absolute permeability, and $c = 1/\sqrt{\varepsilon_0 \mu_0}$ is the speed of light *in vacuo*. Although these equations can be solved directly in simple cases, it is often useful to introduce new variables, called potentials, in terms of which the four first-order Maxwell equations are replaced with two second-order equations. The scalar and vector potentials ϕ and $\mathbf{A}$ are defined by

$$\mathbf{E} = -\nabla \phi - \partial_t \mathbf{A}, \quad \mathbf{B} = \nabla \times \mathbf{A}. \tag{97}$$

With these definitions, the homogeneous equations (94) and (95) are automatically satisfied. The two inhomogeneous Maxwell equations become coupled second-order equations for the potentials:

$$\nabla^2 \phi + \partial_t \nabla \cdot \mathbf{A} = -\frac{\rho}{\varepsilon_0} \tag{98}$$

and

$$\nabla^2 \mathbf{A} - c^{-2} \partial_t^2 \mathbf{A} - \nabla(\nabla \cdot \mathbf{A} + c^{-1} \partial_t \phi) = -\mu_0 \mathbf{J}. \tag{99}$$

There exists a freedom in the definition of ϕ and $\mathbf{A}$. The electric field $\mathbf{E}$ and the magnetic induction $\mathbf{B}$ are unchanged by the replacements

$$\phi \to \phi' = \phi - \partial_t \Lambda \tag{100}$$

and

$$\mathbf{A} \to \mathbf{A}' = \mathbf{A} - \nabla \Lambda, \tag{101}$$

where Λ is a function of $\mathbf{x}$ and t. These replacements are called gauge transformations. Their existence provides sufficient freedom to decouple Eqs. (98) and (99).

It is easiest to approach the differential-geometric formulation of electromagnetism in stages, each incorporating more aspects of the theory. Here, the Maxwell equations for $\mathbf{E}$ and $\mathbf{B}$ are first expressed using the language of differential forms. The structure of the theory as a fiber bundle is then described, thereby incorporating the potentials ϕ and $\mathbf{A}$ and the notion of gauge transformations. To obtain consistent physical dimensionalities within expressions, it is convenient to work with a coordinate $x^0 = ct$ with dimensions of length rather than with the time coordinate t. In what follows, the spatial coordinates (x, y, z) are denoted (x^1, x^2, x^3).

Begin with the identification of the space and time dimensions as a four-dimensional smooth manifold M. The manifold is often taken to be R^4 but this is not essential. The tangent space to M at a point P is also four-dimensional, and a basis for this space is the set $\{\partial_\mu\}$, $\mu = 0, 1, 2, 3$, of derivatives with respect to the four coordinates (x^0, x^1, x^2, x^3). An arbitrary vector can be expanded with respect to this basis. One vector, denoted by j and called the four-vector current, has components j^μ formed from the charge and current densities ρ, $\mathbf{J}$:

$$j = j^\mu \partial_\mu = \left(\frac{\mu_0 \rho}{c}\right) \partial_0 + \mu_0 \mathbf{J} \cdot \nabla. \tag{102}$$

An important tensor field on M is the Minkowskian metric g, defined to have components $g_{\mu\nu}$ in a coordinate basis

forming a matrix given by

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{103}$$

This incorporates the essential elements of special relativity.

The Maxwell equations can be expressed in terms of a two-form field F defined on M. This antisymmetric tensor of type (0,2) is called the field strength. The components $F_{\mu\nu}$ of F are related to the components of the electric field and magnetic induction, and are given by

$$F_{\mu\nu} = \begin{pmatrix} 0 & +E^1/c & +E^2/c & +E^3/c \\ -E^1/c & 0 & -B^3 & +B^2 \\ -E^2/c & +B^3 & 0 & -B^1 \\ -E^3/c & -B^2 & +B^1 & 0 \end{pmatrix}. \tag{104}$$

This assignment of $\mathbf{E}$ and $\mathbf{B}$ is not *a priori* mathematically unique but establishes agreement of the resulting theory with experiment. In terms of the two-form F, the inhomogeneous Maxwell equations can be rewritten as

$$dF = j, \tag{105}$$

and the homogeneous ones become

$$d * F = 0. \tag{106}$$

The two-form $*F$ is called the dual field strength. In component form in a coordinate basis, these equations read

$$\partial_\mu F^{\mu\nu} = j^\nu \tag{107}$$

and

$$e_{\mu\nu\rho\sigma}\, \partial^\rho F^{\mu\nu} = 0. \tag{108}$$

Each of these represents four equations, and an inspection shows they reduce to the usual form of the Maxwell equations upon substitution in F and j of $\mathbf{E}$, $\mathbf{B}$, ρ, and $\mathbf{J}$.

The discussion so far has excluded the potentials ϕ and $\mathbf{A}$. These can be combined to form the components A^μ of a vector, called the gauge potential:

$$A^\mu \partial_\mu = \left(\frac{\phi}{c}\right) \partial_0 + \mathbf{A} \cdot \nabla. \tag{109}$$

The factor of c is introduced to maintain dimensional consistency. The metric g provides the associated one-form

$$A = A_\mu dx^\mu = g_{\mu\nu} A^\nu dx^\mu, \tag{110}$$

with components obtained by lowering the index. A complete description of the differential-geometric role of the gauge potential in electromagnetism requires a framework in which to place its nonuniqueness under gauge transformations. This freedom can be interpreted as a symmetry of Eqs. (98) and (99) expressing electromagnetism in terms of the potentials. It can be shown that this symmetry is a Lie group, called $U(1)$. A natural geometrical framework to express this is a fiber bundle, as is discussed next. For simplicity in what follows, the charge and current densities are taken to vanish. Nonzero distributions can be incorporated consistently with the addition of some extra structure.

The bundle of interest is a principal fiber bundle with the four-dimensional space-time manifold taken as the base manifold B and the symmetry group $U(1)$ of gauge transformations taken for the structure group G. Since the manifold of the group $U(1)$ is a circle S^1, the principal bundle is five-dimensional. Denote the coordinate on S^1 by θ. The introduction of a connection separates the tangent space to a point P in the bundle into a four-dimensional horizontal subspace spanned by the basis $\{D_\mu = \partial_\mu\}$ and a one-dimensional vertical subspace spanned by

the generator $D_\theta = \partial_\theta$ of the Lie algebra of $U(1)$. The dual basis is the set $\{dx^\mu, \Omega^\theta = d\theta\}$. The composite connection form ω is $\omega = \Omega^\theta D_\theta = d\theta\partial_\theta$.

The gauge potential A can be identified with the value of the one-form Ω^θ on a section s of the bundle. Suppose that the surface s through the bundle E is specified in a chart U by choosing the group coordinate θ as a function of the coordinates $\{x^\mu\}$ provided by U. Then the dual form becomes

$$\Omega^\theta \equiv d\theta = \partial_\mu\theta(x)dx^\mu \equiv A_\mu(x)dx^\mu, \tag{111}$$

where the identification of the components of the one-form Ω^θ with the components of the gauge-potential one-form has been made. Under a change of cross section, which is equivalent to the action of a group element with a parameter Λ, say, the potentials A_μ change by an amount $\partial_\mu\Lambda$. This provides the geometrical interpretation for the gauge transformations (100) and (101).

The curvature two-form $d\omega + \omega \wedge \omega$ derived from the connection form ω is denoted by F. Evaluated on a local section, it has components

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{112}$$

In terms of the scalar and vector potentials, this equation reproduces the definitions of Eq. (97). The Bianchi identity in component form in this case can be written

$$d * F = 0, \tag{113}$$

thereby reproducing the homogeneous Maxwell equations. To complete the specification of the bundle, additional equations are needed that explicitly determine in each section the connection and the curvature. These are called equations of motion. Requiring these to transform as usual under Lorentz transformations and to be second-order differential equations for the connection or first-order equations for the curvature significantly restricts the options. An inspection of the Lorentz representation content of the general first-order term $\partial_\lambda F_{\mu\nu}$ shows that the simplest choice is $\partial_\mu F_{\mu\nu} = 0$ or its form equivalent

$$dF = 0. \tag{114}$$

This reveals the geometrical role of the remaining two equations in Maxwell's theory.

In the presence of monopoles, the homogeneous Maxwell equations are modified by the introduction of sources. A geometrical setting for the equations describing the fields of a monopole is provided by a nontrivial principal bundle. It can be shown that the essential physics is contained in a bundle with base space S^2 and structure group $U(1)$. The bundle space E looks like S^3 and the projection map π is called the Hopf map.

2.8 Complex Manifolds

Just as the requirement of differentiability for manifolds introduces many useful structures, a further restriction imposing complex analyticity is of considerable interest. The resulting manifolds, called complex manifolds, look locally like the complex plane. Some of their main features are outlined in this section. Basic methods of complex analysis are assumed here. See ANALYTIC METHODS, Sec. 1, for more details.

The formal definition of a complex manifold M parallels that for a real differentiable manifold presented in Sec. 2.1. The key difference is that the local charts now contain

maps f taking neighborhoods U into C^n, the product of n complex planes C, and that the composition map $f_1 \circ f_2^{-1}$ is required to be holomorphic rather than differentiable. This ensures that the methods of complex analysis can be used on M independently of any choice of chart. The number n is called the complex dimension of M; the real dimension is $2n$. An important feature is that a complex manifold may have two or more incompatible atlases, i.e., the union of two atlases may not satisfy the requirements for an atlas. In this case the atlases are said to define different complex structures. An example is the torus T^2 with two real dimensions; it can be shown that the complex structures on the torus are distinguished by a complex number called the modular parameter.

Denote the n complex coordinates on M in a chart U by $z^j = x^j + iy^j, j = 1, \ldots, n$, with $\bar{z}^j = x^j - iy^j$. The tangent space $T_P M$ at a point P of M^n is spanned by a $2n$-dimensional coordinate basis $\{\partial/\partial x^j, \partial/\partial y^j\}$. It is useful to define

$$\partial_j \equiv \frac{\partial}{\partial z^j} = \frac{1}{2}\left(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\right),$$
$$\partial_{\bar{j}} \equiv \frac{\partial}{\partial \bar{z}^j} = \frac{1}{2}\left(\partial\partial x^j + i\frac{\partial}{\partial y^j}\right). \quad (115)$$

The cotangent space is spanned by the dual basis $\{dx^j, dy^j\}$, or equivalently by

$$\{dz^j = dx^j + idy^j, \quad d\bar{z}^j = dx^j - idy^j\}. \quad (116)$$

Define the linear map $J : T_P M \to T_P M$ by

$$J\partial_j = i\partial_j, \quad J\partial_{\bar{j}} = -i\partial_{\bar{j}}. \quad (117)$$

Note that $J \circ J = -I$. This map is smooth and globally defined on any complex manifold M. It is called the almost complex structure of M. The action of J separates $T_P M$ into two separate vector spaces, one spanned by vectors v such that $Jv = iv$ and the other by vectors such that $Jv = -iv$. It follows that a vector in $T_P M$ can be uniquely decomposed into two pieces, called the holomorphic and antiholomorphic parts. The cotangent space $T_P^* M$ can be separated into two corresponding pieces.

Complex differential forms of degree (p, q) can also be introduced. These are elements of a vector space denoted by $\wedge^{(p,q)} T$. In local coordinates, $\wedge^{(p,q)} T$ is spanned by a coordinate basis containing p factors of dz^j and q factors of $d\bar{z}^j$. The exterior derivative d naturally separates into the sum of two pieces,

$$d = \partial + \bar{\partial}, \quad (118)$$

called the Dolbeault operators. They satisfy

$$\partial\partial = \overline{\partial\partial} = \partial\bar{\partial} + \bar{\partial}\partial = 0. \quad (119)$$

All complex manifolds admit a Hermitian metric. A Riemannian metric g on M is said to be Hermitian if

$$g(Jv_1, Jv_2) = g(v_1, v_2) \quad (120)$$

for all vectors $v_1, v_2 \in T_P M$ at all points P. In a coordinate basis, g can be shown to have the form

$$g = g_{j\bar{k}} dz^j \wedge d\bar{z}^k + g_{\bar{j}k} d\bar{z}^j \wedge dz^k. \quad (121)$$

One can also define a two-form Ω called the Kähler form by

$$\Omega(v_1, v_2) = g(Jv_1, v_2). \quad (122)$$

If the Kähler form is closed, $d\Omega = 0$, the manifold is called a Kähler manifold and the metric g is said to be a Kähler metric. In a chart, the components of a Kähler metric can be written as

$$g_{j\bar{k}} = \partial_j \partial_{\bar{k}} K, \quad (123)$$

where K is a scalar function called the Kähler potential. Compact Kähler manifolds in one complex dimension are called Riemann surfaces and are of great importance in certain branches of physics, notably string theory. Examples of Riemann surfaces are the two-sphere S^2 and the two-torus T^2.

2.9 Global Considerations

Essentially all the differential geometry considered above has involved local concepts. It is also of interest to address the issue of the extent to which the local properties of a manifold determine its global ones. The study of global properties of a manifold forms part of the branch of mathematics called topology (*q.v.*) and as such is tangential to the scope of this article. This section provides a sketch of some connections between the two subjects. Details may be found in the references provided at the end of the article.

One link between the geometry and topology of a differentiable manifold M can be introduced by considering the space of all closed p-forms on M. This space can be separated into classes, each containing closed forms differing from one another only by exact ones. The set of all classes is a vector space called the pth de Rham cohomology group of M and denoted $H^p(M)$. This vector space contains topological information about M. For example, the dimension of H^p, called the pth Betti number, is a topological invariant of M that contains information about the holes in M. The Betti numbers also determine the number of harmonic forms on M.

There are relationships between the number of critical points of functions on a manifold M and the topology of M. This is the subject of the calculus of variations in the large, or Morse theory. Among the results obtained are the Morse inequalities, which relate the number of certain types of critical points of a function to combinations of the Betti numbers on M.

The presence of a metric on M permits other types of global information to be obtained. An example is the Hodge decomposition theorem. This can be viewed as the statement that on a compact orientable Riemannian manifold M without boundary, any p-form can be uniquely decomposed into the sum of an exact form, a coexact form, and a harmonic form.

The issue of describing the global structure of a bundle (not necessarily principal) provides another link between geometry and topology. It is possible to develop measures of the ways in which a given bundle differs from the trivial bundle. The relevant mathematical objects are called characteristic classes. They are elements of the cohomology classes of the base manifold, and are given different names depending on the type of bundle being considered. Among these are Pontrjagin, Euler, and Chern classes, corresponding to orthogonal, special orthogonal, and unitary structure groups. Elements in these classes can be expressed in terms of the curvature two-form of the bundle. Another set of characteristic classes, the Steifel-Whitney classes, determines the orientability of a manifold and whether a spinor field can be consistently defined on it.

There are also relations between certain aspects of differential operators on bundles and the topology of the bundles. These are given by index theorems. An important example is the Gauss-Bonnet theorem, which connects the number of harmonic forms on a manifold (this is a property

of the exterior derivative operator) to an integral over the Euler class (this is a topological quantity). Another important example is the Riemann-Roch theorem for complex manifolds. These are special cases of the Atiyah-Singer index theorem.

2.10 Further Examples

Many sets of smooth physical variables can be viewed as differentiable manifolds, and so differential-geometric concepts such as vectors, tensors, forms, and bundles play key roles in much of theoretical physics. Examples can be found in every major branch of physics. For instance, the modern formulation of the Hamiltonian dynamics of a system proceeds via the investigation of a manifold M called the phase space, with local coordinates corresponding to the generalized coordinates and momenta of the system. A closed nondegenerate two-form called the symplectic form is defined on M, making the phase space a symplectic manifold. The study of the properties of the phase space using the methods of differential geometry provides information about the behavior of the system. An extension of this example occurs in quantum mechanics. Quantization of a system involves the introduction of complex structure on the symplectic manifold. The study of this procedure is called geometric quantization.

Differential geometry is particularly crucial in the development of theories of fundamental interactions and particles. The geometrical constructions presented above for electromagnetism can readily be extended to other theories of fundamental forces. For example, the equations believed to describe the underlying physics of the strong interactions form a theory called chromodynamics. This theory can be expressed geometrically using a principal bundle over space-time but where the structure group is the eight-dimensional Lie group called SU(3) rather than U(1). The presence of a multidimensional group manifold with a nontrivial group composition law means that, unlike the electrodynamic case, the horizontal lifts are inequivalent to the basis for the tangent space to the base manifold. As a result, the structure of the Bianchi identities and the equations of motion are somewhat more complicated in detail. The essential construction, however, remains the same.

Another important physical theory is general relativity, which provides a good description of the gravitational interactions at the classical level. This theory can also be given a geometrical interpretation as a fiber bundle, but it is of a somewhat different kind, called a bundle of frames. Each point on a fiber of this bundle consists of a choice of basis vectors for the tangent space to the space-time manifold, and the symmetry group that plays the role of the structure group of a principal bundle now acts to rotate these bases into one another. A connection form and an associated curvature still exist, and they are closely related to the Christoffel symbols and the Riemann space-time curvature tensor of general relativity. In addition, there exists new freedom arising from the choice of basis vector on the base manifold, which leads to the existence of a second natural one-form on the bundle called the solder form or vierbein. This also has an associated two-form, called the torsion. In Einstein's general relativity the torsion form is specified to be zero, although other possibilities can be envisaged.

Attempts to unify the known fundamental forces and particles make wide

use of geometrical constructions. Examples of such theories in four dimensions are the grand unified theories, describing the strong, weak, and electromagnetic forces in a single framework. The geometrical structures discussed above can be extended to more complicated symmetry groups large enough so that the connection forms include all the force fields needed for these theories. Certain elementary particles play the role of sources for these fields and can also be incorporated in bundles called associated bundles. Many unified theories involve higher-dimensional manifolds, in which physical space-time is a submanifold. These include the so-called Kaluza-Klein theories. Often, the symmetries of the extra dimensions permit them to play the role of the structure group in a principal bundle.

Generalizations of the geometrical framework of gravitation are also possible. For example, if the base manifold for a bundle of frames is generalized in a certain way, it is possible to specify bundles describing extensions of general relativity that include fundamental particles and forces other than gravity and that incorporate enlarged symmetries called supersymmetries. The resulting theories are called supergravities.

String theories are candidate unified theories including gravity that are believed to be consistent with quantum mechanics. In these theories, the fundamental forces and particles are interpreted as objects that are extended in one dimension (hence the name string). As a string propagates in space-time, it sweeps out a two-dimensional surface called the world sheet. A description of the world sheet involves the study of complex manifolds, in particular Riemann surfaces, as well as the notions of global differential geometry.

3
Projective Geometry

In its basic form, projective geometry is essentially the theory of perspective, i.e., the study of those features of geometrical objects that remain the same when the objects are projected from a point onto a line or plane. The elements of projective geometry are implicitly used by artistic painters, designers, and other people who represent three-dimensional objects on a two-dimensional medium. In its generalized form, the subject is fundamental in axiomatic geometry. It can be viewed as subsuming the classical Euclidean and non-Euclidean geometries.

There are two approaches to projective geometry. Synthetic projective geometry seeks to develop the subject as a series of deductions starting from certain axioms, in the Euclidean tradition. Analytical projective geometry introduces homogeneous coordinates and uses analytical techniques to obtain results. The two approaches are complementary, although projective geometries exist for which coordinates cannot be introduced.

A key feature of projective geometry is that parallel lines are assumed to meet in a single point, called the point at infinity, and that parallel planes meet in a single line, called the line at infinity. One advantage of these assumptions is that geometrical statements do not require exceptions for parallelism. For example, it is now true that any two lines in the plane determine a point, and any two planes in three dimensions determine a line.

In a plane, the statement that two lines determine a point is strikingly similar to the statement that two points determine a line. In general, projective-geometric statements involving points and lines in the plane remain valid when the roles of

the points and lines are interchanged. In the plane, points are said to be dual to lines. In three dimensions the notion of duality applies between points and planes, or between lines and lines. A similar concept exists in higher dimensions.

With these ideas, a set of axioms for synthetic projective geometry can be formulated in terms of three basic notions: point, line, and incidence. The latter is meant in the sense of intersection: for example, a point is incident to a line if it lies on the line. The axioms can be expressed in dual pairs, so that propositions deduced necessarily have valid duals.

3.1 Some Theorems

There are several theorems that play a central role both in the development of the basic theory and in its extension to more abstract situations. A key result is Desargues's theorem: Given six distinct points in two sets, $\{A_1, A_2, A_3\}$ and $\{B_1, B_2, B_3\}$ (i.e., the vertices of two triangles), if the lines A_1B_1, A_2B_2, A_3B_3 meet at a point, then the three points C_1, C_2, C_3 given respectively by the pairwise line intersections A_1B_2 and A_2B_1, A_2B_3 and A_3B_2, A_3B_1 and A_1B_3 are collinear. This theorem holds in all projective geometries in three dimensions or more and in certain two-dimensional cases, including the usual plane projective geometry. However, in two dimensions non-Desarguesian geometries also exist.

Another important result that holds for a large class of projective geometries including the usual plane and solid ones is Pappus's theorem: Given two lines a and b lying in a plane and two sets of three distinct points $\{A_1, A_2, A_3\}$ incident to a and $\{B_1, B_2, B_3\}$ incident to b, then the three points C_1, C_2, C_3 given respectively by the pairwise line intersections A_1B_2 and A_2B_1, A_2B_3 and A_3B_2, A_3B_1 and A_1B_3 are collinear. Non-Pappian geometries also exist.

A pencil of lines about a point P is defined as the set of all lines lying in a plane and incident with P. A line s in the plane not incident with P is called a section of the pencil, and the pencil is said to project the section from P. Two pencils can be projectively related through a common section. Two distinct sections are said to be related by a projective transformation from the point P. The fundamental theorem of projective geometry states that a projective transformation is specified when three collinear points and their images are given. The theorem generalizes to projective transformations of higher-dimensional figures.

Conic sections (see Sec. 1.2) have a natural construction in projective geometry, and their theory can be developed entirely within this subject. Since all conics can be generated by projection of a circle from a point onto a plane, the projective approach gives them a unified treatment and consequently several results of analytical geometry can follow from a single projective theorem. Plane-projective definitions also play an important role. For example, the locus of intersections of corresponding lines in two projectively related pencils is a conic. A well-known result in this branch of the subject is Pascal's theorem: Given six points $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ incident to a conic, then the three points B_1, B_2, B_3 given respectively by the pairwise line intersections A_1A_2 and A_4A_5, A_2A_3 and A_5A_6, A_3A_4 and A_6A_1 are collinear. The dual to Pascal's theorem is sometimes called Brianchon's theorem. These methods of projective geometry can also be extended to the study of quadrics and higher-dimensional hypersurfaces.

3.2 Homogeneous Coordinates

In analytical projective geometry, a set of coordinates called homogeneous coordinates is introduced. Consider first homogeneous coordinates on the line. A Cartesian coordinate system assigns a single real number x to each point P. In contrast, a homogeneous coordinate system assigns *two* real numbers (x_0, x_1) to each point, where $x = x_1/x_0$ and at least one of (x_0, x_1) is nonzero. Evidently, the homogeneous coordinates (x_0, x_1) and (cx_0, cx_1), where c is a constant, both represent P. The advantage of homogeneous coordinates is that the point (0,1) at infinity is treated on the same footing as, say, the origin (1,0). It also makes any polynomial equation $f(x) = 0$ homogeneous in (x_0, x_1) without affecting the degree of the equation.

In the plane, the homogeneous coordinates of a point P specified in Cartesian coordinates by (x, y) are three real numbers (x_0, x_1, x_2), not all zero, for which $x = x_1/x_0$, $y = x_2/x_0$. A line in Cartesian coordinates is given by the linear equation $Ax + By + C = 0$. In homogeneous coordinates this becomes the homogeneous linear equation

$$Ax_1 + Bx_2 + Cx_0 = 0. \tag{124}$$

The line at infinity has equation $x_0 = 0$ and is thereby treated on a similar footing to other lines; for example, the x and y coordinate axes have equations $x_2 = 0$ and $x_1 = 0$, respectively. All these ideas generalize to higher dimensions.

In addition to providing a framework in which analytical calculations can be developed, the homogeneous coordinate system offers a simple setting for duality. For example, given Eq. (124), the three numbers (A, B, C) can be viewed as homogeneous coordinates for a line in the plane. Then, coordinate statements about a point are expressed in terms of three numbers (x_0, x_1, x_2), while statements about a line are expressed in terms of a dual set of three numbers (A, B, C). A single equation thus represents a line or a point depending on which three numbers are considered variables.

Any set of three coordinates $(\widehat{x}_0, \widehat{x}_1, \widehat{x}_2)$, obtained from the homogeneous coordinate system (x_0, x_1, x_2) in the plane by an invertible linear transformation

$$\widehat{x}_j = A_{jk} x_k \tag{125}$$

(see ALGEBRAIC METHODS, Sec. 3.3), also leaves unchanged the degree of any polynomial function of the coordinates. The set $(\widehat{x}_0, \widehat{x}_1, \widehat{x}_2)$ can be taken as alternative homogeneous coordinates.

3.3 Group of Projective Transformations

Instead of being taken as a change of coordinates for a fixed point P, the linear transformation (125) can be interpreted as a mapping from a point P at (x_0, x_1, x_2) to another point $\widehat{P}$ at $(\widehat{x}_0, \widehat{x}_1, \widehat{x}_2)$. This provides a mapping of the projective plane onto itself. Such mappings form a group G called the group of projective transformations for the plane. Similarly, groups of projective transformations can be introduced for higher-dimensional cases.

According to the so-called erlangen program, projective geometry can be viewed as the study of properties of figures that are invariant under the action of G. Various other geometries can be obtained by requiring invariance under a subgroup of G. They include the regular Euclidean geometry, as well as affine geometry and

the non-Euclidean elliptic and hyperbolic geometries.

Extensions of projective geometry to fields other than the real numbers exist. For example, complex projective geometry is defined over the complex numbers. The field may be finite or even noncommutative (see ALGEBRAIC METHODS, Sec. 2). For example, a finite geometry in the plane called $PG(2,5)$ can be constructed using 31 points and 31 lines, with six points on each line and six lines through each point. Details of these generalized projective geometries may be found in the references at the end of this article.

4 Algebraic Geometry

Algebraic geometry involves the study of mathematical objects called varieties, which are generalized curves, surfaces, and hypersurfaces. The subject has several levels of abstraction, in each of which the precise meaning of the word variety is different. For the purposes of this article a relatively simple level of sophistication suffices, in which a variety can roughly be viewed as the solution to a set of polynomial equations for variables in a space. Note, however, that the modern definition of variety is considerably more abstract. It uses a branch of mathematics called the theory of schemes, about which more can be found in the references at the end of this article.

This section presents a few simple notions of algebraic geometry in the framework of polynomial equations. The discussion refers to several concepts (e.g., field, polynomial ring, rational functions) that are defined and described in the article ALGEBRAIC METHODS.

4.1 Affine Varieties

Here is a more precise definition of one important type of variety. Consider an algebraically closed field F. An n-dimensional affine space A^n over F is defined as the set of points specified by the coordinates $(f_1, \ldots, f_n)$ with $f_j \in F$. Denote by $F[f_1, \ldots, f_n]$ the polynomial ring in n variables over F. An affine variety V is a subset of A^n given by the common zeros of a set S of polynomials in $F[f_1, \ldots, f_n]$. If S contains only one polynomial, V is called an affine curve for $n = 2$, an affine surface for $n = 3$, and an affine hypersurface for $n > 3$.

A subset of V satisfying the definition of a variety is called a subvariety. If V is the union of two subvarieties, it is called reducible; otherwise, it is irreducible. For example, an irreducible affine curve is one for which the defining polynomial is irreducible (i.e., cannot be factored). An irreducible component of V is defined as a maximal irreducible subvariety of V. One result in this subject is that any variety V can be written uniquely as the union of finitely many distinct irreducible components.

Starting with a variety V, a sequence of irreducible varieties can be constructed such that each member of the sequence is a subvariety of the preceding one. This sequence is of finite length, and the number of subvarieties in it is called the dimension of V.

The unions and finite intersections of subvarieties of V are also subvarieties. This means that the complements of the subvarieties of V can be used as the collection t of subsets for a topological space (see Sec. 2.1). Therefore, A^n and hence also V can be endowed with a topology, called the Zariski topology. This

topology is not Hausdorff but, unlike the usual Hausdorff topology on C^n, it is defined for all affine varieties over F.

4.2 Projective Varieties

Several extensions of the notion of affine variety to more general varieties exist. One generalization uses an approach similar to that taken in the construction of differentiable manifolds: The meaning of variety is extended to include objects constructed by patching together affine varieties. This generalization then looks locally like an affine variety but globally is different. An important result in algebraic geometry is that certain subsets of projective spaces form varieties of this sort, called projective varieties.

An n-dimensional projective space P^n over F can be introduced as the set of points specified by the homogeneous coordinates $(f_0, f_1, \ldots, f_n)$ with $f_j \in F$ not all zero, subject to the restriction that two such sets of homogeneous coordinates related via a single nonzero constant $c \in F$ as

$$(f_0, f_1, \ldots, f_n) = (cf_0, cf_1, \ldots, cf_n) \quad (126)$$

specify the same point (cf. Sec. 3.2). Denote by $H[f_1, \ldots, f_n]$ the ring of homogeneous polynomials in n variables over F. A projective variety V is a subset of P^n given by the common zeros of a set S of polynomials in $H[f_1, \ldots, f_n]$. If S contains only one polynomial, V is called a projective curve for $n = 2$, a projective surface for $n = 3$, and a projective hypersurface for $n > 3$.

4.3 Classification

The ultimate aims of algebraic geometry are the classification and characterization of varieties. These are difficult and unsolved problems in the generic case. To attack the classification problem, a means of relating varieties to one another is needed. This is provided by the notion of a rational map.

A rational map $f : V \to A^n$ from an affine variety to n-dimensional affine space is basically a set of n rational functions f_j. The domain of f is by definition taken as the union of the domains of the n functions f_j. A rational map $f : V_1 \to V_2$ between two affine varieties $V_1 \subset A^{n_1}$ and $V_2 \subset A^{n_2}$ is defined to be a rational map $f : V_1 \to A^{n_1}$ such that the range of f lies in V_2. If the map f also has a rational inverse, it is called a birational equivalence.

The classification problem is approached by seeking a classification up to birational equivalence. Ideally, this means providing discrete and/or continuous numerical quantities that are invariant under birational equivalence and that characterize inequivalent varieties. Then, given a birationally equivalent set of varieties, a standard subset with desirable features (e.g., no singularities) can be sought and a classification attempted. Finally, one can seek some means of measuring the deviation from this standard subset of the remaining members of the birational-equivalence class.

An example is provided by the special case of the algebraic curves over F. For these varieties, a discrete quantity called the genus g can be introduced, which is a nonnegative real number that is invariant under birational equivalence. Curves with $g = 1$ are sometimes called elliptic curves. For each nonzero g the birational-equivalence classes can be labeled by a one-dimensional continuous variable if $g = 1$ and by a $(3g - 3)$-dimensional set of continuous variables otherwise. The

continuous variables are called moduli. They also form an irreducible variety, called moduli space, that can in turn be studied with the methods of algebraic geometry.

If the field F is the field C of complex numbers, the resulting algebraic curves are the Riemann surfaces. A curve with $g = 0$ is topologically a sphere, while one with $g = 1$ is topologically a torus. The set of Riemann surfaces plays an important role in string theories (see Sec. 2.10). For example, at a particular order in perturbation theory for a scattering process the string world sheet is topologically a Riemann surface with punctures for the incoming and outgoing strings. The methods of differential and algebraic geometry play a significant role in the evaluation of such contributions to the scattering amplitudes.

Glossary

Considerations of space prevent an extensive glossary being provided for this article. Instead, the following is a list incorporating important concepts together with the number of the section in which the concept appears.

Abscissa: See Sec. 1.1.
Adjoint Exterior Derivative: See Sec. 2.4.
Affine Space: See Sec. 4.1.
Affine Variety: See Sec. 4.1.
Almost Complex Structure: See Sec. 2.8.
Antisymmetric Symbol: See Sec. 2.4.
Antisymmetric Tensor: See Sec. 2.3.
Applicate: See Sec. 1.5.
Atlas: See Sec. 2.1.
Base Manifold: See Sec. 2.5.
Betti Number: See Sec. 2.9.
Bianchi Identity: See Sec. 2.6.
Birational Equivalence: See Sec. 4.3.
Brianchon's Theorem: See Sec. 3.1.
Bundle of Frames: See Sec. 2.10.
Cartan Exterior Algebra: See Sec. 2.4.
Cartesian Coordinates: See Sec. 1.1.
Characteristic Class: See Sec. 2.9.
Chart: See Sec. 2.1.
Circle: See Sec. 1.2.
Closed Form: See Sec. 2.4.
Coclosed Form: See Sec. 2.4.
Coexact Form: See Sec. 2.4.
Complex Manifold: See Sec. 2.8.
Complex Structure: See Sec. 2.8.
Cone: See Sec. 1.5.
Conic Section: See Sec. 1.2.
Connection: See Sec. 2.6.
Contraction: See Sec. 2.2.
Contravariant Components: See Sec. 2.2.
Coordinate Basis: See Sec. 2.2.
Cosine: See Sec. 1.3.
Cotangent Space: See Sec. 2.2.
Covariant Components: See Sec. 2.2.
Covariant Derivative: See Sec. 2.6.
Cross Section: See Sec. 2.5.
Curvature Form: See Sec. 2.6.
Curvilinear Coordinates: See Sec. 1.4.
Cylinder: See Sec. 1.5.
Cylindrical Coordinates: See Sec. 1.5.
De Rham Cohomology: See Sec. 2.9.
Desargues's Theorem: See Sec. 3.1.
Differential Forms: See Sec. 2.4.
Directrix: See Sec. 1.2.
Discriminant: See Sec. 1.2.
Dolbeault Operator: See Sec. 2.8.
Dual Basis: See Sec. 2.2.
Dual Vector Space: See Sec. 2.2.
Duality, Projective: See Sec. 3.
Eccentricity: See Sec. 1.2.
Einstein Summation Convention: See Sec. 2.2.
Ellipse: See Sec. 1.2.
Ellipsoid: See Sec. 1.5.
Embedding: See Sec. 2.3.
Erlangen Program: See Sec. 3.3.
Euclidean Space: See Sec. 1.5.

Exact Form: See Sec. 2.4.
Exterior Derivative: See Sec. 2.4.
Fiber: See Sec. 2.5.
Fiber Bundle: See Sec. 2.5.
Finite Geometry: See Sec. 3.3.
Focus: See Sec. 1.2.
Fundamental Theorem of Projective Geometry: See Sec. 3.1.
Genus: See Sec. 4.3.
Geometric Quantization: See Sec. 2.10.
Group of Projective Transformations: See Sec. 3.3.
Harmonic Form: See Sec. 2.4.
Hausdorff: See Sec. 2.1.
Hermitian Metric: See Sec. 2.8.
Hodge Decomposition Theorem: See Sec. 2.9.
Hodge Star: See Sec. 2.4.
Homogeneous Coordinates: See Sec. 3.2.
Hopf Map: See Sec. 2.7.
Horizontal Lift: See Sec. 2.6.
Horizontal Subspace: See Sec. 2.6.
Hyperbola: See Sec. 1.2.
Hyperboloid: See Sec. 1.5.
Hypersphere: See Sec. 1.5.
Hypersurface: See Sec. 1.5.
Incidence: See Sec. 3.
Index Theorem: See Sec. 2.9.
Irreducible Variety: See Sec. 4.1.
Kähler Metric: See Sec. 2.8.
Kepler Problem: See Sec. 1.6.
Lie Algebra: See Sec. 2.5.
Lie Group: See Sec. 2.5.
Line At Infinity: See Sec. 3.
Manifold: See Sec. 2.1.
Maxwell Equations: See Sec. 2.7.
Method of Coordinates: See Sec. 1.
Metric: See Sec. 2.3.
Modular Parameter: See Sec. 2.8.
Moduli: See Sec. 4.3.
Morse Theory: See Sec. 2.9.
Neighborhood: See Sec. 2.1.
Ordinate: See Sec. 1.1.
Pappus's Theorem: See Sec. 3.1.
Parabola: See Sec. 1.2.
Paraboloid: See Sec. 1.5.
Parametric Representation: See Sec. 1.5.
Pascal's Theorem: See Sec. 3.1.
Pencil: See Sec. 3.1.
Plane Analytic Geometry: See Sec. 1.5.
Plane Polar Coordinates: See Sec. 1.4.
Point at Infinity: See Sec. 3.
Potentials: See Sec. 2.7.
Principal Bundle: See Sec. 2.5.
Projective Geometry: See Sec. 3.
Projection Map: See Sec. 2.5.
Projective Transformation: See Sec. 3.1.
Projective Variety: See Sec. 4.2.
Quadric: See Sec. 1.5.
Riemann Surface: See Secs. 2.8, 2.10, 4.3.
Riemannian Manifold: See Sec. 2.3.
Scalar Field: See Sec. 2.3.
Sine: See Sec. 1.3.
Slope: See Sec. 1.1.
Solid Analytic Geometry: See Sec. 1.5.
Sphere: See Sec. 1.5.
Spherical Polar Coordinates: See Sec. 1.5.
Spherical Triangle: See Sec. 1.5.
Structure Group: See Sec. 2.5.
Submanifold: See Sec. 2.3.
Symmetric Tensor: See Sec. 2.3.
Symplectic Form: See Sec. 2.10.
Tangent: See Sec. 1.3.
Tangent Space: See Sec. 2.2.
Tensor: See Sec. 2.3.
Tensor Field: See Sec. 2.3.
Topological Space: See Sec. 2.1.
Variety: See Sec. 4.
Vector: See Sec. 2.2.
Vector Field: See Sec. 2.3.
Vertical Subspace: See Sec. 2.6.
Wedge Product: See Sec. 2.4.
Zariski Topology: See Sec. 4.1.

Further Reading

Abraham, R., Marsden, J. (1985), *Foundations of Mechanics*, Reading, MA: Addison-Wesley.

Borsuk, K. (1969), *Multidimensional Analytic Geometry*, Warsaw: PWN–Polish Scientific.
Coxeter, H. (1974), *Projective Geometry*, Toronto: Univ. of Toronto Press.
Eisenhart, L. (1909), *Differential Geometry*, New York: Atheneum.
Eisenhart, L. (1926), *Riemannian Geometry*, Princeton, NJ: Princeton Univ. Press.
Flanders, H. (1963), *Differential Forms*, New York: Academic.
Frampton, P. (1987), *Gauge Field Theories*, Reading, MA: Benjamin Cummings.
Fulton, W. (1974), *Algebraic Curves*, New York: Benjamin.
Goldstein, H. (1980), *Classical Mechanics*, Reading, MA: Addison-Wesley.
Goodman, A. W. (1963), *Analytic Geometry and the Calculus*, New York: Macmillan.
Green, M., Schwarz, J., Witten, E. (1987), Superstring Theory, 2 Vols., Cambridge, U.K.: Cambridge Univ. Press.
Hartshorne, R. (1977), *Algebraic Geometry*, Berlin: Springer-Verlag.
Kobayashi, S., Nomizu, K. (1965), *Foundations of Differential Geometry*, New York: Wiley.
Milnor, J., Stasheff, J. (1974), *Characteristic Classes*, Princeton, NJ: Princeton Univ. Press.
Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists*, New York: Academic.
Ramond, P. (1981), *Field Theory*, Reading, MA: Benjamin Cummings.
Schutz, B. (1980), *Geometrical Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge Univ. Press.
Semple, J., Roth, L. (1985), *Algebraic Geometry*, Oxford: Clarendon.
Spivak, M. (1979), Differential Geometry, Vols. 1–5, Wilmington: Publish or Perish.
Springer, G. (1981), *Riemann Surfaces*, New York: Chelsea.
Steenrod, N. (1974), *The Topology of Fibre Bundles*, Princeton, NJ: Princeton Univ. Press.
Struik, D. (1948), *A Concise History of Mathematics*, New York: Dover.
Veblen, O., Young, J. (1938), Projective Geometry, Vols. I and II, New York: Blaisdell.
Winger, R. (1962), *Projective Geometry*, New York: Dover.
Woodhouse, N. (1992), *Geometric Quantization*, Oxford: Oxford Univ. Press.
Zariski, O. (1971), *Algebraic Surfaces*, Berlin: Springer-Verlag.

Green's Functions

Kazuo Ohtaka
Laboratory of Applied Physics, Faculty of Engineering, Chiba University, Chiba-shi, Japan

Introduction ***160***
1 History of Green's Functions ***162***
2 Construction of Green's Functions ***164***
2.1 One-Dimensional Equation of Sturm-Liouville Type with Dirichlet-Type Boundary Conditions *164*
2.2 Retarded, Advanced, and Causal Green's Functions of the Helmholtz Equation *165*
2.3 Green's Functions Obtained by Fourier Transform *166*
2.3.1 Heat Equation *166*
2.3.2 Time-Dependent Schrödinger Equation *167*
2.3.3 Klein-Gordon Equation *167*
2.4 Green's Functions Matching Homogeneous Boundary Conditions at the Boundary of a Finite Region *168*
2.5 Spectral Representation of Green's Functions *168*
3 Green's Functions used in Solving Initial- and Boundary-Value Problems ***169***
3.1 Dirichlet and Neumann Problems of Poisson's Equation *169*
3.2 Initial- and Boundary-Value Problem for the Heat Equation *170*
4 Boundary-Element Method ***171***
4.1 Practical Boundary-Value Problems *171*
4.2 Poisson's Equation as Treated by the Boundary-Element Method *171*
4.3 Applications of the Boundary-Element Method *172*
4.3.1 Fluid Mechanics *172*
4.3.2 Sound and Electromagnetic Waves *172*
4.3.3 Elasticity *172*
5 Green's Functions Having a Direct Relevance to Physical Reality ***173***
5.1 Wave Front of Radiation Emitted from a Point Source and Huygens's Principle *173*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

5.2 Retarded Green's Function of Schrödinger's Equation *174*
5.3 Dislocations *174*
5.4 Magnetic Field around a Vortex Line in a Type-II Superconductor *175*
6 Perturbational Treatment to Obtain Green's Functions *175*
6.1 Slater-Koster Model *176*
6.2 Scattering Cross Section of a Plane Wave from a Scatterer with Spherical Symmetry *177*
6.3 Band Structure of Electrons and Photons and Diffraction Experiments *178*
7 Green's Functions in Many-Body Theories *179*
7.1 Single- and Two-Particle Green's Functions *179*
7.2 Wick's Theorem and Feynman Diagrams *181*
7.3 Temperature Green's Functions *181*
7.4 Linear Response Theory of Kubo and Temperature Green's Functions *182*
7.5 Green's Functions in Quantum Field Theories *183*
Glossary *185*
List of Works Cited *186*
Further Reading *187*

Introduction

In mathematics the term Green's function is usually given to a solution of an initial- or boundary-value problem of a differential equation with a δ-function inhomogeneous term. Let us be more specific. Consider an ordinary or partial differential equation

$$L_x G(x, x') = \delta(x - x'), \qquad (1)$$

with L_x a linear differential operator with respect to the variable x. Here, x may stand for either the position $\mathbf{r}$, the time t, or the pair $(\mathbf{r},t)$. Then the solution $G(x, x')$ is called the Green's function if it satisfies a given homogeneous boundary condition – a condition relating the value of G to its derivative G_x on the boundary of the domain, a simple example of which is $G(x, x') = 0$ or $G_x(x, x') = 0$. In physics and applied physics, however, the term Green's function is often used without explicitly referring to the boundary condition. For example, a fundamental solution in mathematics is often called simply a Green's function. Thus it is desired to seek an alternative and more generalized definition that enables us to deal with a wider class of functions encountered in various areas of physical science. This can be achieved by formulating the concept of Green's functions in terms of response theory: The Green's function is then a response function that connects the output signal $O(x)$ of the system with the input signal $I(x)$ in the form of a linear integral transform:

$$O(x) = \int G(x, x') I(x')\, dx', \qquad (2)$$

the integral range over x' depending upon the problem under consideration. In the present article the term Green's function is employed in this generalized sense.

When the linear response of a system is described by a linear operator L_x, which

may be differential, integral, integro-differential, or of any other kind, the two signals $O(x)$ and $I(x)$ are related through

$$L_x O(x) = I(x). \tag{3}$$

Comparing this equation with Eq. (2), we see that the Green's function is formally defined by L_x^{-1}. When, in particular, L_x is a differential operator in Eq. (3) and a homogeneous boundary condition, $u(x) = 0$, for example, is imposed on the boundary Γ, the definition used in mathematics is recovered. This is because the superposition over x' in Eq. (2) solves the problem when G satisfies Eq. (1) with the boundary condition $G(x, x') = 0$ for x on Γ. There are, however, many cases where it is difficult to specify the operator L_x for describing the response. The relationship between the responses of a black box to a δ-function type and distributed input signals is shown in Fig. 1.

Although the principle of superposition and hence the validity of the form given by Eq. (2) hold only when the solution satisfies a homogenous boundary condition, Green's functions are also central when one tries to construct a solution of a boundary-value problem with an inhomogeneous boundary condition – for example, a solution having a prescribed nonzero value on the boundary. This is one reason why Green's functions are so widely used.

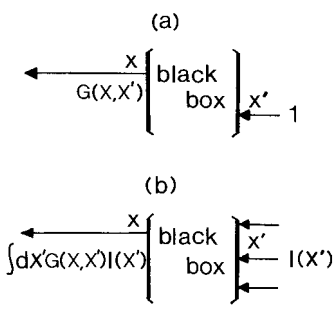


Fig. 1 Response at x of a black box to an input signal: (a) an input signal localized at x' with unit strength, (b) an input signal distributed with magnitude $I(x')$

The quantities called "resolvent", "resolvent kernel", "signal function", "point response function", or "transfer function", encountered in various fields of mathematics, physics, applied physics, and engineering, are nothing but the Green's functions in the generalized definition. We note that in Eq. (1) the Green's function $G(x, x')$ describes the response to a "point" input source and in Eq. (2) it "transfers" the input signal into the output response of the system in question. When one recalls that many problems in physics and applied physics ultimately reduce to finding the output $O(x)$ for a given input $I(x)$, one can understand why Green's functions are very popular today in many fields – hydrodynamics, electrodynamics, acoustics, elasticity, quantum mechanics, solid-state physics, elementary-particle physics, and so on. To imagine how widely they are used, it is enough to remember the diverse names given to them, listed above. Their usefulness is still growing progressively today, as various numerical techniques continue to develop for calculations involving Green's functions.

The subjects of the present article are the definitions, significances, constructions, utilizations, and usefulness of Green's functions. We try to make the description as illustrative as possible. The Green's functions we deal with in this article range from those treated in mathematical textbooks to the ones used in many fields of pure and applied physics.

Although, unless stated otherwise, the concrete forms of the Green's functions will be given for the case of three-dimensional space, the reader should keep in mind that they depend intrinsically on

the dimensionality of the space considered. The reader can consult the monographs quoted in the text for the case of other dimensionalities. Also, it should be noted that in defining a Green's function in this article the coefficient of a δ function of the point input signal is not always unity, reflecting some arbitrariness existing in the literature.

The present article is organized as follows. We sketch in Sec. 1 a brief history of Green's functions including their modern development. In Sec. 2 some typical methods of constructing Green's functions are explained for several differential equations. The usefulness of Green's functions in initial- and boundary-value problems is demonstrated in Sec. 3. The boundary-element method, devised to handle boundary-value problems for a nontrivial geometry, is explained in Sec. 4, together with the presentation of some of its applications. Up to this point, the description is given for the role of Green's functions as a convenient tool for solving a mathematical or physical problem. In Sec. 5, a number of Green's functions are given which have a direct relevance with a physical reality. The treatments as combined with perturbation method are described in Sec. 6. The Green's functions popular in many-body problems are described in Sec. 7, where, among other things, we review their extremely important application in linear response theory of condensed-matter physics. A brief sketch of their use in quantum field theory is also given.

1 History of Green's Functions

In the history of Green's functions, it will be appropriate to go back to 1828, when N. Green put forth Green's formula (Kellog, 1939)

$$\int_\Omega [u(\mathbf{r})\Delta v(\mathbf{r}) - v(\mathbf{r})\Delta u(\mathbf{r})]d^3\mathbf{r} = \int_\Gamma \left(u(\mathbf{r})\frac{\partial}{\partial n}v(\mathbf{r}) - v(\mathbf{r})\frac{\partial}{\partial n}u(\mathbf{r}) \right) ds. \tag{4}$$

It converts the volume integral within a region Ω of the left-hand side into the surface integral on its boundary Γ, with Δ the Laplacian and $(\partial/\partial n)v(\mathbf{r}) = \widehat{\mathbf{n}} \cdot \nabla v(\mathbf{r})$, $\widehat{\mathbf{n}}$ being the outward normal with unit length to the boundary Γ. This formula holds for arbitrary $u(\mathbf{r})$ and $v(\mathbf{r})$. When, in particular, $u(\mathbf{r})$ is a harmonic function, $\Delta u(\mathbf{r}) = 0$, and $v(\mathbf{r})$ is the Green's function of the Laplace equation

$$\Delta v(\mathbf{r}) = -\delta(\mathbf{r} - \mathbf{r}') \tag{5}$$

or

$$v(\mathbf{r}) = \frac{1}{(4\pi|\mathbf{r} - \mathbf{r}'|)}, \tag{6}$$

Green's formula yields for $\mathbf{r}$ within Ω

$$u(\mathbf{r}) = \frac{1}{4\pi}\int_\Gamma \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|}\frac{\partial}{\partial n'}u(\mathbf{r}') - u(\mathbf{r}')\frac{\partial}{\partial n'}\frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) ds'. \tag{7}$$

It shows that, as in Cauchy's formula for regular functions of complex argument $z = x + iy$, we can express a harmonic function inside a region as an integral over its boundary: we may evaluate the value of $u(\mathbf{r})$ inside Ω only if we know values of both $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ on Γ. If we let the argument $\mathbf{r}$ approach the boundary Γ, Eq. (7) becomes a Fredholm integral equation, which allows us to express $u(\mathbf{r})$ in terms of $\partial u(\mathbf{r})/\partial n$. Likewise, differentiating Eq. (7) over $\mathbf{r}$ and letting $\mathbf{r}$ tend to Γ lead to the equation for $\partial u(\mathbf{r})/\partial n$. Thus $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ cannot both be

specified freely on Γ. From this fact stem the Dirichlet and Neumann problems of the Laplace equation, i.e., the problems of finding a harmonic function $u(\mathbf{r})$ that has a prescribed boundary value $u(\mathbf{r})$ and outward derivative $\partial u(\mathbf{r})/\partial n$, respectively, on the boundary Γ.

In the latter half of the 19th century, Green's functions played a fundamental role in the discussion of the existence and uniqueness of the solution of internal or external boundary-value problems (Courant and Hilbert, 1937). In addition to their significance in these basic problems, they were also the key quantities in the practical side of constructing the solutions of boundary-value problems, as analyzed fully by Lyapunov for the potential problem (Smirnov, 1965).

Towards the end of the 19th century and at the beginning of the 20th century, Green's functions were used in examining the completeness property (closure property) of the set of eigenfunctions of self-adjoint operators and in proving the expansion theorem for an arbitrary function in terms of the complete set obtained from, say, a Sturm-Liouville operator (Courant and Hilbert, 1937). In mathematics, these concepts opened the way to functional analysis, which has since refined and generalized greatly the theory of partial differential equations (Yosida, 1965). In the fields of pure and applied physics, not only were they adopted in solving various practical problems, but they were also used in clarifying many fundamental concepts underlying quantum mechanics, which was founded in 1925 and has been developing ever since. Indeed, one of the easiest ways to recognize a marked peculiarity of quantum mechanics contrasting with classical mechanics is to change the action integral for a classical motion into the form involving the Green's function of the Schrödinger equation (Sakurai, 1985). Also, in many quantum-mechanical applications, the Green's functions enable us to take into account a perturbation series to an infinite order to give a deep insight not attainable by a finite-order treatment.

As elementary-particle physics and solid-state physics began to develop rapidly after World War II, the extended applications of Green's functions were actively pursued. One example is seen in many-particle physics in which the Green's functions are defined in terms of field operators and used in conjunction with the graphical representation of many-body processes (see, e.g., Feynman, 1972). In these graphs, which have came to be called Feynman diagrams, each line standing for a Green's function describes temporal and spatial evolution of an elementary particle or excitation. The crossing or branching of the lines represents the interaction among particles, implying that the Green's function carries all the information on the "personal history" of an electron, proton, photon, etc. Because of this characteristic, these Green's functions are more often called "propagators." Their contributions in the development of quantum electrodynamics (Bogoliubov and Shirkov, 1959) and solid-state physics (Abrikosov et al., 1963) have been quite remarkable. Despite the apparent difference in definition, Green's functions defined in terms of Feynman's path integral in quantum field theory constitute the second example belonging to this category (Feynman and Hibbs, 1965; Itzykson and Zuber, 1980). Since the functional integral seems to be the most powerful tool to date to quantize nonlinear Lagrangians, the Green's functions will continue to be a useful tool in the future development of this field. As a last example, we refer the reader to linear response theory applied widely in

condensed-matter physics. In this example. too, Green's functions have been very useful in that the theory of Kubo is most powerful when it is applied in conjunction with the temperature Green's functions introduced by Matsubara (Kubo et al., 1991).

Parallel to such generalizations in pure physics, the Green's functions of traditional usage have been refined in various ways, yielding many important concepts. Especially, many practical problems related to the Laplace, heat, wave, or Schrödinger equation, previously left untouched simply because the boundaries of the domains in question were too complex for analytical treatment, have come to be solved with the help of numerical techniques. The boundary-element method, one of the methods designed for just such problems, takes full advantage of the Green's-function approach (Brebbia, 1978; Brebbia and Walker, 1980). Green's functions are now so widely used everywhere that familiarity with them is becoming more and more important.

2 Construction of Green's Functions

A number of typical methods of constructing Green's functions are illustrated.

2.1 One-Dimensional Equation of Sturm-Liouville Type with Dirichlet-Type Boundary Conditions

The Green's function for the Sturm-Liouville operator satisfies $[p(x) > 0]$

$$L[G(x,x')] \equiv \frac{d}{dx}\left(p(x)\frac{d}{dx}G(x,x')\right) - q(x)G(x,x') = -\delta(x-x'), \tag{8}$$

where the one-dimensional region $0 < x < 1$ is assumed. Suppose that a Dirichlet-type boundary condition is imposed on G:

$$G(x,x') = 0 \quad \text{at } x = 0 \text{ and } 1. \tag{9}$$

The solution of this problem is constructed as follows:

$$G(x,x') = \begin{cases} \gamma u_<(x)u_>(x'), & 0 \le x \le x' \le 1, \\ \gamma u_>(x)u_<(x'), & 0 \le x' \le x \le 1. \end{cases} \tag{10}$$

Here $u_<(x)$ is a solution of $L[u_<(x)] = 0$ with the boundary value $u_<(0) = 0$ at the left boundary, while $u_>(x)$ satisfies $L[u_>(x)] = 0$ with $u_>(1) = 0$ at the right boundary. The constant γ in Eq. (10), independent of x and x', is given by

$$\gamma = \frac{1/p(x')}{u'_<(x')u_>(x') - u'_>(x')u_<(x')}, \tag{11}$$

with $u'_<(x') = \left(\frac{d}{dx'}\right)u_<(x')$. It is determined such that

$$p(x')[G'(x'^+,x') - G'(x'^-,x')] = -1, \tag{12}$$

with $G'(x'^{\pm},x') = (d/dx)G(x,x')|_{x=x'^{\pm}}$, which is the condition obtained by integrating both sides of Eq. (8) in the infinitesimal interval $x'^- < x < x'^+$, where $x'^{\pm} = x' \pm \varepsilon(\varepsilon \to 0+)$. When the left-hand solution $u_<(x)$ happens to satisfy simultaneously the condition $u_<(1) = 0$, i.e., when $u_<(x)$ happens to be the true eigenfunction of the operator L with zero eigenvalue, the constant γ diverges, meaning that the system resonates with the point external force expressed by the δ function in Eq. (8). Still, in this case, one can redefine a generalized Green's function so that Eq. (2) remains valid in taking into account the inhomogeneous term (Smirnov, 1965). In the case of a Neumann- or mixed-type homogeneous

boundary condition in place of Eq. (9), Eq. (10) still provides us with the Green's function if $u_<(x)$ and $u_>(x)$ therein satisfy the given boundary conditions. For various types of Sturm-Liouville operators and their generalizations, Green's functions are tabulated in many books. See, e.g., Butkovskiy (1982).

2.2 Retarded, Advanced, and Causal Green's Functions of the Helmholtz Equation

The Green's function of the Helmholtz equation is defined by

$$(\Delta + \kappa^2)G(\mathbf{r}, \mathbf{r}'; \kappa) = -\delta(\mathbf{r} - \mathbf{r}'). \quad (13)$$

By Fourier transform we find

$$G(\mathbf{r}, \mathbf{r}'; \kappa) = \int \frac{d^3k}{(2\pi)^3} \frac{\exp[i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')]}{k^2 - \kappa^2}. \quad (14)$$

As such the Green's function is undetermined because the integral is divergent. If we add a small imaginary part $i\varepsilon$ to $\kappa (\varepsilon \to 0_+)$, one may construct the following three types of Green's functions:

1. By putting κ equal to $\kappa + i\varepsilon$, one obtains

$$G_R(\mathbf{r}, \mathbf{r}'; \kappa) = \frac{\exp(i\kappa|\mathbf{r} - \mathbf{r}'|)}{4\pi|\mathbf{r} - \mathbf{r}'|}, \quad (15)$$

which is called the retarded Green's function. It is regular in the upper half of the complex κ plane.

2. By putting κ equal to $\kappa - i\varepsilon$, one obtains a complex conjugate of the retarded function,

$$G_A(\mathbf{r}, \mathbf{r}'; \kappa) = \frac{\exp(-i\kappa|\mathbf{r} - \mathbf{r}'|)}{4\pi|\mathbf{r} - \mathbf{r}'|}. \quad (16)$$

This Green's function is called the advanced Green's function and is regular in the lower half of the complex κ plane.

3. By putting κ equal to $\kappa + i\varepsilon$ sgn κ, sgn κ being $\kappa/|\kappa|$, (i.e., κ^2 to $\kappa^2 + i\varepsilon$), one obtains the causal Green's function

$$G_C(\mathbf{r}, \mathbf{r}'; \kappa) = G_R(\mathbf{r}, \mathbf{r}'; \kappa)\theta(\kappa) + G_A(\mathbf{r}, \mathbf{r}'; \kappa)\theta(-\kappa), \quad (17)$$

the Heaviside step function $\theta(\kappa)$ being defined by

$$\theta(\kappa) = \begin{cases} 1, & \kappa > 0, \\ 0, & \kappa < 0. \end{cases} \quad (18)$$

We see that $G_R(G_A)$ is obtained by analytically continuing G_C of the range $\kappa > 0\,(\kappa < 0)$ to the upper (lower) half of the complex κ plane. Except in many-body theories, G_C is seldom used [see Eq. (88)]. The names "retarded" and "advanced" come from the time dependence of the Green's functions of the (time-dependent) wave equation

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta\right) G(\mathbf{r}, t; \mathbf{r}', t') = \delta(\mathbf{r} - \mathbf{r}')\delta(t - t'), \quad (19)$$

c being a positive constant. Upon converting from t to the Fourier space ω, we obtain the Helmholtz equation given by Eq. (13) with $\kappa^2 = \omega^2/c^2$. The inverse Fourier transform back to the variables t and t' of Eq. (14) [or Eq. (15) or (16)] shows that $G_R(\mathbf{r}, t; \mathbf{r}', t')$ has nonzero values only in the case $t > t'$, while G_A is finite in the opposite case, $t < t'$. Namely, when G_R or G_A is substituted for G in the input-output relation (2), it turns out that

$$O(\mathbf{r}, t) = \int_{-\infty}^{t} dt' \times \int d^3\mathbf{r}' G_R(\mathbf{r}, t; \mathbf{r}', t') I(r', t'),$$

$$O(\mathbf{r}, t) = \int_t^\infty dt' \times \int d^3r' G_A(\mathbf{r}, t; \mathbf{r}', t') I(\mathbf{r}', t'). \tag{20}$$

That is, the retarded Green's function duly satisfies the causality condition in the sense that a perturbation due to $I(\mathbf{r}', t')$ precedes its consequence observed at time t. The advanced Green's function describes the time-reversed process of the physical one. These features are also obvious in the time-independent version $G(\mathbf{r}, \mathbf{r}'; \kappa)$, which describes the scattering of, e.g., a monochromatic sound wave with a fixed frequency ω ($\kappa = \omega/c$, c being the velocity of sound). Here G_R given by Eq. (15) leads properly to the outgoing scattered wave, while G_A gives rise to an incoming scattered wave (Sommerfeld, 1949). To summarize, the three Green's functions of wave equation are defined by

$$G_R(\mathbf{r}, t; \mathbf{r}', t') = \lim_{\varepsilon \to 0} \int \frac{d^3k\, d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}') - i\omega(t - t')]}{k^2 - (\omega + i\varepsilon)^2/c^2},$$

$$G_A(\mathbf{r}, t; \mathbf{r}', t') = \lim_{\varepsilon \to 0} \int \frac{d^3k\, d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}') - i\omega(t - t')]}{k^2 - (\omega - i\varepsilon)^2/c^2},$$

$$G_C(\mathbf{r}, t; \mathbf{r}', t') = \lim_{\varepsilon \to 0} \int \frac{d^3k\, d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}') - i\omega(t - t')]}{k^2 - \omega^2/c^2 - i\varepsilon}. \tag{21}$$

After evaluating the integrals, we find

$$G_R(\mathbf{r}, t; \mathbf{r}', t') = \left(\frac{c}{2\pi}\right) \theta(t - t') \times \delta(c^2(t - t')^2 - |\mathbf{r} - \mathbf{r}'|^2),$$

$$G_A(\mathbf{r}, t; \mathbf{r}', t') = \left(\frac{c}{2\pi}\right) \theta(t' - t) \times \delta(c^2(t - t')^2 - |\mathbf{r} - \mathbf{r}'|^2),$$

$$G_C(\mathbf{r}, t; \mathbf{r}', t') = \frac{ic/(2\pi)^2}{|\mathbf{r} - \mathbf{r}|^2 - c^2(t - t')^2 + i\varepsilon}. \tag{22}$$

The first of the three leads to Huygens's principle (See Sec. 5.1).

2.3 Green's Functions Obtained by Fourier Transform

As shown in Sec. 2.2, the Fourier transform is a convenient way to obtain Green's functions. It is powerful only for obtaining the Green's function for an infinite domain, however, i.e., a fundamental solution. Nevertheless, it should be noted that such a Green's function enables us to derive a Green's function subject to a homogeneous boundary condition on the boundary of a finite domain (see Sec. 2.4). Note also that the infinite-domain Green's functions are used very often to solve the problems for a finite domain [the boundary-element method is one of the examples (see Secs. 4.2 and 4.3)]. With this remark in mind, we will in this section give some Green's functions obtained by Fourier transform.

2.3.1 Heat Equation

The Green's function of the heat equation is defined by

$$\left(\frac{\partial}{\partial t} - \sigma^2 \Delta\right) G(\mathbf{r}, t; \mathbf{r}', t') = \delta(\mathbf{r} - r')\delta(t - t'). \tag{23}$$

In Fourier space, it holds that

$$G(\mathbf{r}, t; \mathbf{r}', t') = \int \frac{d^3k\, d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}') - i\omega(t-t')]}{\sigma^2 k^2 - i\omega} = \begin{cases} [4\pi\sigma^2(t-t')]^{-3/2} \times \exp\left\{\dfrac{-|\mathbf{r}-\mathbf{r}'|^2}{4\sigma^2(t-t')}\right\}, & t > t', \\ 0, & t < t', \end{cases} \tag{24}$$

the second relation being obtained by calculating the residue in the ω integral. We should note that the integral over ω is well defined in contrast to the wave equation treated in Sec. 2.2. The finiteness of the Green's function only in the case $t > t'$ is in accord with the law of increase of entropy or the second law of thermodynamics.

2.3.2 Time-Dependent Schrödinger Equation

The Green's function for a free particle with mass m obeying Schrödinger's equation is defined by

$$\left(i\hbar\frac{\partial}{\partial t} + \frac{\hbar^2}{2m}\Delta\right) G(\mathbf{r}, t; \mathbf{r}', t') = i\hbar\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'). \tag{25}$$

Fourier transform then yields

$$G_R(\mathbf{r}, t; \mathbf{r}', t') = i\int \frac{d^3k\, d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}') - i\omega(t-t')]}{\omega - \hbar k^2/2m + i\varepsilon} = \begin{cases} \left[\dfrac{m}{2\pi i\hbar(t-t')}\right]^{3/2} \times \exp\left[\dfrac{im|\mathbf{r}-\mathbf{r}'|^2}{2\hbar(t-t')}\right], & t > t', \\ 0, & t < t'. \end{cases} \tag{26}$$

In the ω integral, we have replaced ω by $\omega + i\varepsilon$ to obtain the retarded Green's function. The advanced Green's function, finite for $t < t'$, is obtained by putting ω equal to $\omega - i\varepsilon$. Thus the Schrödinger equation allows a time-reversed solution, like what we have seen for the wave equation [Eq. (19)].

2.3.3 Klein-Gordon Equation

The Green's functions of the Klein-Gordon equation are defined by

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta + \mu^2\right) G(\mathbf{r}, t; \mathbf{r}', t') = \delta(\mathbf{r}-\mathbf{r}')\delta(t-t'), \tag{27}$$

c and μ being two positive constants. Replacing μ^2 by $-\mu^2$ defines the Green's function of the telegraphic equation. By Fourier transform we find

$$G(\mathbf{r}, t; \mathbf{r}', t') = \int \frac{d^3k d\omega}{(2\pi)^4} \times \frac{\exp[i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}') - i\omega(t-t')]}{k^2 + \mu^2 - (\omega/c)^2}. \tag{28}$$

Since the integral is not well defined, we can construct three Green's functions, G_R, G_A, and G_C, by replacing ω as in Eq. (21). They are obtained, respectively, as (Bogoliubov and Shirkov, 1959)

$$\begin{aligned} G_R(x) &= -\theta(t-t')\Delta(x;\mu^2), \\ G_A(x) &= \theta(t'-t)\Delta(x;\mu^2), \\ G_C(x) &= \overline{\Delta}(x;\mu^2) + \left(\frac{i}{2}\right)\Delta^{(1)}(x;\mu^2). \end{aligned} \tag{29}$$

Here $G(\mathbf{r}, t; \mathbf{r}', t')$ is expressed simply as $G(x)$, x standing for $(c(t-t'), \mathbf{r}-\mathbf{r}')$, and $\theta(t-t')$ is the step function defined by Eq. (18). The other quantities are defined by

$$\Delta(x;\mu^2) = \left(\frac{-c}{2\pi}\right)\mathrm{sgn}(t-t')$$
$$\times\left[\delta(x^2) - \theta(x^2)\left(\frac{\mu^2}{2}\right)J_1\frac{(\mu x)}{(\mu x)}\right],$$
$$\overline{\Delta}(x;\mu^2) = -\frac{1}{2}\mathrm{sgn}(t-t')\Delta(x;\mu^2),$$
$$\Delta^{(1)}(x;\mu^2) = \left(\frac{c}{4\pi|\mathbf{r}-\mathbf{r}'|}\right)\left(\frac{\partial}{\partial r}\right)$$
$$\times[\theta(x^2)N_0(\mu x) - i\theta(-x^2)$$
$$\times H_0(i\mu(-x^2)^{1/2})], \qquad (30)$$

where $\mathrm{sgn}(t) = t/|t|$, $x^2 = c^2(t-t')^2 - |\mathbf{r}-\mathbf{r}'|^2$, $x = (x^2)^{1/2}$, J_1 is the first-order Bessel function, and N_0 and H_0 are the Neumann and first-kind Hankel functions of the zeroth order, respectively (Abramowitz and Stegun, 1965). In quantum field theory, the Green's functions in Fourier space [$(\mathbf{k},\omega)$ representation] are more often used than the $(\mathbf{r},t)$ representation given above. The Green's functions treated in Sec. 2.2 are reproduced by taking the limit $\mu \to 0$ in the above.

2.4 Green's Functions Matching Homogeneous Boundary Conditions at the Boundary of a Finite Region

As an illustration, let us consider the Green's function of the Laplace equation for a region Ω:

$$\Delta G(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}'), \qquad (31)$$

with a homogeneous Dirichlet condition imposed on the boundary Γ:

$$G(\mathbf{r},\mathbf{r}')|_{\mathbf{r}\text{ on }\Gamma} = 0. \qquad (32)$$

Were it not for the restriction (32), the Green's function would be nothing more than the fundamental solution of the Laplace equation, the Coulomb potential given by Eq. (6). It satisfies the boundary condition $G(\mathbf{r},\mathbf{r}') \to 0$, as $r \to \infty$. To match the boundary condition (32), the Green's function must have the form

$$G(\mathbf{r},\mathbf{r}') = g(\mathbf{r},\mathbf{r}') + \frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|}, \qquad (33)$$

where $g(\mathbf{r},\mathbf{r}')$ is the solution of the homogeneous equation

$$\Delta g(\mathbf{r},\mathbf{r}') = 0 \qquad (34)$$

subject to the boundary condition

$$g(\mathbf{r},\mathbf{r}')|_{\mathbf{r}\text{ on }\Gamma} = \left.\frac{-1}{4\pi|\mathbf{r}-\mathbf{r}'|}\right|_{\mathbf{r}\text{ on }\Gamma}. \qquad (35)$$

The second term of Eq. (33) takes account of the δ function of the Poisson equation, while the first term $g(\mathbf{r},\mathbf{r}')$ incorporates the boundary condition. The problem of finding the Green's function thus reduces to an orthodox Dirichlet problem of Eqs. (34) and (35) of finding a harmonic function $g(\mathbf{r},\mathbf{r}')$ satisfying the inhomogeneous boundary condition. Although the existence and uniqueness of the solution is well established, a concrete expression for $g(\mathbf{r},\mathbf{r}')$ is hard to obtain analytically, unless the boundary Γ has a good symmetry (Kellog, 1939; Smirnov, 1965). Nevertheless, this method of obtaining the Green's functions satisfying the boundary condition has a wide applicability in many differential equations, and is not restricted to the Laplace equation treated here.

2.5 Spectral Representation of Green's Functions

As an example of Green's functions of three-dimensional self-adjoint operators, we treat here that of the Schrödinger equation defined by

$$[E - H(\mathbf{r})]G(\mathbf{r},\mathbf{r}';E) = i\hbar\delta(\mathbf{r}-\mathbf{r}'). \qquad (36)$$

For an electron in a hydrogen atom, for example, the Hamiltonian H is given by

$$H(\mathbf{r}) = -\frac{\hbar^2}{2m}\Delta - \frac{e^2}{4\pi\varepsilon_0 r}, \tag{37}$$

the proton being taken as the origin of coordinates. From Eq. (36) one obtains

$$G(\mathbf{r}, \mathbf{r}'; E) = \sum_n \frac{i\hbar\psi_n(\mathbf{r})\psi_n(\mathbf{r}')^*}{E - E_n}, \tag{38}$$

with the eigenfunction ψ_n satisfying

$$H(\mathbf{r})\psi_n(\mathbf{r}) = E_n\psi_n(\mathbf{r}). \tag{39}$$

Equation (38) can easily be verified by applying $E - H(\mathbf{r})$ to both sides and using the completeness of $\{\psi_n\}$ for the self-adjoint operator $H(\mathbf{r})$. The set of states n includes not only the states with discrete energy eigenvalues but also the states within continuous spectra, if there are any, as in the case of a hydrogen atom. To describe a physical process occurring in a hydrogen atom, we must use the retarded version of the Green's function obtained by changing E to $E + i\varepsilon$ in Eq. (38), in accordance with the remark made in 2.2.

Even in a finite-domain problem subject to a homogeneous restriction imposed on the boundary, the expression (38) remains valid, with the understanding that the ψ_n's are now the solutions for the eigenvalue problem with that boundary condition.

3 Green's Functions used in Solving Initial- and Boundary-Value Problems

In solving an initial- or boundary-value problem, the Green's function is useful in taking account of not only an inhomogeneous term but also an inhomogeneous initial or boundary condition. Some examples are given in Kellog (1939) and Morse and Feshbach (1953).

3.1 Dirichlet and Neumann Problems of Poisson's Equation

The internal Dirichlet problem of the Poisson equation is defined by

$$\Delta u(\mathbf{r}) = -f(\mathbf{r}), \tag{40}$$

with the inhomogeneous boundary condition imposed on Γ,

$$u(\mathbf{r})|_\Gamma = g(\mathbf{r}). \tag{41}$$

The solution $u(\mathbf{r})$ of this problem may be written down if we know the Green's function $G_1(\mathbf{r}, \mathbf{r}')$ of the Laplace equation satisfying the boundary condition $G_1(\mathbf{r}, \mathbf{r}')|_\Gamma = 0$ [i.e., the Green's function given by Eq. (33)]. It reads

$$u(\mathbf{r}) = \int_\Omega G_1(\mathbf{r}, \mathbf{r}')f(\mathbf{r}')d^3r' - \int_\Gamma \left(\frac{\partial}{\partial n'}G_1(\mathbf{r}, \mathbf{r}')\right) g(\mathbf{r}')\, ds', \tag{42}$$

where n defines the outward normal to Γ as in Eq. (4). The right-hand side is being written solely in terms of the given boundary value $g(\mathbf{r})$. The reader can easily convince himself that this formula is correct by noting that the first term satisfies the Poisson equation with the boundary value $u(\mathbf{r})|_\Gamma = 0$, while the second is the solution of the Laplace equation with $u(\mathbf{r})|_\Gamma = g(\mathbf{r})$, as can be checked by the use of the Green's formula (4) with $u = u(\mathbf{r})$ and $v = G_1(\mathbf{r}, \mathbf{r}')$.

Let $G_2(\mathbf{r}, \mathbf{r}')$ be the Green's function satisfying the homogeneous Neumann-type condition, $(\partial/\partial n)G_2(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r}$ on Γ. If we employ G_2 in place of G_1 and

replace $-\partial G_1/\partial n'$ by G_2, the expression given by Eq. (42) gives the solution of the Poisson equation subject to the inhomogeneous Neumann-type boundary condition, $(\partial/\partial n)u(\mathbf{r})|_\Gamma = g(\mathbf{r})$, in place of Eq. (41). The external problems are treated analogously.

3.2 Initial- and Boundary-Value Problem for the Heat Equation

The Green's formula (4) is generalized to an arbitrary second-order differential operator L – to that of the heat equation, $L = \partial/\partial t - \sigma^2\Delta$, for example. By using it, we can express the solution for, say, the following problem of the heat equation for $\mathbf{r} \in \Omega$ and $t > 0$:

$$\left(\frac{\partial}{\partial t} - \sigma^2\Delta\right) u(\mathbf{r}, t) = f(\mathbf{r}, t), \tag{43}$$

with the initial temperature distribution specified by

$$u(\mathbf{r}, 0) = g(\mathbf{r}) \tag{44}$$

and the boundary condition of Dirichlet type given by

$$u(\mathbf{r}, t)|_\Gamma = h(\mathbf{r}, t). \tag{45}$$

Suppose we already happen to know the Green's function $H_1(\mathbf{r}, t; \mathbf{r}', t')$ for the operator $M(= -\partial/\partial t - \sigma^2\Delta)$, the adjoint of the operator L:

$$\left(-\frac{\partial}{\partial t} - \sigma^2\Delta\right) H_1(\mathbf{r}, t; \mathbf{r}', t') = \delta(\mathbf{r} - \mathbf{r}')\delta(t - t'), \tag{46}$$

which satisfies the homogeneous boundary condition

$$H_1(\mathbf{r}, t; \mathbf{r}', t')|_{\mathbf{r}\text{ on }\Gamma} = 0. \tag{47}$$

Then it is shown that the solution of the problem (43)–(45) is given by

$$\begin{aligned} u(\mathbf{r}, t) = &\int_0^t dt' \int_\Omega d^3r' H_1(\mathbf{r}', t'; \mathbf{r}, t) f(\mathbf{r}', t') \\ &+ \int_\Omega d^3r' H_1(\mathbf{r}', 0; \mathbf{r}, t) g(\mathbf{r}') \\ &- \sigma^2 \int_0^t dt' \int_\Gamma ds' \frac{\partial H_1(\mathbf{r}', t', \mathbf{r}, t)}{\partial n'} \\ &\times h(\mathbf{r}', t'). \end{aligned} \tag{48}$$

If we employ, in place of H_1, another Green's function H_2 satisfying the boundary condition $[\partial H_2(\mathbf{r}, \mathbf{r}')/\partial n]|_\Gamma = 0$ instead of Eq. (47), we obtain the solution of Eqs. (43) and (44) with, in place of Eq. (45), the Neumann-type inhomogeneous boundary condition, $(\partial/\partial n) \times u(\mathbf{r})|_\Gamma = h(\mathbf{r}, t)$. We should note that, as in Sec. 3.1, $-\partial H_1/\partial n'$ in Eq. (48) must be replaced by H_2.

The present examples given for operators that are not self-adjoint will suffice to illustrate the practical value of Green's functions in a rather wide class of boundary-value problems. An important point is that the Green's function used in the input-output relation (2) is not defined by a δ-function inhomogeneous term for the operator L but by the one for its adjoint operator M. Also, note that the reciprocity relation of Green's functions is in general established between the Green's functions G and H for the operator L and its adjoint M. Namely, for the Green's function G and H for the operators L and M, respectively, it holds that

$$G_i(\mathbf{r}, t; \mathbf{r}', t') = H_i(\mathbf{r}', t'; \mathbf{r}, t), \tag{49}$$

$i = 1$ and 2 corresponding to the Dirichlet and Neumann boundary conditions, respectively. If the operator L is self-adjoint, with the relation $L = M$, the Green's function H is automatically identical to G, leading to the well-known reciprocity

relation:

$$G_i(\mathbf{r}, t; \mathbf{r}', t') = G_i(\mathbf{r}', t'; \mathbf{r}, t). \qquad (50)$$

The operator L for the Laplace, wave, Klein-Gordon, or Schrödinger equation is self-adjoint but that for the heat equation is not (Courant and Hilbert, 1937).

4 Boundary-Element Method

4.1 Practical Boundary-Value Problems

In actual situations, we often encounter a complex boundary Γ. If we insist on applying the formulas given in Sec. 3, we will be forced to solve additional boundary-value problems in order to find Green's functions G_1, H_1, etc., as the example in Sec. 2.4 shows. Therefore these formulas are not very helpful in such problems, and more direct methods, taking full advantage of numerical techniques, are more often employed in practice. The difference method and finite-element method are two such popular examples. The boundary-element method, developed and applied widely in recent years, also belongs to this class. In contrast to the former two, which have nothing to do with Green's functions, this method is related deeply to Green's formula and hence Green's functions. Conceptually, it is a revival of the old method of expressing the solution of the Laplace equation in the form of the potential caused by a monopole or dipole layer on the boundary Γ, the unknown density of which is determined by solving the Fredholm integral equation (Courant and Hilbert, 1937). The Green's functions involved are the fundamental solution (6) (in the case of the Laplace equation), instead of the complicated Green's functions G_1, etc. Let us briefly see the characteristic points of the boundary-element method (BEM) through the following examples (Brebbia and Walker, 1980).

4.2 Poisson's Equation as Treated by the Boundary-Element Method

We return to the Poisson equation treated in Sec. 3.1. If the fundamental solution of the Laplace equation [Eq. (6)] is substituted for $\nu(\mathbf{r})$ in Green's formula (4), we can express the solution for the Poisson equation by using the boundary values of both $u(\mathbf{r})$ and $\partial u(\mathbf{r})/\partial n$ on Γ. The result is the extension of Eq. (7) to the Poisson equation. For $\mathbf{r} \in \Omega$, we find

$$u(\mathbf{r}) = \int_\Omega G(\mathbf{r}, \mathbf{r}') f(\mathbf{r}') d^3\mathbf{r}' + \int_\Gamma \left(G(\mathbf{r}, \mathbf{r}') \frac{\partial u(\mathbf{r}')}{\partial n'} - \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} u(\mathbf{r}') \right) ds' \qquad (51)$$

with $G(\mathbf{r}, \mathbf{r}') = 1/4\pi|\mathbf{r} - \mathbf{r}'|$. Since we know the value of $u(\mathbf{r})$ on Γ through the Dirichlet condition (41), $u(\mathbf{r}) = g(\mathbf{r})$, this formula provides us with the solution of the original problem, if we somehow find the value of $\partial u(\mathbf{r})/\partial n$ on Γ on the right-hand side. The procedure characterizing the BEM is that the unknown quantity $\partial u/\partial n$ is determined from Eq. (51) by letting $\mathbf{r}$ tend to a point on Γ and setting $u(\mathbf{r}) = g(\mathbf{r})$. The result is a Fredholm integral equation of the first kind for the unknown function $\partial u(\mathbf{r})/\partial n$ on Γ:

$$\frac{1}{2} g(\mathbf{r}) - \int_\Omega G(\mathbf{r}, \mathbf{r}') f(\mathbf{r}') d^3\mathbf{r}' + \int_\Gamma \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} g(\mathbf{r}')\, ds' = \int_\Gamma G(\mathbf{r}, \mathbf{r}') \frac{\partial u(\mathbf{r}')}{\partial n'}\, ds'. \qquad (52)$$

Here the factor $\frac{1}{2}$ takes account of the discontinuous nature of $\partial G/\partial n$ in letting $\mathbf{r}$ tend to the boundary. What Eq. (52) shows is that we need to make a numerical calculation of the unknown quantity $[\partial u(\mathbf{r})/\partial n]|_\Gamma$. This can be carried out readily by discretizing the integral on Γ using the well-established algorithm of the finite-element method. Following similar steps, the internal Neumann problem and the external problems are eventually reduced to an integral equation on the boundary Γ as in this example.

4.3 Applications of the Boundary-Element Method

4.3.1 Fluid Mechanics

In fluid mechanics, this method has been known as the surface-singularity method. For an incompressible and irrotational fluid, it is well known that the velocity potential satisfies the Laplace equation. Hence Eq. (52), with the inhomogeneous term $f(\mathbf{r})$ dropped, is the key equation in analyzing various boundary problems for perfect fluids. In practical problems, such as the analyses for the air flow around an aircraft or a space shuttle flying with relatively low velocity, a distribution of vortices must often be taken into account. In such cases the final integral equation like Eq. (52) needs to be modified, but the BEM is still quite powerful. See for details the report by Morino et al. (1975).

4.3.2 Sound and Electromagnetic Waves

If the retarded Green's function $\boldsymbol{G}_R$ for the Helmholtz equation [Eq. (15)] is used in place of G, Eq. (52) turns out to be the key equation for the wave equation [here $f(\mathbf{r})$ therein is an inhomogeneous term of the wave equation]. Then if we let $u(\mathbf{r})$ stand for the velocity potential associated with a sound wave, the boundary values for $u(\mathbf{r})$ and $\partial \boldsymbol{u}(\mathbf{r})/\partial \boldsymbol{n}$ will be related, respectively, to the pressure and the velocity on Γ. For a region bounded by a rigid wall, it holds that $[\partial \boldsymbol{u}(\mathbf{r})/\partial n]|_\Gamma = 0$. When $\boldsymbol{f}(\mathbf{r}) = 0$, Eq. (52) becomes a homogeneous integral equation for $\boldsymbol{g}(\mathbf{r})$, yielding the eigen-frequencies for the sound modes established in that region, which can be, in an actual problem, an auditorium or a complicated resonator such as that of a violin. By converting the integral equation to the linear coupled equations, we can find the eigenvalues ω [involved in the Green's function through $\kappa = \omega/c$ in Eq. (15)].

An external problem may be formulated similarly to deal with sound propagation from a source with a complicated shape. Needless to say, the BEM for the wave equation is not limited to acoustics.

4.3.3 Elasticity

The final example we give on the application of the BEM is the problem of determining the strain tensor of an elastic body caused by a body force $\mathbf{f}(\mathbf{r})$ and a surface force $\mathbf{p}(\mathbf{r})$, both applied externally. Since the basic equation of elasticity is rather complicated, an analytical treatment is possible only for the exceptional case of very simple $\mathbf{f}(\mathbf{r})$ and $\mathbf{p}(\mathbf{r})$, applied to an elastic body whose shape, too, is very simple. In the BEM, these restrictions may be largely relaxed. First we need the Green's function for an infinite elastic body. For an isotropic and homogeneous system the tensor of the Green's functions satisfies (Landau and Lifshitz, 1986)

$$\Delta G_{ij}(\mathbf{r},\mathbf{r}') + \frac{1}{1-2\sigma}\sum_k \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_k} G_{kj}(\mathbf{r},\mathbf{r}') = -\frac{2(1+\sigma)}{E}\delta_{ij}\delta(\mathbf{r}-\mathbf{r}') \qquad (53)$$

with Young's modulus E, Poisson's ratio σ, and Kronecker's delta δ_{ij}. Here G_{ij} is the ith component of the deformation field at $\mathbf{r}$ induced by the j-directed point force at $\mathbf{r}'$ with unit strength. The solution to Eq. (53) is known as Kelvin's solution, which reads

$$G_{ij}(\mathbf{r},\mathbf{r}') = \frac{1+\sigma}{8\pi E(1-\sigma)}[(3-4\sigma)\delta_{ij} + n_i n_j] \times \frac{1}{|\mathbf{r}-\mathbf{r}'|}, \tag{54}$$

where $\mathbf{n} = (\mathbf{r}-\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$. In terms of G_{ij} one may obtain the basic integral equation on the external surface Γ, involving the body force $\mathbf{f}(\mathbf{r})$, an analog of Eq. (52) for the strain tensor $u_{ij}(\mathbf{r})$ (Brebbia and Walker, 1980). The boundary value for $\Sigma_j \partial u_{ij}/\partial x_j$ on Γ may be related to the given surface force $\mathbf{p}(\mathbf{r})$.

5 Green's Functions Having a Direct Relevance to Physical Reality

The Green's functions treated in Secs. 3 and 4 were used mainly as a tool for solving a partial differential equation. The reader will recall that by definition they describe an output signal in response to a point input signal. This suggests that they are also usually related to a physical reality. That this is indeed so will be seen through the examples presented below.

5.1 Wave Front of Radiation Emitted from a Point Source and Huygens's Principle

The retarded Green's function given by Eq. (22) for the wave equation shows where the wave front of the radiation is found at time t, when it is emitted at a former time t' from the point source located at $\mathbf{r}'$. If we consider conversely a point of observation fixed at $\mathbf{r}$ and point sources distributed around that point, we can regard the radiation observed there at time t as a composition of the propagating wavelets that leave at $t = 0$ the various point sources, whose distance from point $\mathbf{r}$ is ct (Fig. 2). In fact, this situation is well expressed by the solution of the following initial-value problem for the wave equation in three-dimensional space:

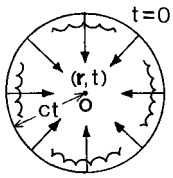


Fig. 2 Huygens's principle for wave propagation. The wave observed at the observation point $O(\mathbf{r},t)$ is the sum of the wavelets leaving, at $t = 0$, the sources distributed on the sphere with radius ct, c being the velocity of the wave

$$\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \Delta\right) u(\mathbf{r},t) = h(\mathbf{r},t) \tag{55}$$

with the initial conditions

$$u(\mathbf{r},0) = f(\mathbf{r}), \qquad u_t(\mathbf{r},0) = g(\mathbf{r}). \tag{56}$$

The solution is given by Kirchhoff's formula (Baker and Copson, 1950; Morse and Feshbach, 1953)

$$u(\mathbf{r},t) = \int_{|\mathbf{r}-\mathbf{r}'|<ct} d^3r' \frac{h(\mathbf{r}', t - |\mathbf{r}-\mathbf{r}'|/c)}{4\pi|\mathbf{r}-\mathbf{r}'|} + \frac{\partial}{\partial t}[tf(\mathbf{r})^{ct}] + tg(\mathbf{r})^{ct}. \tag{57}$$

Here the first term takes into account the inhomogeneous term of Eq. (55) in the form of a retarded potential. The quantities f^{ct} and g^{ct} are the averages of their values on the surface of the sphere centered at $\mathbf{r}$

with radius ct: for example,

$$f(\mathbf{r})^{ct} = \int \frac{d\Omega f(\mathbf{r} + ct\hat{\mathbf{r}}_0)}{4\pi}, \tag{58}$$

the integral being over the solid angle in the direction $\widehat{\mathbf{r}}_0$ from the point $\mathbf{r}$. Equation (57) demonstrates in mathematical form the well-known Huygens's principle. Although the detailed derivation is omitted, the important role of the Green's function in this principle will be understood if the constraint imposed by Green's functions in Eq. (22) is compared with those involved in the integrals in Eqs. (57) and (58). The name "propagators" given very often to the Green's functions of the wave equation or its generalizations stems from the concepts involved in Huygens's principle.

5.2 Retarded Green's Function of Schrödinger's Equation

The inverse Fourier transform of the retarded version of the Green's function given by Eq. (38) is

$$G_R(\mathbf{r}, t; \mathbf{r}', t') = \begin{cases} \Sigma_n \psi_n(\mathbf{r}) \psi_n^*(\mathbf{r}') \\ \times \exp\left[-\left(\frac{i}{\hbar}\right) E_n(t - t')\right], & t > t', \\ 0, & t < t'. \end{cases} \tag{59}$$

This quantity provides the probability amplitude for a particle observed at $(\mathbf{r}', t')$ to be found at $(\mathbf{r}, t)$. To see this, we note that Eq. (59) is the solution of the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{r}, t) = H(\mathbf{r}) \Psi(\mathbf{r}, t), \tag{60}$$

subject to the condition that at time t'

$$\Psi(\mathbf{r}, t') = \delta(\mathbf{r} - \mathbf{r}'). \tag{61}$$

This interpretation applies to an arbitrary Hamiltonian and is not restricted to that of an electron in a hydrogen atom, which we had in mind in Sec. 2.5. The above argument, when taken together with the brief explanation to be given in Sec. 7.5 on the path-integral approach to quantum mechanics, shows that the Green's function of the Schrödinger equation bridges the gap between quantum mechanics and classical dynamics (Sakurai, 1985). Also, in the perturbational treatment, we can follow an actual process of electron scattering in $(\mathbf{r}, t)$ space, by making use of the Green's function (see Secs. 6.2 and 6.3).

5.3 Dislocations

A dislocation is a line of singularities in the deformation field of an elastic body. Equation (53) thus describes a dislocation if the strength and distribution of the δ functions of the right-hand side are specified appropriately. For example, around a straight edge dislocation on the z axis with a y-directed Burgers vector $\mathbf{b}$ ($|\mathbf{b}| = b$), the elastic deformation $\mathbf{u}(\rho)$ as a function of the two-dimensional coordinate $\rho = (x, y)$ is given by (Landau and Lifshitz, 1986)

$$\mathbf{u}(\rho) = \mathbf{u}_0(\rho) + \mathbf{w}(\rho). \tag{62}$$

Here the vector $\mathbf{u}_0(\rho) = (u_{0x}(\rho), u_{0y}(\rho), u_{0z}(\rho))$ takes account of the multivaluedness of the deformation field and is defined by $u_{0z}(\rho) = 0$ and

$$u_{0x}(\rho) + iu_{0y}(\rho) = i\left(\frac{b}{2\pi}\right) \ln(x - iy). \tag{63}$$

It is essentially a fundamental solution of the two-dimensional Laplace equation. On the other hand, the vector $\mathbf{w}(\rho)$ in Eq. (62)

is obtained as

$$w_i(\rho) = \frac{Eb}{(1+\sigma)} \int_{-\infty}^{\infty} G_{i\gamma}(\mathbf{r}, \mathbf{r}')\, dz', \quad (64)$$

with $\mathbf{r} = (\rho, 0), \mathbf{r}' = (0, 0, z')$, and G_{ij}, Kelvin's solution, given by Eq. (54). The integral over z' takes account of the line of δ-function singularities on the z axis. From Eqs. (63) and (64), one can determine the strain and stress fields around a straight edge dislocation.

5.4
Magnetic Field around a Vortex Line in a Type-II Superconductor

It is well known, as the Meissner effect, that a magnetic field applied externally cannot penetrate into a superconductor. In a type-II superconductor, however, superconducting and nonsuperconducting regions coexist in a phase called the mixed phase, which is realized above a certain threshold strength for the magnetic field: As the field strength increases to pass the threshold, the magnetic flux begins to penetrate the sample in the form of a quantized flux line, called a vortex. Well below the upper critical field, above which the superconducting phase can no longer exist, the density of such vortices is so low that one may treat each of them as independent. It is shown that around an isolated rectilinear vortex, the microscopic magnetic field configuration is determined by (in cgs units following the convection of this field)

$$(\Delta - \lambda^{-2})\mathbf{h}(\mathbf{r}) = -\phi_0 \lambda^{-2} \widehat{\mathbf{z}} \delta(\rho), \quad (65)$$

$\phi_0 = hc/2e$ being the flux quantum and λ the penetration depth (Fetter and Hohenberg, 1969). Here $\widehat{\mathbf{z}}$ is the unit vector in the z direction and the vortex line is assumed to be situated on the z axis $\rho = 0$, $\rho = (x, y)$ being the two-dimensional position vector. The field $\mathbf{h}(\mathbf{r})$ is thus expressed by the Green's function of the two-dimensional Helmholtz equation [note, however, that κ^2 in Eq. (13) is replaced by $-\lambda^{-2}$]:

$$\mathbf{h}(\mathbf{r}) = \left(\frac{\phi_0 \lambda^{-2}}{2\pi}\right) \widehat{\mathbf{z}} K_0\left(\frac{\rho}{\lambda}\right), \quad (66)$$

with K_0 the modified Hankel function of order zero. If the right-hand side of Eq. (65) is replaced by a two-dimensional periodic distribution of δ functions, one may determine the magnetic field set up in a vortex lattice and calculate the gain in the free-energy density. The discussion of the equilibrium lattice structure that minimizes the free-energy density provides us with a basis for the more complete analysis due to Abrikosov et al., based on the Ginzburg-Landau equations. See for details Fetter and Hohenberg (1969).

6
Perturbational Treatment to Obtain Green's Functions

When an operator can be divided into two parts and the Green's function for one of the two is obtained straightforwardly, one may treat the remaining part perturbatively to obtain the full Green's function. For illustration, let us consider the stationary solution of the Schrödinger equation for the Hamiltonian of the following form:

$$H = H_0 + V. \quad (67)$$

It is very convenient to define the Green's function using the operator $(E - H)^{-1}$,

which is sometimes called the Greenian, and to examine it as a function of the energy variable E. One may show that the quantity $\langle \mathbf{r}|(E-H)^{-1}|\mathbf{r}'\rangle$, the $(\mathbf{r},\mathbf{r}')$ matrix element of the Greenian, behaves as a Green's function, and one may reasonably denote it simply as $G(\mathbf{r},\mathbf{r}';E)$ (see, e.g., Schiff, 1968). [The present definition leads to the Green's function $(i\hbar)^{-1}$ times that defined in Eq. (38)]. Rewriting the operator $(E-H)^{-1}$ as a power series with respect to the perturbation V, we obtain the following integral equation for G:

$$G(\mathbf{r},\mathbf{r}';E) = G_0(\mathbf{r},\mathbf{r}';E) + \int d^3r_1 \times G_0(\mathbf{r},\mathbf{r}_1;E)V(\mathbf{r}_1)G(\mathbf{r}_1,\mathbf{r}';E), \quad (68)$$

$G_0(\mathbf{r},\mathbf{r}';E)$ being the unperturbed Green's function $\langle \mathbf{r}|(E-H_0)^{-1}|\mathbf{r}'\rangle$. The expression is given for a local operator V, i.e., $\langle \mathbf{r}|V|\mathbf{r}'\rangle = V(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}')$. Iterating the right-hand side, we find

$$\begin{aligned} G &= G_0 + G_0VG_0 + G_0VG_0VG_0 + \cdots \\ &= G_0 + G_0TG_0. \end{aligned} \quad (69)$$

Here the simplified notation of G_0VG_0, etc., should be interpreted as in the second term of Eq. (68) by supplementing the arguments and integrals. The series in the first equality, the Born series, defines the operator T, which is usually called the t matrix by identifying its $(\mathbf{r},\mathbf{r}')$ matrix element $T(\mathbf{r},\mathbf{r}';E)$ with the operator itself. It is an effective scattering potential, taking into account the multiple scattering effect to all orders of V through the relation

$$\begin{aligned} T &= V + VG_0V + VG_0VG_0V + \cdots \\ &= V + VGV. \end{aligned} \quad (70)$$

Note that the Green's function in the second equality is G, not G_0, which already takes full account of the effect of V.

6.1 Slater-Koster Model

As shown by Eqs. (38) and (39), the eigenvalues for a Hamiltonian operator H are given by the poles of the Green's function $\langle \mathbf{r}|(E-H)^{-1}|\mathbf{r}'\rangle$. The Slater-Koster model is a typical model in which the series (69) is exactly calculable. Usually, it is defined by the model Hamiltonian which has a δ-function-type (contact-type) perturbation, $V(\mathbf{r}) = v_0\delta(\mathbf{r})$, v_0 being a constant for the potential strength. In this case, Eq. (69) leads to

$$G(\mathbf{r},\mathbf{r}';E) = G_0(\mathbf{r},\mathbf{r}';E) + G_0(\mathbf{r},0;E) \times \left\{\frac{v_0}{1-v_0G_0(0,0;E)}\right\} G_0(0,\mathbf{r}';E). \quad (71)$$

Thus, the zeros of the denominator of the second term as a function of E provide the eigenvalues for the full Hamiltonian:

$$1 - v_0G_0(0,0;E) = 0. \quad (72)$$

Considering that G_0 has poles at the unperturbed energy eigenvalues [the eigenvalues of the Hamiltonian H_0 in Eq. (67)], we see that this equation has a form quite adequate to see how these eigenvalues are modified by the perturbation. Using Eq. (72), one can discuss, for example, how a plane-wave state $\mathbf{k}$ is affected by the potential $V(\mathbf{r})$ and whether or not a bound state appears for an attractive potential $V(\mathbf{r})$ for a given unperturbed energy spectrum. In condensed-matter physics these problems are very frequently encountered in dealing with electrons, phonons, magnons, etc. For a detailed analysis, see, e.g., the book by Economou (1990).

The treatment for a photon along this line is just the perturbational treatment of the Maxwell equations. For example, the reflection and transmission of light

for a rough solid surface was analyzed successfully by this approach (see, e.g., Maradudin et al., 1990).

6.2 Scattering Cross Section of a Plane Wave from a Scatterer with Spherical Symmetry

In the scattering of an incident wave ψ_0 of, say, an electron with energy E by a perturbing potential $V(\mathbf{r})$, the total wave function for the Schrödinger equation $H\psi = E\psi$ with H given by Eq. (67) may be expressed as (Schiff, 1968)

$$\begin{aligned}\psi(\mathbf{r}) &= \psi_0(\mathbf{r}) + \int d^3r' G_0(\mathbf{r}, \mathbf{r}'; E) \\ &\quad \times V(\mathbf{r}')\psi(\mathbf{r}') \\ &= \psi_0(\mathbf{r}) + \int\int d^3r_1 d^3r_2 G_0(\mathbf{r}, \mathbf{r}_1; E) \\ &\quad \times T(\mathbf{r}_1, \mathbf{r}_2; E)\psi_0(\mathbf{r}_2). \qquad (73)\end{aligned}$$

The second equality is obtained by iterating the first equality and using the definition (70) for the t matrix T. For a spherically symmetric potential $V(\mathbf{r}) = V(r)$, we may go further by resolving $\psi(\mathbf{r})$ and $\psi_0(\mathbf{r})$ into partial waves according to the magnitudes of angular momentum $\hbar l$. The partial-wave expansion of the incident plane wave with wave vector $\mathbf{k}_i$ is

$$\begin{aligned}\psi_0(\mathbf{r}) &= e^{i\mathbf{k}_i\cdot\mathbf{r}} \\ &= \sum_{l=0}^{\infty}(2l+1)i^l j_l(k_i r)P_l(\cos\theta), \qquad (74)\end{aligned}$$

where $j_l(k_i r)$ is the spherical Bessel function of order l, the scattering angle θ is the angle between $\mathbf{k}_i$ and $\mathbf{r}$, and $P_l(\cos\theta)$ is the Legendre function of order l. Asymptotically, the solution ψ in Eq. (73) behaves as

$$\psi(\mathbf{r}) = \psi_0(\mathbf{r}) + \frac{f(\theta)e^{ik_s r}}{r}, \qquad (75)$$

with k_s the wave number of the scattered spherical wave. Comparing Eqs. (73) and (75) and using the form given by Eq. (15), we find that the amplitude of the outgoing spherical wave is given by

$$f(\theta) = -\left(\frac{m}{2\pi\hbar^2}\right) T(\mathbf{k}_s, \mathbf{k}_i; E), \qquad (76)$$

where

$$\begin{aligned}T(\mathbf{k}_s, \mathbf{k}_i; E) &= \int\int e^{-i\mathbf{k}_s\cdot\mathbf{r}} T(\mathbf{r}, \mathbf{r}'; E) \\ &\quad \times e^{i\mathbf{k}_i\cdot\mathbf{r}'} d^3r\, d^3r', \qquad (77)\end{aligned}$$

with $\mathbf{k}_s$ the wave vector of the scattered wave, directed from the origin to the observation point $\mathbf{r}$ ($k_s = k_i$ by energy conservation). The t matrix $T(\mathbf{k}_s, \mathbf{k}_i; E)$ for the two wave vectors, both related to the incident energy E through the relation $k_i = k_s = (2mE/\hbar^2)^{1/2}$, is called the t matrix on the energy shell and is known to be expressed in terms of the phase shift $\delta_l(E)$ of the partial wave l. In this way, $f(\theta)$ is finally expressed as

$$f(\theta) = \frac{1}{2ik_i}\sum_{l=0}^{\infty}(2l+1)(e^{2i\delta_l} - 1)P_l(\cos\theta). \qquad (78)$$

The differential and total cross sections are then obtained from $|f(\theta)|^2$.

To carry out the partial-wave expansion for the Maxwell equations, we need the tensor of the Green's functions (Mahan, 1990) and the vector spherical wave for each l (Stratton, 1941). With the difference that we now require two kinds of phase shifts for each l because of the character of light as a transverse vector field, the Mie and Rayleigh scatterings of light (Born and Wolf, 1975) may be treated compactly as in the electron case.

6.3 Band Structure of Electrons and Photons and Diffraction Experiments

The band structure of electron states in solids may be viewed as arising from the electron scattering by atoms arrayed periodically in a lattice. With the difference that the perturbation V is now due to a periodic array of scatterers, Eqs. (67)–(70) still hold here without modification. Since the poles of the full Green's function are identical to those of the t matrix [Eq. (69)], the calculation of the t matrix for arrayed scatterers is equivalent to the band-structure calculation. If we denote the multiple-scattering effect of the kth scatterer by the t matrix t_k, the scattering from the array as a whole is described by the following t matrix:

$$T = \sum_k t_k + \sum_{\substack{k,k' \\ (k \neq k')}} t_k G_0 t_{k'} + \cdots, \qquad (79)$$

where G_0 is the Green's function for free motion. The first term describes the scattering from a single site. The second exhibits the process where an electron, once scattered by the kth scatterer, propagates to another site k' and undergoes a multiple scattering there. The constraint $k \neq k'$ eliminates double counting, because the term $t_k G_0 t_k$ to be removed is already involved in the single t matrix t_k [see the series expansion of the first equality of Eq. (70)]. The higher-order terms in Eq. (79) incorporate processes involving three or more scatterers. These scattering processes are exhibited in Fig. 3. If the potentials are spherical and nonoverlapping with each other, as is usually the case to a good approximation, only the on-the-energy-shell values are involved in each t_k (see Sec. 6.2). Since G_0 sandwiched between t_k and $t_{k'}$ in Eq. (79) describes the free propagation of an electron from one atom k to another k', it depends solely upon the structure of the lattice. Therefore Eq. (79) shows that the band structure of electron energies in solids is determined quite generally by two quantities: the phase shifts of the atoms and the structure factor of the periodic array they form. This is indeed explicit in the Korringa-Kohn-Rostoker (KKR) eigenvalue equation for the band structure of electrons. See the monograph by Lukes (1969) for the derivation of the KKR equation based on Eq. (79).

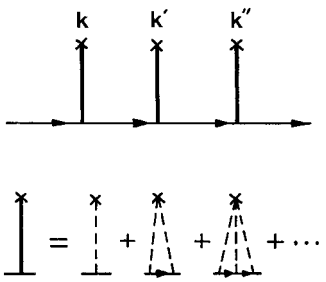


Fig. 3 Scattering of an electron in a solid. The vertical solid line at site k is the t matrix t_k used in Eq. (79) and the arrowed horizontal lines show G_0. The t matrix t_k is defined by the lower graph, which shows the series given by Eq. (70) with the dotted lines for the atomic potential V

When we apply Eq. (73) to a periodic array of scatterers, the solution $\psi(\mathbf{r})$ describes how the atomic array responds to an incident electron. When, in particular, an incident electron from the outside has an energy that is not very different from the positions of the poles of the t matrix, the amplitude of the scattered wave reflects, through $T(\mathbf{r}_1, \mathbf{r}_2; E)$ of Eq. (73), the detailed band structure as a function of incident energy E and wave vector. This is essentially the origin of the fine structure observed in the usual low-energy-electron diffraction (Tong, 1975).

Similar fine structure is expected to arise in the diffraction of visible light, i.e., in the light scattered from a periodic array of

scatterers with periodicity in the visible range. X-ray diffraction is simply its high-energy limit. For details of the diffraction of visible light reflecting the photon band structure, see Ohtaka (1980) and Inoue and Ohtaka (1982) and the analysis by Yablonovitch and Gmitter (1989).

7 Green's Functions in Many-Body Theories

The developments of condensed-matter physics and elementary-particle physics owe considerably to the use of Green's functions in the treatment of many-body interactions. How they are defined and why they have been so widely used will be shown through several examples. For more details, see the books by, e.g., Abrikosov et al. (1963), Fetter and Walecka (1971), Economou (1990), and Mahan (1990). For illustration we have in mind a model Hamiltonian for interacting spinless particles, either bosons or fermions:

$$\begin{aligned} K &= \int d^3r \Psi^\dagger(\mathbf{r})[\left(\frac{-\hbar^2}{2m}\right)\Delta - \mu]\Psi(\mathbf{r}) \\ &\quad + \frac{1}{2}\int\int d^3r\, d^3r' \Psi^\dagger(\mathbf{r})\Psi^\dagger(\mathbf{r}') \\ &\quad \times v(|\mathbf{r}-\mathbf{r}'|)\Psi(\mathbf{r}')\Psi(\mathbf{r}) \\ &= K_0 + K' \end{aligned} \tag{80}$$

expressed in the second quantized form. Here $\Psi^\dagger(\mathbf{r})$ and $\Psi(\mathbf{r})$ are the field operators for creating and annihilating, respectively, a particle at position $\mathbf{r}$. The term proportional to the chemical potential μ in the first term on the right-hand side shows that we are considering the grand canonical ensemble. The two-body correlation described by the operator K' may be treated using Green's functions.

7.1 Single- and Two-Particle Green's Functions

The single-particle Green's function is defined in many-body theories by

$$G(x, x') = -\frac{i\langle\Phi_0|T(\Psi(x)\Psi^\dagger(x'))|\Phi_0\rangle}{\langle\Phi_0|\Phi_0\rangle}. \tag{81}$$

Here x and x' denote $(\mathbf{r},t)$ and $(\mathbf{r}', t')$, respectively, Φ_0 is the exact ground state for the Hamiltonian given by Eq. (80), and the time-dependent operators $\Psi(x)$ etc. are the operators in the Heisenberg picture defined by

$$\Psi(x) = e^{iKt/\hbar}\Psi(\mathbf{r})e^{-iKt/\hbar}. \tag{82}$$

The symbol T in Eq. (81) is the time-ordering operator that is defined by the rule

$$T(\Psi(x)\Psi^\dagger(x')) = \begin{cases} \Psi(x)\Psi^\dagger(x'), & t > t', \\ \pm\Psi^\dagger(x')\Psi(x), & t < t', \end{cases} \tag{83}$$

the upper (lower) sign referring to bosons (fermions). Aside from the presence of the time-ordering operator T, the Green's function defined by Eq. (81) can be given the physical interpretation of a probability amplitude, analogous to that given by Eq. (59). Indeed, when the particle interaction vanishes, i.e., $K = K_0$ in Eq. (80), G tends to the unperturbed Green's function

$$G_0(x, x') = -i\langle 0|T(\Psi(x)\Psi^\dagger(x'))|0\rangle, \tag{84}$$

$|0\rangle$ being the normalized ground state for K_0. It satisfies

$$\left(i\hbar\frac{\partial}{\partial t} + \frac{\hbar^2}{2m}\Delta + \mu\right)G_0(x, x') = \hbar\delta(x - x'). \tag{85}$$

The inhomogeneous term comes from the presence of the operator T, the

origin of the step function via Eq. (83), and the commutation (anticommutation) relation of the field operators for bosons (fermions). Through Eq. (85), we see obviously that iG_0 is the many-body analog of the single-particle Green's function treated in Sec. 5.2.

The first reason why the single-particle Green's function is so useful is that many important observables such as the ground-state energy, number or current density, etc. of the many-body system, are expressed in terms of the Green's function (81). The second reason is that the effect of K' may be handled by applying Wick's theorem and interpreted visually with the help of Feynman diagrams (see Sec. 7.2). Third, part of the interaction process is taken into account to all orders with respect to K'.

For example, the modification of a single-particle energy due to the mutual interaction K' is taken into account by Dyson's equation

$$G(\mathbf{k}, \omega) = G_0(\mathbf{k}, \omega) + G_0(\mathbf{k}, \omega)\Sigma(\mathbf{k}, \omega)G(\mathbf{k}, \omega) \quad (86)$$

in the Fourier space $(\mathbf{k}, \omega)$. From this, we find

$$G(\mathbf{k}, \omega) = \left[\omega - \left(\frac{\hbar k^2}{2m}\right) + \left(\frac{\mu}{\hbar}\right) - \Sigma(\mathbf{k}, \omega)\right]^{-1}, \quad (87)$$

where we have used

$$G_0(k, \omega) = \left[\omega - \left(\frac{\hbar k^2}{2m}\right) + \left(\frac{\mu}{\hbar}\right) + i\varepsilon\ \mathrm{sgn}(\omega)\right]^{-1} \quad (88)$$

given by Eq. (85). Note that G_0 has a causal form (see Sec. 2.2). The complex quantity $\Sigma(\mathbf{k}, \omega)$ is called the self-energy part. The form (87) provides an exact expression for G, if the self-energy part is exactly given. In this sense, Dyson's equation (86) is one of the key equations in treating many-body interactions. The imaginary part of $\Sigma(\mathbf{k}, \omega)$ determines the lifetime of a plane-wave state $\mathbf{k}$, caused by the many-body interaction. Although it is generally a hopeless task to attempt to make an exact calculation of $\Sigma(\mathbf{k}, \omega)$ with all possible many-body processes included, an important subset may usually be taken into account. For a system of electrons interacting with Coulomb repulsion (the electron-gas model), for example, we may now say that some of the physical quantities have so far been obtained almost exactly.

The two-particle Green's function is defined by

$$G(x_1, x_2; x_1', x_2') = \frac{(-i)^2\langle\Phi_0|T(\Psi(x_1)\Psi(x_2)\Psi^\dagger(x_2')\Psi^\dagger(x_1'))|\Phi_0\rangle}{\langle\Phi_0|\Phi_0\rangle}. \quad (89)$$

It is so called because it deals with the two particles created at x_1' and x_2'. Since a correlation function between two physical quantities is usually expressed by the two-particle Green's function, the latter is an important quantity, yielding transport coefficients, conductivity, and dielectric constant, for example (see Sec. 7.4). As mentioned before, Wick's theorem is a key theorem for calculating the Green's function. The names "random-phase approximation", "ladder approximation", etc., are each assigned to a special choice of infinite series of Feynman diagrams considered in two-particle Green's functions. We give in Fig. 4 two typical examples, which exhibit how the two particles interact with each other, sometimes involving other particles.

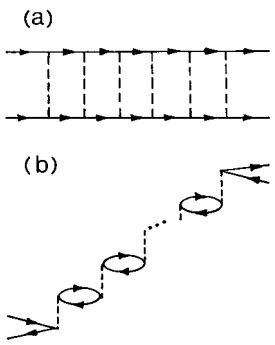


Fig. 4 Two-particle Green's function taking into account the particle interaction: (a) ladder diagram, (b) random-phase approximation. The dotted lines show the particle interaction, ν of Eq. (80), and the arrowed solid lines show the unperturbed Green's function G_0, defined by Eq. (84). When ν is an attractive short-range interaction for an narrow energy range, the ladder diagram for a singlet pair of electrons leads to superconductivity. When ν is a repulsive Coulomb potential, the random-phase approximation leads, in the low-density limit, to the exact ground-state energy of interacting electrons

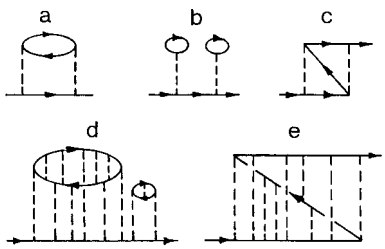


Fig. 5 Examples of Feynman diagrams. The graphs *a* through *c* show examples of second-order corrections to the single-particle Green's functions for the Hamiltonian given by Eq. (80). The vertical dotted lines show the particle interaction ν. The graphs *d* and *e* are two higher-order corrections

7.2 Wick's Theorem and Feynman Diagrams

The Green's functions (81) and (89) are defined in terms of the operators in the Heisenberg picture for the full Hamiltonian K. In calculating them perturbationally, using the relationships between Φ_0 and $|0\rangle$ and between the operators in the Heisenberg and interaction representations, we encounter the following types of quantities:

$$\langle 0|T(\Psi_0(x_1)\Psi_0(x_2)\Psi_0(x_3)\cdots\Psi_0^\dagger(x_{2n-1}) \times \Psi_0^\dagger(x_{2n}))|0\rangle, \tag{90}$$

where $\Psi_0(x)$ etc. are the operators in the interaction picture defined by

$$\Psi_0(x) = e^{iK_0t/\hbar}\Psi(\mathbf{r})e^{-iK_0t/\hbar}. \tag{91}$$

Quantities such as (90) are the expectation values in the state $|0\rangle$ of the time-ordered product of an equal number of operators $\Psi_0(x)$ and $\Psi_0^\dagger(x)$. Wick's theorem handles just such quantities and guarantees that the quantity given above may be resolved into a sum of products of n unperturbed Green's functions G_0's, formed by pairing a Ψ_0 with a $\Psi_0^\dagger$ in all possible ways. For a rigorous proof, see, e.g., Fetter and Walecka (1971). In this way the perturbation series for a Green's function, either a single-particle one or a two- or many-particle one, is expressed as a sum of products of a number of G_0's, and each product is given a graphical representation called Feynman diagram (Feynman graph), where a line represents an unperturbed propagator G_0. Some examples were already given for the two-particle Green's functions in Fig. 4. For K given by Eq. (80), several second- and higher-order corrections for the single-particle Green's function are shown in Fig. 5.

7.3 Temperature Green's Functions

The temperature Green's functions are essentially the Green's functions treated in

Sec. 7.1 with the time argument t and t' replaced by the imaginary time $-i\tau$ and $-i\tau'$, respectively, and with the expectation value $\langle\Phi_0|\cdots|\Phi_0\rangle$ for the ground state replaced by the thermal average $\langle\cdots\rangle$. The temperature Green's function also goes under the name of Matsubara, who first introduced them (Kubo et al., 1991). They are convenient tools for calculating various physical quantities in thermal equilibrium – for instance, the single-particle energy and thermodynamic potential. Wick's theorem holds here again. Because of the periodic properties with respect to the argument $\tau - \tau'$, the Fourier components of the Green's function $G(\tau - \tau')$ are defined at the discrete frequencies $\omega_n = (2n+1)\pi k_B T/\hbar$ for fermions and $\omega_n = 2n\pi k_B T/\hbar$ for bosons, n being an integer, T the temperature, and k_B the Boltzmann constant. For those frequencies, the Dyson equation (86) holds and is solved algebraically. The single-particle energy is obtained by analytically continuing $G(\omega_n)$ from the discrete points $i\omega_n$ on the imaginary axis to the complex ω plane. For the BCS Hamiltonian of superconductivity, for example, all the predictions of the BCS theory (Bardeen et al., 1957) have been reproduced straightforwardly in this way and the Ginzburg-Landau equations, originally introduced semiphenomenologically, have been given an unambiguous microscopic foundation. The developments of the theory and experiment of superconductivity owed much to the use of temperature Green's functions (see, e.g., Abrikosov et al., 1963; Parks, 1969; Mahan, 1990). The advantage of the two-particle temperature Green's functions can be best appreciated in connection with the linear response theory of Kubo treated below.

7.4 Linear Response Theory of Kubo and Temperature Green's Functions

How the system responds to a small external signal is summarized by the Kubo formula of linear response theory. Its essence lies in that the response function is given by a spatial and temporal correlation function of fluctuations, calculated in thermal equilibrium without the external perturbation. For example, when a weak electric field, whose ν component is $E_\nu(\mathbf{r}', t')$, of frequency ω is an input signal $I(\mathbf{r}', t')$ in the input-output relation given by Eq. (2), an Ohmic electric current at the position $(\mathbf{r},t)$ is the output signal $O(\mathbf{r},t)$. The response function, the Green's function $G(x, x')$ in Eq. (2), is in this case the ac conductivity tensor $\sigma_{\mu\nu}(x, x')$. Here the fluctuation in the Kubo formula is that of the current density. Letting $j_\alpha(x)$ be the operator for the current density in the α direction in the Heisenberg picture, the fluctuation is defined by $j_\alpha(x) - \langle j_\alpha(x)\rangle$ [$= j_\alpha(x)$, because $\langle j_\alpha(x)\rangle = 0$ in equilibrium]. The response function is then expressed by

$$\sigma_{\mu\nu}(x, x') = i\left(\frac{ne^2}{m\omega}\right)\delta_{\mu\nu}\delta(\mathbf{r}-\mathbf{r}') + \left(\frac{-1}{\hbar\omega}\right)\langle[j_\mu(x), j_\nu(x')]\rangle\theta(t-t'), \quad (92)$$

with n the electron density, m the mass, and the Heaviside step function in the second term guaranteeing causality (see Sec. 2.2). The square bracket in the second term is the commutator and $\langle\rangle$ denotes the thermal average. Because of the presence of the step function, quantities like that in the second term of Eq. (92) are called retarded Green's functions in linear response theory, in analogy with their counterpart for the wave equation in Sec. 2.2.

The way of calculating the Fourier transform of the correlation function involved in the second term of Eq. (92) is summarized by the theorem, sometimes called the Abrikosov-Gor'kov-Dzyaloshinski-Fradkin theorem, which relates a retarded Green's function to the corresponding temperature Green's function (Abrikosov et al., 1963). According to this theorem, the first step is the calculation of the two-particle temperature Green's function

$$\widetilde{G}_2(\mathbf{r}, \tau; \mathbf{r}', \tau') = \langle T_\tau(j_\mu(\mathbf{r}, \tau) j_\nu(\mathbf{r}', \tau'))\rangle \tag{93}$$

for the Fourier component $\widetilde{G}_2(\mathbf{k}, \omega_n)$, and the second is the analytic continuation of the result to the real frequency ω, carried out simply by changing $i\omega_n$ to $\omega + i\varepsilon$. The first step is carried out with the help of Wick's theorem, and the second is a procedure analogous to that used in obtaining a retarded Green's function in Sec. 2.2 (Abrikosov et al., 1963).

Since this method of calculation using the temperature Green's functions was introduced, the usefulness of the Kubo formula has increased remarkably. This was indeed one of the key steps in the development of condensed-matter physics.

7.5 Green's Functions in Quantum Field Theories

One typical way of quantizing a classical Lagrangian is based on the functional integral. One of the merits of this method is that we may treat the field theory and statistical physics on the same footing (Amit, 1984).

The Green's functions and the related quantities developed in this field are outlined here by taking an interacting scalar Bose field $\phi(x)$ in four-dimensional space as an example (Ramond, 1981). The Lagrangian density $L(\phi(x))$ for the model called the ϕ^4 model is defined [in natural units, where $c = \hbar = 1$ and the dimension of the field $\phi(x)$ is (length)$^{-1}$] by

$$L(\phi) = L_0(\phi) + L_{\text{int}}(\phi), \tag{94}$$

with

$$L_0(\phi) = \phi_t(x)^2 - \phi_x(x)^2 - \mu^2\phi(x)^2,$$
$$L_{\text{int}}(\phi) = \left(\frac{\lambda}{4!}\right)\phi^4(x), \tag{95}$$

ϕ_t and ϕ_x being $\partial\phi/\partial t$ and $\partial\phi/\partial x$, respectively. The functional $Z(J)$ for generating the Green's functions is defined by

$$Z(J) = \mathscr{N}^{-1}\int \mathscr{D}\phi \exp\left(i\int [L(\phi) + J\phi]d^4x\right). \tag{96}$$

Here $J(x)$ is a fictitious external field linearly coupled to $\phi(x)$, the factor $\mathscr{N}^{-1}$ normalizes $Z(J)$ such that $Z(0) = 1$, and the measure $\mathscr{D}\phi$ of the functional integral is defined by

$$\mathscr{D}\phi = \lim_{N\to\infty}\prod_{n=1}^{N} d\phi(x_n). \tag{97}$$

Here it is understood that the integral over x in Eq. (96) is treated as a sum over the integrand at N discrete points $\{x_n\}$ in four-dimensional space, and the limit for N is taken after the integrals over the N independent variables $\{\phi(x_n)\}$ have been carried out. Equation (96) is a generalization of the quantization scheme for the classical one-dimensional motion of a particle carried out through the quantity

$$F(t_2, t_1) = \int \mathscr{D}q \exp\left(\frac{i}{\hbar}\int_{t_1}^{t_2} L(q(t), q'(t))\, dt\right), \tag{98}$$

where the classical Lagrangian L is defined in terms of $q(t)$ and $q'(t)$, the particle position and velocity at time t, respectively. With the restriction $q(t_1) = q_1$ and $q(t_2) = q_2$ imposed on the path, it may be shown that, if the measure $\mathscr{D}q$ is defined appropriately (Sakurai, 1985), Eq. (98) is identical to the quantum-mechanical probability amplitude of a particle, just described by the Green's function treated in Secs. 2.3.2 and 5.2. Thus the quantization through the functional integral is well established.

From Eq. (96), the Green's function is obtained as follows:

$$G(x_1, x_2) = \left.\frac{\delta^2 Z(J)}{\delta J(x_1)\delta J(x_2)}\right|_{J=0} = -\mathscr{N}^{-1} \int \mathscr{D}\phi T\left[\phi(x_1)\phi(x_2) \times \exp\left(i\int L(x)d^4x\right)\right], \quad (99)$$

where the functional derivative $\delta Z(J)/\delta J(x_1)$ is defined by the ratio of the induced change of $Z(J)$ to the infinitesimal variation $\delta J(x_1)$ at x_1 of the external field. Many-point correlation functions, $G_2(x_1, x_2, x_3, x_4)$, etc., are defined similarly. The unperturbed Green's function is then calculated from $Z_0(J)$, the generating function for $L = L_0$. It is shown that

$$Z_0(J) = \exp\left(\frac{i}{2}\int\int J(x)g(x-y) \times J(y)d^4x\, d^4y\right), \quad (100)$$

with $g(x-y)$ the Green's function for the Klein-Gordon equation, G_R, G_A, or G_C of Eq. (29), according to the way of avoiding the poles in the ω integral (see Secs. 2.2 and 2.3.3). From Eq. (99), the unperturbed Green's function G_0 is then obtained as

$$G_0(x_1, x_2) = ig(x_1 - x_2). \quad (101)$$

In order to take into account L_{int}, the following relation is useful:

$$Z(J) = \mathscr{N}^{-1} \exp\left[i\int d^4x L_{\text{int}}\left(-\frac{i\delta}{\delta J(x)}\right)\right] \times Z_0(J), \quad (102)$$

$Z_0(J)$ being given by Eq. (100). With Eqs. (100) and (102), we can calculate Green's functions to any desired order. For example, to obtain the perturbation expansion of G with respect to the parameter λ involved in L_{int}, we have only to expand the exponential of Eq. (102) in terms of L_{int}.

It is remarkable that Wick's theorem, Dyson's equation, Feynman diagrams, etc., all apply here without any modification. The difficulty of ultraviolet divergence is thus handled and then a number of physical quantities are defined by a procedure that extracts the finite part of every term in the perturbation series. This is the regularization and renormalization program.

To connect field theory with statistical physics, we need only to change the time t to the imaginary time $-i\tau$ (called in this field the Wick rotation), just as in introducing the temperature Green's functions in Sec. 7.3. The point is that the generating function $Z(J)$ converts itself essentially to a partition function. In our example this will be seen by noting that the Lagrangian is then transformed to minus the Hamiltonian. Taking advantage of this remarkable connection, and combining it with renormalization-group procedures, the critical properties of a statistical system near a second-order phase transition can be discussed using field-theoretical methods. For example, the critical index of the specific heat of the Ising model

can be successfully obtained in this way. See, for more detail, e.g., Amit (1984) and Itzykson and Drouffe (1989).

Glossary

Adjoint Operator: For a second-order differential operator L with respect to x, y, and z, we may transform the integral in the region Ω,

$$\int_\Omega v(\mathbf{r})L[u(\mathbf{r})]d^3r,$$

by integrating by parts until all the derivatives on $u(\mathbf{r})$ in the volume integral are transferred to $v(\mathbf{r})$. Finally the volume integral reduces to the form

$$\int_\Omega u(\mathbf{r})M[v(\mathbf{r})]d^3r,$$

with a new differential operator M, called the adjoint operator of L. Thus we obtain the following relation, the generalized Green's formula:

$$\int_\Omega (vL[u] - uM[v])d^3r = \int_\Gamma (Px_n + Qy_n + Rz_n)\, ds.$$

The right-hand side shows the surface integral that remains in the integration by parts, where x_n, y_n, and z_n are the directional cosines of the outward normal $\hat{\mathbf{n}}$ to the boundary Γ, and P, Q, and R are functions of u, v, and their derivatives, determined dependent upon the form of L (see, e.g., Courant and Hilbert, 1937).

Dirichlet Problem: A boundary-value problem of a differential equation that seeks a solution with a specified boundary value. These types of boundary conditions are called the Dirichlet condition.

Fredholm Integral Equation: For an unknown function $u(\mathbf{r})$ in the domain Ω, the following integral equation is called the Fredholm integral equation of the second kind:

$$u(\mathbf{r}) - \int_\Omega K(\mathbf{r}, \mathbf{r}')u(\mathbf{r}')d^3r' = f(\mathbf{r}),$$

$f(\mathbf{r})$ being a given function and K the integral kernel. The Fredholm integral equation of the first kind is the one with the first term $u(\mathbf{r})$ missing on the left-hand side.

Generalized Green's Formula: For the Laplacian operator the Green's formula is given by Eq. (4). It is generalized to an arbitrary second-order differential operator L and its adjoint M in the form shown in the definition of the adjoint operator.

Homogeneous Boundary Condition: The boundary condition such as $u(\mathbf{r})|_\Gamma = 0$ or $(\partial/\partial n)u(\mathbf{r})|_\Gamma = 0$ on the boundary Γ of the domain under consideration is a homogeneous boundary condition. In general, homogeneous boundary conditions consist of relations between the value of $u(\mathbf{r})|_\Gamma$ and its derivative $(\partial/\partial n)u(\mathbf{r})|_\Gamma$. If $u(\mathbf{r})$ satisfies the homogeneous boundary condition, so does $cu(\mathbf{r})$, c being an arbitrary constant. Boundary conditions for which this does not hold are called inhomogeneous boundary conditions. Examples are $u(\mathbf{r})|_\Gamma = f(\mathbf{r})$ or $(\partial/\partial n)u(\mathbf{r})|_\Gamma = g(\mathbf{r})$, with given nonzero $f(\mathbf{r})$ or $g(\mathbf{r})$ defined on the boundary.

Internal or External Boundary-Value Problem: When the solution of a boundary-value problem is sought inside the boundary on which a boundary condition is imposed, it is called the internal boundary-value problem. The problem for the outside region is the external problem.

Neumann Problem: The boundary condition specifying $\partial u/\partial n$, the derivative in

the direction of the outward normal to the boundary, is called the Neumann condition. A boundary-value problem with a Neumann condition is a Neumann problem.

Vector Spherical Waves: Let Φ be a scalar function satisfying the scalar Helmholtz equation

$$(\Delta + k^2)\Phi = 0.$$

From Φ one may construct three vector fields, $\mathbf{L} = \mathrm{grad}\Phi$, $\mathbf{M} = \mathrm{rot}(\mathbf{a}\Phi)$, and $\mathbf{N} = (1/k)\mathrm{rot}\mathbf{M}$, with $\mathbf{a}$ any constant vector. They all satisfy the vector Helmholtz equation. The two vectors $\mathbf{M}$ and $\mathbf{N}$ are solenoidal, while $\mathbf{L}$ is irrotational. For $\Phi = C_l(kr)Y_{lm}(\widehat{\mathbf{n}})$, C_l being the lth cylindrical function and $Y_{lm}(\widehat{\mathbf{n}})$ the spherical harmonic, the vectors $\mathbf{L}$, $\mathbf{M}$, and $\mathbf{N}$ are called the lth vector spherical waves.

Wick's Theorem: The theorem transforming the time-ordered product $T(\Psi(x)\Psi(x')\cdots)$ of any number of field operators in the interaction picture into a sum of products of simpler quantities (see, e.g., Fetter and Walecka, 1971). The expectation value of a product of field operators in the noninteracting state may be calculated by use of this theorem.

List of Works Cited

Abramowitz, M., Stegun, I. A. (1965), *Handbook of Mathematical Functions*, New York: Dover.

Abrikosov, A. A., Gor'kov, L. P., Dzyaloshinski, I. E. (1963), *Methods of Quantum Field Theory in Statistical Physics*, Englewood Cliffs, NJ: Prentice-Hall.

Amit, D. J. (1984), *Field Theory, The Renormalization Group, and Critical Phenomena*, 2nd ed., Singapore: World Scientific.

Baker, B. B., Copson, E. T. (1950), *The Mathematical Theory of Huygens' Principle*, Oxford: Clarendon.

Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957), *Phys. Rev.* **108**, 1175–1204.

Bogoliubov, N. N., Shirkov, D. V. (1959), *Introduction to The Theory of Quantized Fields*, New York: Interscience.

Born, M., Wolf, E. (1975), *Principles of Optics*, 5th ed., New York: Pergamon.

Brebbia, C. A. (1978), *The Boundary Element Method for Engineers*, New York: Pentech.

Brebbia, C. A., Walker, S. (1980), *Boundary Element Techniques in Engineering*, New York: Newns-Butterworths.

Butkovskiy, A. G. (1982), *Green's Functions and Transfer Functions*, New York: Ellis Horwood.

Courant, R., Hilbert, D. (1937), *Methoden der Mathematischen Physik*, Berlin: Springer [reprint: (1989), New York: Interscience], Vols. 1 and 2.

Economou, E. N. (1990), *Green's Functions in Quantum Physics*, 2nd ed., Berlin: Springer.

Fetter, A. L., Hohenberg, P. C. (1969), in: R. D. Parks (Ed.), *Superconductivity*, New York: Dekker, Vol. 2.

Fetter, A. L., Walecka, J. D. (1971), *Quantum Theory of Many-Particle Systems*, New York: McGraw-Hill.

Feynman, R. P. (1972), *Statistical Mechanics*, New York: Benjamin.

Feynman, R. P., Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill.

Inoue, M., Ohtaka, K. (1982), *Phys. Rev. B* **26**, 3487–3509.

Itzykson, C., Drouffe, J.-M. (1989), *Statistical Field Theory*, New York: Cambridge University Press, Vols. 1 and 2.

Itzykson, C., Zuber, J.-B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Kellog, O. D. (1939), *Foundation of Potential Theory*, Berlin: Springer [reprint: (1953), New York: Dover].

Kubo, R., Toda, M., Hashitsume, N. (1991), *Statistical Physics II*, 2nd ed., Berlin: Springer.

Landau, L. D., Lifshitz, E. M. (1986), *Theory of Elasticity*, 3rd ed., Oxford: Oxford University Press.

Lukes, T. (1969), in: P. T. Landsberg (Ed.), *Solid State Theory*, New York: Wiley-Interscience.

Mahan, G. D. (1990), *Many-Particle Physics*, 2nd ed., New York: Plenum.

Maradudin, A. A., Michel, T., McGurn, A. R., Méndez, E. R. (1990), *Ann. Phys. (N.Y.)* **203**, 255–307.

Morino, L., Chen, L. T., Suciu, E. O. (1975), *AIAA J.* **13**, 368–374.
Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill.
Ohtaka, K. (1980), *J. Phys. C* **13**, 667–680.
Parks, R. D. (1969), *Superconductivity*, New York: Dekker, Vols. 1 and 2.
Ramond, P. (1981), *Field Theory*, Reading, MA: Benjamin.
Sakurai, J. J. (1985), *Modern Quantum Mechanics*, Reading, MA: Addison-Wesley.
Schiff, L. I. (1968), *Quantum Mechanics*, 3rd ed., New York: McGraw-Hill.
Smirnov, V. I. (1965), *A Course of Higher Mathematics*, Reading, MA: Addison-Wesley, Vol. IV.
Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.
Stratton, J. A. (1941), *Electromagnetic Theory*, New York: McGraw–Hill.
Tong, S. Y. (1975), *Prog. Surf. Sci.* **7**, 1–48.
Yablonovitch, E., Gmitter, T. J. (1989), *Phys. Rev. Lett.* **63**, 1950–1953.
Yosida, K. (1965), *Functional Analysis*, Berlin: Springer.

Further Reading

Bateman, H. (1932), *Partial Differential Equations of Mathematical Physics*, New York: Cambridge. An overview of Green's functions concentrating on the mathematical aspects.
Doniach, S., Sondheimer, E. H. (1974), *Green's Functions for Solid State Physicists*, Reading, MA: Addison-Wesley. A readable textbook for the Green's functions used in solid-state physics.
Hedin, L., Lundqvist, S. (1969), in: H. Ehrenreich, F. Seitz, D. Turnbull (Eds.), *Solid State Physics*, New York: Academic, Vol. 23. A monograph for detailed use of Green's functions in many-electron systems.
Itzykson, C., Zuber, J.-B. (1980), *Quantum Field Theory*, New York: McGraw-Hill. A detailed textbook for quantum field theory.
Jackson, J. D. (1967), *Classical Electrodynamics*, New York: Wiley. Some examples of Green's functions as applied to electrodynamics are found.
Jaswon, M. A., Symm, G. T. (1977), *Integral Equation Methods in Potential and Electrostatics*, New York: Academic. A textbook for the foundation of the boundary-value problem.
Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill. A detailed textbook for Green's functions used in many fields of theoretical physics.
Newton, R. G. (1966), *Scattering Theory of Waves and Particles*, New York: McGraw-Hill. A detailed overview of scattering theory.

Group Theory

M. Hamermesh
School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota, USA

Introduction ***190***
1 Elementary Definitions ***190***
1.1 Transformation Groups *190*
1.1.1 Subgroups *192*
1.1.2 Cosets *193*
1.1.3 Conjugate Classes *193*
1.1.4 Invariant Subgroups *194*
1.1.5 Isomorphic Groups *194*
1.1.6 Homomorphic Groups *195*
1.1.7 Factor Groups *195*
1.1.8 Direct Product *195*
1.1.9 Finite Groups *195*
1.1.10 Infinite Discrete Groups *196*
1.2 Continuous Groups *196*
1.3 Lie Groups *197*
2 Linear Representations of Lie Groups ***200***
2.1 Group Theory and Quantum Mechanics *200*
2.2 Construction of Representations *201*
2.2.1 Equivalent Representations *203*
2.2.2 Addition of Representations. Reducibility and Irreducibility *203*
2.2.3 Invariance of Functions and Operators *204*
2.3 General Theorems *205*
2.4 Kronecker Product of Representations *206*
2.5 Analysis of Representations *206*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

3 Applications *209*
3.1 Atomic Structure *209*
3.2 Nuclear Structure *210*
3.3 Solid State and Crystal-Field Theory *211*
4 Summary *212*
List of Works Cited *212*
Further Reading *212*

Introduction

Group theory is a mathematical technique for dealing with problems of symmetry. Such problems appear repeatedly in all branches of physics. The use of formal group theory is a very recent development, but notions of symmetry had been used extensively 1000 years ago. The design of ornaments with symmetries, the observation of periodic patterns, and the regular appearance of the Sun and other astronomical objects showed that symmetry was a useful concept. The first modern uses of symmetry were in crystallography. The first clear statement of the importance of symmetry was made by Pierre Curie around 1870. Since then group theory has become the principal tool for dealing with difficult problems in solid-state theory, relativity theory, atomic and nuclear spectroscopy, and the theory of elementary particles. In these problems we assert (or assume) that the laws describing the interactions of particles have some symmetry. In the simpler cases, such as the Coulomb field, the symmetry is easy to see. In nuclear physics the charge symmetry between neutrons and protons required a careful and bold extrapolation of experimental results. In elementary-particle physics we have very little clear understanding of the forces that determine the structure of the fundamental particles and their relations with one another. This has required us to assume some symmetry of the interactions, even though we know almost nothing about the details of the laws governing them. Once a symmetry is assumed we can then make predictions about the families of particles that are related to one another.

1 Elementary Definitions

1.1 Transformation Groups

We first give the definition of a transformation group, because these are the groups of direct importance for physics. For example, if we are considering the motion of a particle in a central field we realize that a rotation about any axis through the center will take a given orbit into another orbit with the same energy. So we want to study rotations around the center. Two successive rotations around the center again give a rotation to an orbit with the same energy. A reverse rotation would bring the orbit back to its original position. We see that the set of rotations of the three-dimensional space form a set of transformations that is closed under composition and contains all inverses. Now we give the rigorous definition. We have a

linear vector space on which a set of transformations act. The transformations form a group if

1. The "product" of any two elements a and b of the set is also a transformation in the set: $ab = c$. By product we mean the transformation c that results from the sequence of transformations b followed by a.
2. The product is associative, i.e., $((ab)c) = (a(bc))$. The product of a whole sequence of transformations gives the same final result if we split the sequence into any clumps that preserve the order.
3. The set of transformations includes the identity transformation e, that leaves all the coordinates unchanged.
4. If the set includes the transformation a, it must also contain the inverse transformation b, such that $ab = ba = e$. We usually write the inverse element of a as a^{-1} so that $aa^{-1} = a^{-1}a = e$.

In the example given above we note that in general, if we reverse the order of rotations a and b around different axes we get different results. In general the multiplication is not commutative, i.e., $ab \neq ba$. If the product of any two elements of the set is commutative ($ab = ba$ for all a, b in the set), we say that the group is *Abelian* (the group is a commutative group). In general the transformations of a group will not all commute with one another. Then we say that the group is non-Abelian. If we take the product of an element a of the group with itself, i.e., we form aa, we write the product as a^2, the "square" of a. Similarly we write successive products of a with itself as a^3, a^4, etc. – powers of a. If we perform repeated transformations using the inverse a^{-1} of a, we get the negative powers a^{-2}, a^{-3}, etc. The total number of distinct elements in the group is called the *order* g of the group. If the order of the group is finite we say that the group is a *finite group*. An infinite discrete group is one in which the distinct elements can be labeled by the integers. Often the elements of the group can only be labeled by one or more continuous parameters. Then we say that the group is a *continuous group*. If we talk about a group of objects with some product that satisfies our definitions we have an *abstract group*. We give some examples of groups.

The numbers 0,1,2,3 form a group if the product is addition modulo 4 : $2 + 1 = 1 + 2 = 3, 2 + 3 = 3 + 2 \cong 5 - 4 = 1, 2 + 2 \cong 4 - 4 = 0$, etc. The inverse of 2 is 2, the inverse of 1 is 3. This is the group Z_4, an Abelian group of order 4. Similarly the numbers $0, 1, \ldots, n-1$ give the Abelian group Z_n of order n. If we take all the integers $\ldots, -2, -1, 0, 1, 2, \ldots$ we get the infinite discrete Abelian group Z, in which the identity element is 0 and the inverse of s is $-s$.

Cyclic group. This is a group that consists of positive powers of some element a, for which there is a finite positive n such that $a^n = e$. Thus a^{n-1} is the inverse of a. So the group consists of $a, a^2, a^3, \ldots, a^n = e$. One example of a cyclic group is the set of rotations about some axis through angles $\theta = 2\pi m/n$ where n is a fixed integer and $m = 1, 2, \ldots, n$. This is the cyclic crystal group C_n. Here the product of rotations with $m = r$ and $m = s$ is the rotation through the angle $(2\pi/n)(r + s)$. This group C_n has the same abstract structure as Z_n. Clearly cyclic groups are Abelian.

An example of an infinite Abelian group is the group of translations along the x axis given by $x' = x + n\delta$, where δ is fixed and n is an integer, positive or negative or zero (which is the identity translation).

This group has the same abstract structure as the group Z.

Permutation groups. Another important class of groups comprises the permutation groups on n letters. In these groups each letter is replaced by some letter, while no two are replaced by the same letter. If each letter is replaced by itself, we get the identity permutation e. We first illustrate for the case of $n = 6$. A permutation of the numbers 1 to 6 replaces each number by one of the six numbers, and no two numbers are replaced by the same number. A simple notation for a permutation is

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ i_1 & i_2 & i_3 & i_4 & i_5 & i_6 \end{pmatrix},$$

where each number in the upper line is replaced by the number appearing below it. For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 2 & 4 & 5 & 3 & 1 \end{pmatrix}$$

replaces 1 by 6 ($1 \to 6$), $2 \to 2$, $3 \to 4$, $4 \to 5$, $5 \to 3$, $6 \to 1$. Note that $2 \to 2$, so 2 is unchanged by this permutation. There are $6! = 720$ different permutations. (If all the symbols are left unchanged we get the identity permutation with $1 \to 1$, $2 \to 2$, etc.). This same permutation can be written in *cyclic notation:* we start with any symbol, say 1. We record 1, and note that $1 \to 6$, so we write $1 \to 6$. $6 \to 1$, and we have a cycle (closed loop), so we put 1 and 6 in parentheses: (1 6). The number 2 is unchanged so we write (2). Next we start with 3 and find $3 \to 4 \to 5 \to 3$, giving the cycle (3 4 5) [which is the same as (4 5 3) or (5 3 4)]. Thus the original permutation is recorded as (16) (2) (3 4 5). The symbols inside a parenthesis are distinct. This particular permutation has the *cycle structure* [3, 2, 1]: it contains a cycle of three symbols, a cycle of two symbols, and a cycle of one symbol – a 3-cycle, a 2-cycle, and a 1-cycle. Often the permutation is written omitting its 1-cycles, so this would be (16) (345). Other examples are (134652) which consists of one 6-cycle, or (142)(365), with two 3-cycles (notation $[3^2]$), or (15)(26)(34) which contains three 2-cycles (notation $[2^3]$), or (135), where the three 1-cycles (unchanged symbols) are omitted, so its cycle structure is $[3\ 1^3]$. To find the inverse of a permutation expressed in cycle notation, we read the numbers from right to left, so the inverse of (142)(365) is (241)(563). The group that contains all the permutations on six symbols is called the symmetric group on six symbols S_6. We can form permutations on n letters and form the symmetric group S_n whose order is $n!$ This group S_n is very important in spectroscopy and especially when we consider identical (indistinguishable) particles, where the physical situation is unchanged when we make any permutation of the identical particles.

1.1.1 Subgroups

Suppose that we have a subset H of elements in the group $G (H \subset G)$. If the elements of H form a group under the same product law as in G, we say that H is a *subgroup* of G. For example, the group S_6 contains all the permutations on five letters (e.g., we drop all permutations containing the number 6, or those that omit any single one of the six letters). Thus S_6 contains six subgroups that have the same structure as S_5. S_6 also contains many differently appearing subgroups that have the same structure as S_4. (We omit permutations that contain some two of the numbers 1–6). Such subgroups are isomorphic to one another. Clearly the group G is a subgroup of G. Also the group element e,

the identity of G, is necessarily a subgroup. These two subgroups of G are said to be improper, while all other subgroups in G are called proper subgroups of G.

Suppose that H is a subgroup in G. If H does not exhaust G, take any element a not contained in H and form the set aH, which contains all elements of the form ah where h runs through H. The set aH contains no elements of H, since, if $ah = h'$ where h' is in H, applying h^{-1} on the right we would get $a = h'h^{-1}$, which says that a is a product of elements in H and is therefore in H. If the elements of G are not all included in H and aH, we choose one of the residual elements k of G and form kH. We continue this process until we have obtained all the elements of G. So G consists of the subsets of elements $H, aH, kH, \ldots, sH$, which have no elements in common and each contain m elements of G. We see then that the number m of elements in H (the order of H) must divide the order of G. This is called *Lagrange's theorem*. Note that the subgroups S_5 in S_6 have order $5! = 120$, which divides the order $6! = 720$ of S_6.

1.1.2 Cosets

The individual pieces in this decomposition of G are called the left cosets of H in G. We write this symbolically as $G = H + a_2H + a_3H + \cdots + a_rH$. If instead we formed the cosets by multiplying H from the right, giving disjoint sets Ha, we would get the right coset decomposition of G with respect to the subgroup H.

1.1.3 Conjugate Classes

If a and b are elements of the group G, the elements $bab^{-1} = c$, where b is any element of G, are said to be *conjugate* to a in G. We denote this by writing $c \sim a$. If we let b run through all the elements of G, we get the set of *transforms* of a, the *conjugate class* of a in G. The determination of transforms gives a decomposition of G into conjugate classes that have no elements in common, i.e., the conjugation operation is an equivalence relation:

1. $a \sim a$.
2. If $a \sim b$ and $b \sim c$, then $a \sim c$.
3. If $a \sim b$, then $b \sim a$.

We prove **2.** $a \sim b$ means that $a = kbk^{-1}$, where k is in G. Similarly $b \sim c$ means that $b = k'ck'^{-1}$, where k' is in G. So $a = kbk^{-1} = k(k'ck'^{-1})k^{-1} = (kk')c(k'^{-1}k^{-1}) = (kk') \times c(kk')^{-1}$, so $a \sim c$.

For geometrical transformations the conjugate has a simple interpretation. Suppose that a is a rotation through angle θ about a direction along the unit vector $\mathbf{n}$: $a = (\mathbf{n}, \theta)$. Let b be a rotation that turns the unit vector $\mathbf{n}$ into the direction of $\mathbf{n}'$. Then, since b turns $\mathbf{n}$ into $\mathbf{n}'$, in the conjugate bab^{-1}, b^{-1} first turns $\mathbf{n}'$ into $\mathbf{n}$, then a rotates about $\mathbf{n}$ through θ, then b brings $\mathbf{n}$ back to $\mathbf{n}'$, so the net result is a rotation through the angle θ about the unit vector $\mathbf{n}'$. Thus the conjugate of a rotation is always a rotation through the same angle θ about some other axis. Similarly, if a were a translation through a vector $\mathbf{s}$, the conjugate bab^{-1} would be a translation through the vector $\mathbf{s}'$, into which b takes $\mathbf{s}$.

For the permutation groups, the conjugate is also obtained simply. Suppose we want to find the conjugate c of the permutation a that results from taking $c = bab^{-1}$, where $a = (135)(24)$, $b = (325)$. Then $b^{-1} = (523)$, and we get $c = (325)(135)(24)(523) = (54)(123)$. We note that the conjugate of a has the same cycle structure as a; we get it by letting the permutation b act on the cycle symbol for a: apply (325) to a; it replaces 3 by 2, 2 by 5, and 5 by 3, so a is changed to

(123)(54), where 4 and 1 are untouched, since b does not contain 1 or 4. Another example would be the transform of $a = (142)(365)$ by $b = (16)(345)(2)$. Applying the permutation b to the cycle symbol for a replaces $1 \to 6 \to 1, 3 \to 4 \to 5 \to 3$, giving $bab^{-1} = (652)(413)$.

We emphasize that the transforming element b must be in the group G. If the group is Abelian, then all its elements commute, and the conjugate class of an element a is just a itself. If an element a commutes with all the elements of the group G, its class consists of a alone. This is always the case for the identity element e for any group – the identity is a class containing the single element e. Suppose we consider the permutation group S_3 on three letters 1, 2, 3. There are six elements in this group: e, (12),(23),(13); (123),(132). Only permutations with the same structure can be in the same conjugate class, so we have three classes: e; (12),(23),(13); (123),(132); with one, three, and two members, and cycle structures $[1^3]$, [2 1], and [3], respectively.

1.1.4 Invariant Subgroups

If H is a subgroup of G, we consider the set aHa^{-1}, containing all elements aha^{-1}, where h runs through the members of H. The elements of this set are all distinct, and the product of any two, $ah_1a^{-1}ah_2a^{-1} = ah_1h_2a^{-1}$, is also in aHa^{-1}, since h_1h_2 is in H. Clearly the set aHa^{-1} is a subgroup that looks like H in form. It is *a conjugate subgroup of H*. If all the conjugate subgroups of H in G are the same as H, we say that H is an invariant subgroup in G. But if $aHa^{-1} = H$ for all elements a, then multiplying on the right by a, we get $aH = Ha$, i.e., the sets of right and left cosets of H in G are identical. We also note that the invariant subgroup contains complete classes. Thus, in the example of the permutations on three letters, the group S_3 contains the proper subgroup e, (123),(132). Since it consists of complete classes, it is an invariant subgroup. For the permutation group on four letters S_4, the group contains $4! = 24$ elements. The different classes are e; type (12), type (123), type (1234), and type (12)(34), with numbers of elements 1, 6, 8, 6, 3, respectively. The classes correspond to the possible cycle structures given by the partitions of 4 into sums of positive integers. Thus e has four 1-cycles, partition $[1^4]$; (12) has one 2-cycle and two 1-cycles (partition $[21^2]$), (123) has one 3-cycle and one 1-cycle, (partition [31]), (1234) has one 4-cycle (partition [4]), and (12)(34) has two 2-cycles (partition $[2^2]$).

As another example, we consider the group of all rotations and translations in a space with dimension m. We denote the elements by $(R|a)$, where this means that we first perform the rotation R and then the translation a. The translations a form a subgroup (Abelian), and we showed earlier that the transform of any a by a rotation R is again a translation. Thus the subgroup of the translations is invariant in the whole group.

1.1.5 Isomorphic Groups

Two groups G and G' are said to be *isomorphic* if their elements can be put into a one-to-one correspondence so that if $a \leftrightarrow a'$ and $b \leftrightarrow b'$ then $ab \leftrightarrow a'b'$. For example, consider the group consisting of the identity e and the inversion of coordinates i. Here $i^2 = e$, so the group contains two elements. Now look at the group consisting of the numbers 1 and -1 with the product being ordinary multiplication, so that $(-1)(-1) = 1$. We see that these groups can be put into one-to-one correspondence $e \leftrightarrow 1, i \leftrightarrow -1$, with $ii = e$ corresponding to $(-1)(-1) = 1$. There can be many

groups with the same abstract structure. They are isomorphic to one another.

1.1.6 Homomorphic Groups

Instead of a one-to-one correspondence, we may have two groups with a correspondence of products, but where several elements of the group G correspond to a single element of G'. We say that there is a *homomorphism* of the group G onto G'. For example, in the cyclic group of order 4, generated by an element a with $a^4 = e$, the elements a^2 and $a^4 = e$ are a subgroup H in G. If we take the odd powers a and a^3 and map them onto the number -1, and map the subgroup H onto 1, we have a two-to-one correspondence between the cyclic group and the two-element group.

1.1.7 Factor Groups

If the subgroup H is an invariant subgroup in G, we can write G as a "sum" of cosets of H in $G : G = H \dot{+} a_2 H \dot{+} a_3 H \dot{+} \cdots \dot{+} a_r H$, where $a_i H = H a_i$. But then the product of cosets $(a_i H)(a_j H) = a_i (H a_j) H = a_i (a_j H) H = (a_i a_j) H$, so the cosets themselves form a group, for which the identity is the coset H itself. This new group is called the factor group F of G by the (invariant) subgroup H. There is a homomorphism of the group G onto the subgroup H where all the elements in a coset are mapped on the same element of the factor group $F = G/H$.

1.1.8 Direct Product

If we have two groups, G and H, and the elements of one commute with all the elements of the other, we can form a new group, their direct product, by taking pairs (g, h) of elements g from G and h from H. The product of two elements (g_1, h_1) and (g_2, h_2) in this new group is the ordered pair $(g_1 g_2, h_1 h_2)$. This group is called the direct product $G \otimes H$ of the two groups G and H. If we have a number of groups $G, H, K, \ldots$, and the elements of any one commute with all the elements of the others, we can form the direct product $G \otimes H \otimes K \otimes \ldots$, whose elements are obtained by taking one element from each group: $(g, h, k, \ldots)$. Clearly all the elements $(g, e_i, e_j, \ldots)$, where the $e_i, e_j, \ldots$ are the identity elements for the other factors, form a group isomorphic to G. This group can be identified with G and is an invariant subgroup of the direct product. The direct product is thus a product of the groups, $G, H, K \ldots$, and each of them is an invariant subgroup of the direct product.

For the crystal group that we described earlier, with elements $(R|a)$, the translations are an invariant subgroup, since RaR^{-1} is again a translation, while the subgroup of rotations is not invariant. Instead the rotations act on the translations and transform the translation vectors. Such a group is called the semidirect product of the groups R and A, where R contains all the pure rotations and A all the pure translations, and is written as $A \otimes R$, where the invariant subgroup is the first symbol.

1.1.9 Finite Groups

For a finite group one can describe the group structure by recording the Cayley table for the group. We write the elements of the group as a matrix with each column headed by an element of the group, and similarly, each row. We then get a square matrix of products by placing the product $g_i g_j$ at the i, j position.

The group formed by taking all products of r and s, where $r^3 = s^2 = (rs)^2 = e$, is a group with six elements. Its Cayley table is shown in Table 1. Note that each row contains all six elements of the group, but they are rearranged. In other words, applying an element of the group from the

Tab. 1 A Cayley Table

e	r	r^2	s	sr	sr^2
r	r^2	e	sr^2	s	sr
r^2	e	r	sr	sr^2	s
s	sr	sr^2	e	r	r^2
sr	sr^2	s	r^2	e	r
sr^2	s	sr	r	r^2	e

left permutes the elements. We thus have a group of permutations that is isomorphic to the group in the table.

This is a group of permutations of six elements, and is a subgroup of the symmetric group S_6. In the same way, every finite group of order n is isomorphic to a subgroup of the symmetry group S_n. This result is called *Cayley's theorem*. An important collection of finite groups consists of the so-called point groups: rotations through submultiples of 2π that keep the origin fixed and reflections in planes passing through the origin (or inversion in the origin). These finite groups describe the point symmetry of crystals.

1.1.10 Infinite Discrete Groups

In three dimensions the translations through a displacement $\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$, where the n_i are integers and the vectors $\mathbf{a}_i$ are fixed, form an infinite discrete group. This is the translation group of a crystal lattice. Similar groups can be constructed for any space dimension. We can also have crystal space groups, that contain elements $(R|a)$ that preserve the crystal lattice. If the translations appearing in $(R|a)$ are always lattice vectors the crystal group is said to be *symmorphic*. There are some crystals in which there are additional symmetries which involve elements $(R|a)$ where the translation is some rational fraction of a lattice vector along the direction of the rotation axis, giving a screw motion. Or we may have glide motions where we translate parallel to some crystal plane through a rational fraction of a lattice vector and then reflect in the crystal plane. Crystals that have such symmetry elements are said to be *nonsymmorphic*. [For further information about space groups see Burns 1977, Chap. 11.] More detailed treatments of Sec. 1.1 can be found in Burns (1977), Chaps. 1, 2, and 13; Elliott and Dauber, (1979), Chaps. 1 and 2; and Hamermesh (1989), Chaps. 1 and 2.

1.2 Continuous Groups

When we consider continuous groups we combine the algebraic concept of "group" with the topological concept of "nearness." The law of combination of elements a, b of G now requires the product ab to depend continuously on its factors, and the inverse a^{-1} must depend continuously on a. We shall not discuss general topological groups but will assume that there is a metric (a measure of distance between group elements) on the group. So the group itself is a metric space on which the product and inverse are defined and continuous. We can look upon the group as a space in which the points are the elements of the group. We shall deal only with groups of transformations, so the group elements form a space whose points are the transformations of the group. Multiplying the element b on the left by some transformation a of the group moves the point b in the group space to the point ab. Thus the group elements can be regarded as operators that act on the group space itself and rearrange the points. The changed points also fill the space since any element of

the group space c can be reached from b by applying the element cb^{-1} to it. Thus the group space looks the same everywhere. It is a homogeneous space, and we can obtain all our information by, e.g., working in the neighborhood of the identity e.

A group is said to be a mixed continuous group if the elements depend on discrete labels as well as continuous ones. For example, the translation group on the line G is the set of transformation $x' = x + a$, $(-\infty \leq a \leq \infty)$. If we adjoin the single transformation $x' = -x$ we also get the transformations $x' = -x + a$. The transformations $x' = x + a$ are in one piece and can be reached continuously from the identity. Thus this piece forms a group. The second piece, containing the transformations $x' = -x + a$, cannot be reached continuously from the identity, and is not a group. This piece is the coset of the first piece, obtained by multiplying elements of the first piece by the transformation $x' = -x$.

A *connected group* is a continuous group in which we can join any two points of the group by a continuous arc in the group space. Our last example showed a group consisting of two pieces (components). The piece that contains the identity is called the component of the identity. If a group is connected it consists of a single component. One basic theorem is the following: In a connected group G, every neighborhood of the identity generates the group. By taking products of elements near the identity we keep expanding the set of products and fill the whole connected piece containing the identity. For example, in the group of translations $x' = x + a$, if we take some neighborhood of the identity $a = 0$, say $-\varepsilon \leq a \leq \varepsilon$, and keep applying these translations, we get the whole line.

1.3 Lie Groups

A Lie group is a group in which the neighborhood of the identity looks like (is homomorphic to) a piece of an r-dimensional Euclidean space (an r-parameter Lie group). In other words, in the neighborhood of the identity the group elements can be parametrized using r parameters (coordinates on the group manifold). Thus each element can be labeled by r parameters $a_1, a_2, \ldots, a_r$. Since we want continuity, this parametrization may not be possible over the whole group space, but we can use several maps that overlap to cover the whole space without singularities. (For example, to cover a sphere with continuous maps requires more than one map.) An element of the group is $R(a) = R(a_1, \ldots, a_r)$ where the parameters $a_1, \ldots, a_r$ are essential, i.e., all r parameters are needed to describe the space. The product of two elements with parameters a and b, respectively, $R(a)R(b)$, is the element $R(c)$ with parameters $c_1, \ldots, c_r$ that are functions of $a_1, \ldots, a_r$ and $b_1, \ldots, b_2$, i.e.,

$$c_i = \varphi_i(a_1, \ldots, a_r; b_1, \ldots, b_r), \text{ or, symbolically,}$$
$$c = \varphi(a; b). \tag{1}$$

The simplest assumption is that the functions φ are analytic functions of the a's and b's. This was the original requirement for a Lie group. But this requirement is too strong. In fact, every parametric group is a Lie group. The group structure combined with continuity implies the analyticity of φ (Hilbert's "Fifth Problem").

For Lie groups with a finite number of parameters, the neighborhood of any point is bounded and contains all its limit points (the group is *locally compact*). If the parameter space is unbounded this

may not be true as we move out to large distances in the parameter space. For most of the groups used in physics the group space is compact, but we shall see some cases of noncompact groups.

Consider transformations in n-space:

$$x_i' = f_i(x_1, \ldots, x_n), \qquad i = 1, \ldots, n, \tag{2}$$

or, symbolically, $x' = f(x)$. Suppose that we have a set of f's labeled by r parameters:

$$x_i' = f_i(x_1, \ldots, x_n; a_1, \ldots, a_r), \tag{3}$$

forming a Lie group. The transformation a is $x \to x' = f(x; a)$, and b is $x' \to x'' = f(x'; b)$. If we first apply a and then b, we get

$$\begin{aligned} x'' &= f(x'; b) = f(f(x; a); b) \\ &= f(x; \varphi(a, b)) = f(x; c), \end{aligned} \tag{4}$$

where $c = \varphi(a, b)$ and φ is an analytic function of the parameters a and b. We give examples of Lie groups that comprise most of the groups of interest to physics.

1. The dilation group. In one dimension this is the group of transformations $x' = ax$, $a \neq 0$. The identity has $a = 1$. All distances on the line are multiplied by the number a. The transformation with $a = -1$ reflects the line about the origin. The parameter a for the inverse is $\bar{a} = 1/a$, and $c = \varphi(a, b) = ab$. The group space is the real line with the origin cut out, $\mathbb{R}\backslash\{0\}$. It consists of the two separate pieces $x \gtrless 0$. The group space for the dilations in two dimensions, $x' = ax$, $y' = ay$, is the real plane with the origin removed (the punctured plane $\mathbb{R}^2\backslash\{0\}$). In this case the group space remains a single connected piece. It is convenient to assign the parameter 0 to the identity. We write the group as $x' = (1 + \alpha)x$, with$\alpha \neq -1$.
The group $x' = ax$, $y' = a^2 y$, with $a \neq 0$, has only one parameter, but now the transformations are in a space of dimension 2. We note that the number of parameters is not determined by the dimension of the space on which the transformations act.
2. GL(n), the general linear group in n dimensions, is the set of nonsingular linear transformations in n-dimensional space. For real entries we write GL(n, R), for complex entries GL(n, C): $x' = Ax$, or $x_i' = a_{ij}x_j (i = 1, \ldots, n)$, where we sum over the repeated index, and det $A \neq 0$ (the nonsingular transformations). The number of real parameters is n^2 for GL(n, R), and $2n^2$ for GL(n, C). $\overline{A} = A^{-1}$, $\varphi(A, B) = BA$. The elements a_{ij} vary over an infinite range ($-\infty \leq a_{ij} \leq \infty$), so the group is not compact. The elements of GL(n, C) are the subset of the $n \times n$ complex matrices with determinant $\neq 0$.
3. SL(n) is the special linear group (unimodular group) with det $= 1$. SL(n, C) and SL(n, R) are subgroups of GL(n, C). Thus SL(2) is the collection of 2×2 matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ with } ad - bc = 1.$$

4. O(n) is the orthogonal group, a subgroup of GL(n), where we impose the requirement that the transformations leave scalar products of any two vectors x, y unchanged; i.e., $(Ox, Oy) = (x, y)$, where the scalar product is

$$(x, y) = \sum_i x_i y_i.$$

Then $(x, y) = (Ox, Oy) = (x, \widetilde{O}Oy)$, where $\widetilde{O}$ is the transpose of the matrix

O, so that

$$\widetilde{O}O = 1, \text{ the unit matrix, or}$$
$$\widetilde{O}_{ij}O_{jk} = \delta_{ik} = O_{ji}O_{jk}. \qquad (5)$$

Taking determinants, $(\det O)^2 = 1$, or $\det O = \pm 1$. The column vectors of the orthogonal matrix are orthogonal unit vectors. Similarly $O\widetilde{O} = 1$, so the rows are orthogonal unit vectors. For $O(n, R)$, there are n conditions from the norm and $n(n-1)/2$ from the orthogonality, so the number of parameters is $r = n(n-1)/2$. The subgroup with positive determinant $+1$ is called $O^+(n, R)$ or $\mathrm{SO}(n, R)$, the special orthogonal group (or proper orthogonal group).

5. $U(n)$ is the unitary group that acts in a complex unitary space, where we have an inner product $(x, y) = \Sigma_i x_i^* y_i$, and the asterisk denotes the complex conjugate, so that $(x, y) = (Ux, Uy) = (x, U^\dagger Uy)$, where $U^\dagger = \widetilde{U}^*$ is the adjoint of U, so $U^\dagger U = 1 = UU^\dagger$, $\Sigma_i U_{ij}^* U_{ik} = \delta_{jk}$, and $|\det U|^2 = 1$, so $\det U = \exp(i\varepsilon)$. Note that for $j = k$ there are n conditions $\Sigma_i |U_{ij}|^2 = 1$, while for $j \neq k$ there are $2n(n-1)/2$ orthogonality conditions (real and imaginary parts both equal to zero). Thus the number of parameters is $r = n^2$. The sum of the absolute squares of all the matrix entries is n, so the group space is bounded and $U(n)$ is a compact group.
6. $SU(n)$ is the unitary unimodular group (special unitary group), a subgroup of $U(n)$ having $\det U = 1$, so $r = n^2 - 1$.
7. The *Euclidean group E(n)* in real space, $x' = Ax + a$, where A is an orthogonal matrix and **a** is a translation vector, preserves the Euclidean distance, so $\widetilde{A}A = 1$, and the number of parameters is $r = n(n+1)/2$. For $n = 3, r = 6$. This group $E(3)$ can be regarded as a group of block matrices (A, a), where A is a 3×3 orthogonal matrix, O is the (1×3) null matrix,

$$\begin{pmatrix} & a_1 \\ A & a_2 \\ & a_3 \\ O & 1 \end{pmatrix},$$

and (A, a) acts on the column vectors

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}.$$

This group describes the rigid-body displacements in 3-space.

In obtaining $O(n, R)$ we required the invariance of the scalar product (x, y). If we had required the invariance of any positive definite symmetric bilinear form $g_{ij}x_i y_j = (x, gy)$, with $g = \widetilde{g}$, we could bring it back to the unit form by a change of basis. Thus the group $O(n, R)$ is obtained for all of these. If we extend this argument to include all nonsingular symmetric bilinear forms, a change of basis will bring them (Sylvester's law of inertia) to diagonal form with p 1's and $q(-1)$'s, where $p + q = n$. We get the pseudo-orthogonal groups $O(p,q)$, defined over the real numbers. If we define a matrix

$$S_{pq} = \begin{pmatrix} 1_p & 0 \\ 0 & -1_q \end{pmatrix},$$

where the 1_p and 1_q are the unit matrices for dimension p and q, these groups are defined by the requirement that $(x, sy) = (Ox, sOy) = (x, \widetilde{O}sOy)$ for arbitrary x and y, so that $\widetilde{O}sO = s$. Again we have $(\det O)^2 = 1$, so by choosing those transformations with $\det = 1$, we can define the special groups $\mathrm{SO}(n-q, q)$. An important example of this kind of group is the *Lorentz group*, where we require

the invariance of $x^2 + y^2 + z^2 - \tau^2$ (where $\tau = ct$) under the group.

What about skew-symmetric bilinear forms? Given a nondegenerate skew-symmetric bilinear form, with

$$\{x, y\} = g_{ij}x_i y_j \quad \text{and} \quad g_{ij} = -g_{ji}, \tag{6}$$

$g = -\widetilde{g}$, $\det g = (-1)^m \det g$, so the form will be nondegenerate only for even dimension m. If we want the transformations to preserve $\{x, y\}$, we must have $\widetilde{O}gO = g$, so $(\det O)^2 = 1$. By a change of basis we can bring g to the canonical form

$$g = \begin{pmatrix} 0 & 1_n \\ -1_n & 0 \end{pmatrix},$$

where O is the $n \times n$ null matrix and 1_n is the $n \times n$ unit matrix. These groups are the symplectic groups $\mathrm{Sp}(2n)$. For further details and examples, see Hamermesh (1989), p. 283ff.

2 Linear Representations of Lie Groups

In Sec. 1 we gave an introduction to group theory and a description of some of the important transformation groups that have application to physics. Now we want to show how groups can be used in the solution of physical problems.

2.1 Group Theory and Quantum Mechanics

In quantum mechanics the states of a system of n particles are described by wave functions (state vectors) Ψ that are functions of the time t, the coordinate vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ of the n-particle system, and any other internal coordinates of the particles. The changes of the system with time (the equations of motion) are governed by the Hamiltonian operator $H = T + V$, where T is the kinetic energy and V is the potential energy of the system, which includes the interaction of the particles with one another and with any external fields. V is a function of the coordinates, while T involves both the coordinates and derivatives with respect to the coordinates. The time behavior of the system will be given by the Schrödinger equation:

$$H\Psi = -\frac{\hbar}{i}\frac{\delta\Psi}{\partial t}, \tag{7}$$

where $\hbar = (1/2\pi) \times$ (the Planck constant h). For most problems we look for the stationary states, i.e., states whose time dependence is given by an exponential factor $\exp(-iEt/\hbar)$:

$$\Psi(r_1, \ldots, r_n; t) = \psi(r_1, \ldots, r_n) \times \exp\left(\frac{-iEt}{\hbar}\right). \tag{8a}$$

E is the energy of this stationary state (energy eigenstate) and $H\psi = E\psi$. A particular problem will have many different solutions ψ_i with corresponding energies E_i. The set of all the eigenvalues E_i gives the spectrum of energy levels of the system. If there are several independent state vectors for a given energy level, we say that the state is degenerate. The theory is linear: we can superpose any states ψ_i to get a new state $\psi = \alpha\psi_r + \beta\psi_s$, where α and β are arbitrary complex numbers. If we assume that the state at time $t = 0$ is given by

$$\psi = \sum_{i=1}^{m} c_i\psi_i,$$

then at time t, it will have developed into the function

$$\psi(t) = \sum_{i=1}^{m} c_i\psi_i(r_1, \ldots, r_n)\exp\left(\frac{-iE_it}{\hbar}\right). \tag{8b}$$

The quantities that are measured for a state ψ (the observables, such as position of the nth particle, the dipole moment of the system, etc.) are operators that act on the state vector, and in general change it to a new state. If the operator S leaves the state ψ unchanged except for a complex factor, so that $S\psi = s\psi$, we say that the state is an eigenstate of the operator with eigenvalue s. If we measure S on a state $\varphi = \Sigma_i b_i \varphi_i$, where the φ_i are normalized eigenstates of the operator S, the result will be one of the s_i, with a probability $|b_i|^2$, and the average value over many measurements (the expectation value of S in this state) will be

$$\langle S \rangle = |(\varphi, S\varphi)|^2 = \sum_i |b_i|^2 s_i. \qquad (9)$$

In addition to observables we can introduce other operators that act on the states of the system. For example, if the coordinates are changed by some transformation, there will be a corresponding change in the state vector $\psi(r)$. The new state vector ψ' will be obtained by acting on the state vector ψ with an operator related to the coordinate transformation. Some coordinate transformations may leave the physical situation unchanged. For example, for a particle in a central field, a rotation R about any axis through the center has no apparent effect on the physics, since the field depends only on the radius vector to the particle. The operator O_R corresponding to the rotation R acts on the state vector ψ to give a new state $\psi' = O_R\psi$ that looks the same in the rotated coordinates $r' = Rr$ as did the original ψ in terms of the coordinates r:

$$O_R\psi(r') = \psi(r),$$

$$O_R\psi(Rr) = \psi(r),$$

or

$$O_R\psi(r) = \psi(R^{-1}r). \qquad (10a)$$

If the state ψ is an eigenstate of the Hamiltonian H with eigenvalue E, $H\psi = E\psi$, $O_R\psi$ will also be an eigenstate of H with the same energy E, i.e., $HO_R\psi = EO_R\psi$, so the states ψ and $O_R\psi$ will be degenerate (or differ by a phase). Applying the operator O_R^{-1} on the left, we get

$$O_R^{-1}HO_R\psi = E\psi = H\psi. \qquad (10b)$$

Since this equation holds for any eigenstate, it holds for all states ψ, so the operators on both sides coincide:

$$O_R^{-1}HO_R = H; \qquad HO_R = O_R H. \qquad (10c)$$

Thus the operator O_R commutes with the Hamiltonian H. It is a symmetry operator for this Hamiltonian. If we choose some other rotation S we get a corresponding symmetry operator O_S that acts on the state vectors. The product of the operators $O_S O_R$ is the operator O_{SR} that corresponds to the result of successive transformations by R and then by S. In general, if we have a collection of symmetry operators their products and inverses will also be symmetry operators. The identity operator is a symmetry operator and the operators satisfy the associative law. Thus the operators obtained by using Eq. (10a) for all the symmetry transformations will form a group of operators that is isomorphic to the group of symmetry transformations. The group of operators acting on some space of functions provides us with a representation of the symmetry group.

2.2 Construction of Representations

As a simple example of the construction of representations, we again consider the problem of a particle in a central field. The group of symmetry operators O_R will take

an eigenfunction ψ into an eigenfunction $O_R\psi$ with the same energy E. If we apply all the rotation operators to the same ψ we will not always get a linearly independent $O_R\psi$ as the result. For example, if our $\psi(x, y, z)$ depends only on the radius r, $O_R\psi$ will leave ψ unchanged, since this state vector is spherically symmetric. If $\psi(x, y, z) = x$, the rotation operators will give only x, y, z or linear combinations of them. So this will be a representation in a three-dimensional carrier space, and we say that this is a three-dimensional representation of the rotation group. Calling these basis functions $\psi_i^{(1)} (i = 1, 2, 3)$, $O_R\psi_i^{(1)}$ will be a linear combination of $\psi_i^{(1)}$:

$$O_R\psi_i^{(1)} = \sum_{j=1}^{3} D_{ji}^{(1)}(R)\psi_j^{(1)}, \quad (i = 1, 2, 3), \tag{11}$$

where the superscript (1) labels the representation $D^{(1)}$, and $D_{ij}^{(1)}(R)$ is the matrix representative of the rotation in this representation. If we perform a second rotation S, we get

$$\begin{aligned} O_s(O_r\psi_i^{(1)}) &= \sum_j D_{ji}^{(1)}(R)O_s\psi_j^{(1)} \\ &= \sum_{k,j} D_{ji}^{(1)}(R)D_{kj}^{(1)}(R)\psi_k^{(1)} \\ &= \sum_k \left[\sum_j D_{kj}^{(1)}(R)D_{ji}^{(1)}(R)\right]\psi_k^{(1)} \\ &= \sum_k D_{ki}^{(1)}(SR)\psi_k^{(1)}. \end{aligned} \tag{12}$$

The matrices $D^{(1)}$ thus give a three-dimensional representation of the rotation group. If we take any linear subspace of this three-dimensional space, applying the rotation operators will automatically generate the rest of the three-dimensional space. There is no proper subspace that is invariant, i.e., is reproduced by the rotation operators. We say that the representation $D^{(1)}$ is an *irreducible* three-dimensional representation of the rotation group.

If we start with quadratic expressions in x, y, z: $xy, yz, zx, x^2, y^2, z^2$, we will obtain a six-dimensional carrier space on which the rotation operators will act, giving us a six-dimensional representation D. But this representation is *reducible:* we can find a proper subspace which is invariant. It is the one-dimensional subspace of the multiples of the function $\psi_{11}^{(0)} = x^2 + y^2 + z^2$, which is unchanged by rotations (orthogonal transformations). The remaining five-dimensional space with basis $xy, yz, zx, x^2 - z^2, y^2 - z^2$, contains no invariant subspace, so the matrices obtained by applying the rotation operators O_R to this subspace will give an irreducible five-dimensional representation of the rotation group. Thus by changing the basis we can decompose the representation into two independent irreducible representations.

In general, for any transformation group G with elements R, starting with some functions ψ, we can apply the corresponding operators O_R to the ψ's, find an independent basis set, and obtain an equation just like Eq. (11):

$$O_R\psi_i^{(\mu)} = \sum_{j=1}^{d_\mu} D_{ji}^{(\mu)}(R)\psi_j^{(\mu)}, \quad (i, j = 1, \ldots, d_\mu), \tag{13}$$

where μ labels the representation $D^{(\mu)}$, with dimension d_μ.

We mention again the two important steps:

1. We start with some functions and find the space of functions generated by the operator group.

2. We make changes of basis and look for invariant subspaces. If there are no invariant subspaces, the representation is irreducible. The basis functions must be degenerate with one another. If there are invariant subspaces, we check these separately. We finally end up with a collection of irreducible representations. The basis functions of each of these form a necessarily degenerate set.

In describing this process, it is convenient in general to use a "coordinate space" of dimension $d_\mu = m$, and call the basis set $y_1, \ldots, y_m$. The transformations R of the group G are represented by operators O_R that act on the coordinates $(y_1, \ldots, y_m)$ or by their matrix representatives $D^{(\mu)}(R)$.

An important method for constructing representations for transformation groups is to realize that the group supplies us with a representation of itself: each element R of the group is its own representative $D(R)$. If n is the dimension of the coordinate space on which the group acts, the representatives will be $n \times n$ matrices $D^{(1)}(R)$. This is the vector representation of the group G. Any quantity that transforms according to this representation is called a vector relative to the group G. The product of representatives of R and S is $D^{(1)}(RS) = D^{(1)}(R)D^{(1)}(S)$. We also sometimes refer to the basis of this representation as a tensor of rank 1 (relative to G). We can define tensors of higher rank relative to G as follows: If x and y are vectors we can form the n^2 products $x_i y_j$. Under the transformations of G these will transform so that $x'_i y'_j = D^{(1)}_{ik}(R) D^{(1)}_{jl}(R) x_k y_l$, where we always use the convention of summing over any repeated index. (Note that the same transformation R is applied to each factor.) Any set of quantities T_{ij} that transform like this product is called a second-rank tensor: $T'_{ij} = D^{(1)}_{ik}(R) D^{(1)}_{jl}(R) T_{kl}$. Similarly we can construct tensors of any rank:

$$T'_{i_1,\ldots,i_n} = D^{(1)}_{i_1 j_1}(R) D^{(1)}_{i_2 j_2}(R) \cdots \cdots D^{(1)}_{i_m j_m}(R) T_{j_1}, \ldots, j_m. \quad (14)$$

These representations will be reducible, but can be reduced by the methods described in Hamermesh (1989), Chap. 10.

2.2.1 Equivalent Representations

If we have a representation $D(R)$ of the group G and make a transformation of basis in the space of the representation, using some matrix P, we can form matrices $PD(R)P^{-1} = D'(R)$, and it is easy to verify that these also form a representation of G with the same dimension as D. We also see that $D'(E) = 1$, and $D'(R)D'(S) = D'(RS)$, since $PD(R)P^{-1} \times PD(S)P^{-1} = PD(R)D(S)P^{-1} = PD(RS)P^{-1}$. We say that the two representations D and D' are equivalent. They are isomorphic and represent only a change of basis. Equivalent representations have the same set of eigenvalues. The diagonal sum (trace) is the same for $D(R)$ and $D'(R)$. We call this quantity which is the same for any two equivalent representations the character $\chi(R)$ of R in the representation D: $\mathrm{Tr}\, D' = D'_{ii} = D_{ii} = \mathrm{Tr}\, D$. If two elements R and S of the group G are in the same conjugate class, so that there is an element P of the group such that $S = PRP^{-1}$, then $D(S) = D(P)D(R)D(P^{-1})$ and, taking the trace, we get $\chi(S) = \chi(R)$. Thus all elements in the same class in G have the same character.

2.2.2 Addition of Representations. Reducibility and Irreducibility

Suppose that we have two representations $D^{(1)}$ and $D^{(2)}$ of the same group G, where

$D^{(1)}$ has dimension n_1 and acts on coordinates $x_i, i = 1, \ldots, n_1$, and $D^{(2)}$ has dimension n_2 and acts on some other space with coordinates $x_i, i = n_1 + 1, \ldots, n_1 + n_2$. Thus the matrices $D^{(1)}(R)$ are $n_1 \times n_1$ matrices and the $D^{(2)}(R)$ are $n_2 \times n_2$. For each R in G, we construct a new representation D that has dimension $n_1 + n_2$,

$$D(R) = \begin{pmatrix} D^{(1)}(R) & 0 \\ 0 & D^{(2)}(R) \end{pmatrix}$$

and acts on the column vectors

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{n_1} \\ x_{n_1+1} \\ \vdots \\ x_{n_1+n_2} \end{bmatrix}$$

and $D(R) = D^{(1)}(R) + D^{(2)}(R)$. The two representations act separately in the two subspaces. If we make a change of basis in the $(n_1 + n_2)$-dimensional space so that the x's mix, the equivalent representation will no longer have this block form, and it will not be evident that the representation is actually a "sum" of two independent representations. Now we proceed in the reverse directions. Given a representation of the group in some space of dimension N, we ask whether it is possible to find a simultaneous decomposition of the matrices for all R into invariant subspaces of dimensions n_1, n_2 such that $n_1 + n_2 = N$, so that $D(R)$ acting on vectors in subspace 1 gives vectors in that subspace, and similarly for subspace 2. If this is true for all R in G we say that the representation D of dimension N is reducible and has been decomposed into the sum of $D^{(1)}$ and $D^{(2)}$. Next we try to find decompositions of the subspaces 1 and 2 into invariant subspaces. This process must end in a finite number of steps, and we finally have a decomposition of the original representation space into a sum of subspaces, and $D = D^{(1)} + D^{(2)} + \cdots + D^{(r)}$, where none of the representations $D^{(i)}$ has an invariant subspace. Then we say that the original representation $D(R)$ has been *fully reduced*. If the carrier space of D contains no proper invariant subspace, we say that the representation is irreducible. Note that the character of the representation $D(R)$ is the sum of the characters of its irreducible component representations: $\chi(R) = \Sigma_i \chi^{(i)}(R)$. The same irreducible representation can occur several times in the decomposition, so that, in general, $D = \Sigma_i a_i D^{(i)}$, and $\chi = \Sigma_i a_i \chi^{(i)}$, where a_i are positive integers.

2.2.3 Invariance of Functions and Operators

We found earlier that $O_R\psi(x) = \psi(R^{-1}x)$, so $O_R\psi$ is not the same as ψ. If $O_R\psi$ is the same as ψ, so that $\psi(Rx) = \psi(x)$, we see that the function ψ is invariant under the transformations. To test for invariance we replace the arguments x of any function $\psi(x)$ by Rx and see whether we get the same expression. If an operator T acts on a function ψ we get a function $\varphi = T\psi$, and applying the operator O_R to $T\psi$ we get

$$\begin{aligned} O_R T\psi(x) &= \varphi(R^{-1}x) \\ &= T(R^{-1}x)\psi(R^{-1}x), \end{aligned} \tag{15}$$

$$\begin{aligned} O_R T(x) O_R^{-1} O_R \psi(x) &= T(R^{-1}x) O_R \psi(x) \\ &= T'(x) O_R \psi(x), \end{aligned} \tag{16}$$

where $T'(x) = T(R^{-1}x)$. In general the operators T and T' are different. If they

are the same, i.e., if

$$O_R T(x) O_R^{-1} = T(x), \qquad (17)$$

the operator T is invariant under the transformation. In other words the operator T commutes with O_R. For example, the function $x^2 + y^2$ is invariant under rotations around the z axis, and also under inversion. The operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$ is also invariant under these transformations. A function describing the state of a system of identical particles is invariant under any permutation of the particles.

Now we can see the connection of group theory with physics. The Hamiltonian H of a physical problem may be invariant under some operators T. The collection of these operators and all possible products form a group G – the symmetry group of the Hamiltonian H. The basis functions for an irreducible representation of this symmetry group must be transformed into vectors in this same space by the operators $D^{(\mu)}(R)$. This m-dimensional space provides us with an m-fold degeneracy of the particular energy level. For a reducible representation of the symmetry group G more states seem to be transformed among themselves, but we can reduce the representation by finding the irreducible components. The basis functions for an irreducible representation *must* transform into one another under the operations of the symmetry group. It may happen that a problem gives degeneracies that are greater than expected from the assumed symmetry group. We must then search for some symmetry operation beyond those assumed. Often this occurrence is labeled as "accidental" degeneracy. Note that the basis function $\psi_i^{(\mu)}$ for the irreducible representation $D^{(\mu)}$ is said to belong to the ith row of the representation.

2.3 General Theorems

We now list without proof the general theorems that enable us to reduce any representation of a group G into its irreducible constituents.

Schur Lemma 1. If D and D' are two irreducible representations of a group G, having different dimensions, and the matrix T satisfies $D(R)T = TD'(R)$ for all R in G, then the matrix T must be the null matrix $T = 0$.

Schur Lemma 1a. If D and D' have the same dimensions and are irreducible representations of G, and if $D(R)T = TD'(R)$ for all R, then either D and D' are equivalent or $T = 0$.

Schur Lemma 2. If the matrices $D(R)$ are an irreducible representation of G and if $TD(R) = D(R)T$ for all R, then T is a multiple of the unit matrix: $T =$ const 1. This lemma gives an easy test of irreducibility.

Next we present the orthogonality relations. The quantities $D_{ij}^{(\mu)}(R)$ for fixed μ, i, j form a vector (whose components are labeled by R) in a g-dimensional space (where g is the order of G). If $D^{(\mu)}$ and $D^{(\nu)}$ are two nonequivalent irreducible representations of G, then

$$\sum_R D_{il}^{(\mu)}(R) D_{mj}^{(\nu)}(R^{-1}) = \frac{g}{n_\mu} \delta_{\mu\nu} \delta_{ij} \delta_{lm}, \qquad (18)$$

where n_μ is the dimension of the representation $D^{(\mu)}$. If $D^{(\mu)}$ and $D^{(\nu)}$ are the same D, then

$$\sum_R D_{il}(R) D_{mj}(R^{-1}) = \frac{g}{n} \delta_{ij} \delta_{lm}. \qquad (19)$$

If the representations are unitary, we replace $D_{mj}(R^{-1})$ by $D_{jm}^*(R)$. Thus each irreducible representation $D^{(\mu)}$ gives us n_μ^2 vectors $D_{ij}^{(\mu)}(R)(i, j = 1, \ldots, n_\mu)$ that

are orthogonal to one another and to all such vectors formed from nonequivalent representations. Since the number of orthogonal vectors in a g-dimensional space cannot exceed g, we must have $\Sigma_\mu n_\mu^2 \leq g$. We can obtain similar formulas for the characters $\chi^{(\mu)}$ by setting $i = l$ and $j = m$ and summing over i and j:

$$\sum_R \chi^{(\mu)}(R)\chi^{(\nu)}(R^{-1}) = g\delta_{\mu\nu}, \qquad (20)$$

or

$$\sum_R \chi^{(\mu)}(R)\chi^{(\nu)*}(R) = g\delta_{\mu\nu} \qquad (21)$$

if the representation is unitary. We saw earlier that all the group elements in the same class have the same character. We label the classes K_1 to K_r, and denote by g_i the number in the class K_i. Then the last equation becomes

$$\sum_{i=1}^{r} \chi_i^{(\mu)}\chi_i^{(\nu)*}g_i = g\delta_{\mu\nu}, \qquad (22)$$

where $\chi^\mu(R) = \chi_i^{(\mu)}$ for all elements in the class K_i.

2.4 Kronecker Product of Representations

If we have two irreducible representations $D^{(\mu)}$ and $D^{(\nu)}$ of the symmetry group, we can take products of their basis functions and get the Kronecker product representation $D^{(\mu\times\nu)}(R)$ with matrices

$$[D^{(\mu\times\nu)}(R)]_{ik,jl} = D_{ij}^{(\mu)}(R)D_{kl}^{(\nu)}(R). \qquad (23)$$

The character of $D^{(\mu\times\nu)}$ can be found by setting $j = i$, $l = k$, and summing over i and k:

$$\chi^{(\mu\times\nu)}(R) = \chi^{(\mu)}(R)\chi^{(\nu)}(R). \qquad (24)$$

All elements R in the same class K_i will have the same character $\chi_i^{(\mu\times\nu)}$. The scalar product of the basis functions is

$$(\psi_i^{(\mu)}, \varphi_j^{(\nu)}) = \int d\tau\, \psi_i^{(\mu)*}\varphi_j^{(\nu)}. \qquad (25)$$

For unitary representations this expression is identical with $(D^{(\mu)}(R)\psi_i^{(\mu)},\ D^{(\nu)}(R)\varphi_j^{(\nu)})$ for any R. If we use Eq. (13) and the orthogonality relation of Eq. (18) we find

$$(\psi_i^{(\mu)}, \varphi_j^{(\nu)} = \frac{1}{n_\mu}\sum_k (\psi_k^{(\mu)}, \varphi_k^{(\nu)})\delta_{\mu\nu}\delta_{ij}. \qquad (26)$$

The scalar product is zero for $i \neq j$, i.e., basis functions belonging to different rows are orthogonal. Setting $\mu = \nu$ and $i = j$, we find that $(\psi_i^{(\mu)}, \varphi_i^{(\mu)})$ is independent of i. This means that the scalar product of two functions belonging to the same row of an irreducible representation is independent of the row. We shall see that this result is the basis of the use of perturbation theory and selection rules throughout the applications of group theory.

2.5 Analysis of Representations

If we know the characters of the irreducible representations of the group G, we can use the above theorems to find how a given representation decomposes into irreducible constituents. We found (see p. 376) the equation

$$D(R) = \sum_\mu a_\mu D^{(\mu)}(R). \qquad (27)$$

Taking the trace for an element R in the class K_i, we get

$$\chi_i = \sum_\mu a_\mu \chi_i^{(\mu)}. \qquad (28)$$

Next we multiply by $\chi_i^{(\nu)*} g_i$ and sum over i. Using the orthogonality relations found earlier, this gives

$$a_\nu = \frac{1}{g} \sum_i g_i \chi_i^{(\nu)*} \chi_i. \tag{29}$$

Thus the number of times a given irreducible representation is contained in D is given by this formula. In particular this shows that if two representations have the same set of characters, they are equivalent. Again, if we multiply Eq. (29) by g times its complex conjugate equation and sum over i, we find

$$\sum_i g_i |\chi_i|^2 = g \sum_\mu a_\mu^2. \tag{30}$$

If the representation is irreducible, all the a_μ must be zero, except for one which is equal to 1. So if the representation is irreducible, its characters must satisfy the equation.

$$\sum_i g_i |\chi_i|^2 = g, \tag{31}$$

which gives a simple test for irreducibility. Finally one can show that the number of inequivalent irreducible representations is precisely equal to the number of conjugate classes in the group, and that

$$g = \sum_{\mu=1}^{r} n_\mu^2, \tag{32}$$

i.e., the sum of the squares of the dimensions of all the nonequivalent irreducible representations is equal to the order of the group.

We give some examples of finding characters for some groups.

1. Cyclic groups: $a, \ldots, a^n = e$. These groups are Abelian, so all irreducible representations have dimension 1. The matrices are 1×1 and the representative is just the character, a complex number. Since $a^n = e$, the nth power of the character $D(a)$ must equal 1, so $D(a) = \exp(2\pi i m/n)$, $m = 1, \ldots, n$, and $D(a^r) = \exp(2\pi i m r/n)$.
2. General Abelian group. Again all irreducible representations are one-dimensional. We choose any element and take its powers. This gives a cyclic subgroup of G. We repeat this process with some other element. We see that the group is a direct product of cyclic subgroups. For example, if $G = C_2 \otimes C_3 \otimes C_5$, with $g = 30$, we have generators a, b, c, with $a^2 = b^3 = c^5 = e$, and the character of any element $a^m b^n c^p$ is $\exp[2\pi i(mr/2 + ns/3 + pt/5)]$.
3. Point groups. These are the groups of rotations about a fixed point (the origin) and reflections in planes through the origin. For crystals, only the rotations through multiples of $2\pi(\frac{1}{2}; \frac{1}{3}; \frac{1}{4}; \frac{1}{6})$ are permitted, i.e., only cyclic subgroups C_2, C_3, C_4, C_6. As a typical example, we treat the octahedral (cubic) group O. This is the group of rotations about axes through the center of the cube that take the cube into itself. It consists of 24 rotations in 5 classes: 6 rotations C_4, C_4^3 around lines joining the midpoints of opposite faces; 3 rotations C_4^2 around these same axes; 6 rotations C_2 around lines joining the midpoints of opposite edges; 8 rotations C_3, C_3^2 around lines joining opposite vertices of the cube; and the identity E. There are five nonequivalent irreducible representations. Using Eq. (32) we find

$$24 = \sum_{\mu=1}^{5} n_\mu^2$$

which has the unique solution $n_\mu = 3, 3, 2, 1, 1$. The character table for this

group is

	E	$C_3, C_3^2(8)$	$C_4^2(3)$	$C_2(6)$	$C_4, C_4^3(6)$
Γ_1	1	1	1	1	1
Γ_2	1	1	1	-1	-1
Γ_3	2	-1	2	0	0
Γ_4	3	0	-1	1	-1
Γ_5	3	0	-1	-1	1

where $\Gamma_1, \ldots, \Gamma_5$ label the irreducible representations. The column vectors are orthogonal:

$$1(1) + 1(1) + 2(-1) + 3(0) + 3(0) = 0, \text{ etc.},$$

and normalized to $g = 24$:

$$1^2 + 1^2 + 2^2 + 3^2 + 3^2 = 24,$$
$$8[1^2 + 1^2 + (-1)^2] = 24, \text{ etc.}$$

The row vectors are also orthonormal when we include the factors g_i:

$$1(2) + 8(1)(-1) + 3(1)(2) + 6(-1)(0) + 6(-1)(0) = 0,$$
$$3^2 + 8(0)(0) + 3(-1)^2 + 6(1)^2 + 6(-1)^2 = 24, \text{ etc.}$$

4. The permutation groups S_n. These finite groups are important for dealing with identical particles, construction of irreducible tensors, and finding states in atomic or nuclear shell models. Earlier we described the conjugate classes in S_n. We found that there is a class corresponding to each partition of n. For example, for $n = 3$, we have $r = 3$ and $g = 3!$, so $6 = \Sigma_i n_i^2$, so $n_i = 2, 1, 1$. There are two one-dimensional and one two-dimensional irreducible representations. The character table is

	E	$(12)_3$	$(123)_2$
Γ_1	1	1	1
Γ_2	1	-1	1
Γ_3	2	0	-1

For large n, the simple procedure becomes impractical. Instead we use Young diagrams. For each partition of $n = n_1 + n_2 + \cdots + n_r$, with $n_1 \geq n_2 \geq \ldots \geq n_r$, we draw a diagram with n_1 dots in the top row, n_2 in the second, etc. For example, for $n = 5 = 3 + 1 + 1$ we get the Young diagram

$$\begin{array}{lll} \bullet & \bullet & \bullet \\ \bullet & & \\ \bullet & & \end{array}$$

Each such partition gives an irreducible representation of the group S_n. Next we enter the digits 1–5 in the boxes in all possible ways that keep the entries increasing to the right in rows and down in columns. These arrangements are the standard Young tableaux:

123	124	125	134	135	145
4	3	3	2	2	2
5	5	4	5	4	3.

There are six standard tableaux so the dimension of this irreducible representation is 6. The use of the symmetric group S_n and the construction of irreducible representations is discussed in Hamermesh (1989), Chaps. 7 and 10.

5. SO(3) is the group of pure rotations in three dimensions. All rotations through a given angle θ about any axis $\mathbf{n}$ are in the same class (see p. 368), so if we choose the z axis for $\mathbf{n}$ the rotation matrix is

$$\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and the character for the vector representation is $\chi^{(1)}(\theta) = 1 + 2\cos\theta = e^{i\theta} + e^0 + e^{-i\theta}$. One can show that there is a single irreducible representation for

each integer $l = 0, 1, \ldots$ with

$$\chi^{(l)}(\theta) = \sum_{m=-l}^{+l} e^{im\theta}, \qquad (33)$$

and the dimension of the representation is $\chi^{(l)}(0) = 2l + 1$. There are also irreducible representations of SO(3) for $l = \frac{1}{2}, \frac{3}{2}$, etc. These are double-valued representations that can be derived by using the group SU(2), which is homomorphic to SO(3) with two elements corresponding to each element of SO(3). These irreducible representations give all information needed about states in a central field. For a detailed treatment of group representations, see Elliott and Dauber (1979), Chaps. 4 and 7; Hamermesh (1989), Chaps. 3, 4, and 5.

3 Applications

3.1 Atomic Structure

The application of group theory to most physical problems requires the use of some model. It should be evident that we cannot consider the quantum problem exactly if we have more than two entities interacting with one another. We must find some method of successive approximations that enables us finally to get satisfactory agreement with experimental results. In atomic physics our procedure is first to consider the individual electrons to be moving in the field of the nucleus and some spherically symmetric averaged field of the other electrons in the atom. In this central field the Hamiltonian for the individual electron has the symmetry group SO(3), so the states of a single electron have quantum numbers l and $m = -l, \ldots, +l$. A single-particle level with quantum number l has degeneracy $2l + 1$. There will be many states with a given lm, with energies $E = b/n^2$ approximately, where $n = 1, 2, \ldots$ giving us a triple of labels n, l, m, with $n \geq l + 1$, and $l = 0, 1, \ldots$, where l is the orbital angular momentum. Often we use letters in place of l:

$l =$	0	1	2	3	4
state label	s	p	d	f	g.

For $l = 0$, we have $1s, 2s, 3s, \ldots$; For $l = 1$, we have $2p, 3p, \ldots$; For $l = 2$, we have $3d, 4d, \ldots$, etc. The energy levels are given approximately by $E = b/n^2$, so the level ns is nondegenerate ($m = 0$), the level np has $m = \pm 1, 0$, etc. In addition we know that each electron carries an internal angular momentum (spin) with $s = \frac{1}{2}$. In light atoms the spin results in there being two states (with $m = \pm\frac{1}{2}$). This results in doubling the occupancy of each level. Thus we have a level sequence

Level	$1s$	$2s$	$2p$	$3s$	$3p$	$3d$	$4s$	$4p$	$4d$	$4f$
Degeneracy	2	2	6	2	6	10	2	6	10	14,

etc., where the atomic shells are labeled by n, and have a total number of electrons $= 2, 8, 18, \ldots$. The order of the levels goes with n, but some changes occur when the $3d$ and $4f$ levels become partly filled.

To study the spectrum of levels for an atom with atomic number Z, we fill the single-particle levels successively. For example, for $Z = 16$, the only unfilled shell would be $3p$, with two electrons. This state would be described by $(1s)^2(2s)^2(2p)^6(3s)^2(3p)^2$. The closed inner shells have spherical symmetry, so we consider only the two electrons in the

$3p$ shell. The orbital states of the two electrons have $l_1 = l_2 = 1$, while the spin states are $s_1 = s_2 = \frac{1}{2}$. For light atoms we use Russell-Saunders coupling, in which we first find the Kronecker product of all the space functions, and separately of all the spin functions, and then couple the two resultants. To keep the Coulomb force between the electrons unchanged requires that the same rotation R be applied to both electrons, so that we are taking the product representation $D^{(l_1)}(R)D^{(l_2)}(R)$. The basis functions are $\psi^{(1)}_{m_1}(1)\psi^{(1)}_{m_2}(2)$, giving nine coupled wave functions. This reducible representation can be decomposed into $L = 2, 1, 0$. The two spins will couple to $S = 1, 0$. Finally we include the weaker coupling of orbit and spin to give the states of total angular momentum $J = L + S$, and find states ${}^3D, {}^3P, {}^3S$ and ${}^1D, {}^1P, {}^1S$. But we must also consider that the electrons are identical particles (fermions), and so the Pauli principle requires that the total wave function must be antisymmetric under the interchange of all coordinates of the two electrons. For this simple case of two electrons the nine functions $\psi^{(1)}_{m_1}(1)\psi^{(1)}_{m_2}(2)$ can be split into a symmetric second-rank tensor (with trace zero), an antisymmetric second-rank tensor, and a scalar. The orbital states with $L = 2, 0$ are symmetric while the state with $L = 1$ is antisymmetric. The spin states with $S = 1$ are symmetric (triplet states) while the $S = 0$ singlet states are antisymmetric. Thus the Pauli principle allows only the states ${}^3P, {}^1D, {}^1S$. Note that this argument would not apply to the case of $(3p)(4p)$ because then the states of the two electrons are not the same, so the Pauli principle has no effect.

This simple procedure fails if there are more than two electrons in the same subshell. For example, for the case of $(3p)^3$ we would have 27 product functions $\psi^{(1)}_{m_1}(1)\psi^{(1)}_{m_2}(2)\psi^{(1)}_{m_3}(3)$ while the spin part would be the product of three spin-$\frac{1}{2}$ functions. The general procedure requires the use of irreducible tensors. If we have r electrons in a subshell with angular momentum l our spatial wave function for one electron is $\psi^{(l)}_m$, and for the r electrons the product is $\Psi = \psi^{(l)}_{m_1}(1)\ldots\psi^{(l)}_{m_r}(r)$ with $m_i = -l, \ldots, +l$. Thus Ψ is an rth-rank tensor in a space of dimension $2l + 1$. Since all these functions are degenerate, the transformations of the group $\mathrm{SU}(2l + 1)$ act on these ψ's. We then must classify them according to the irreducible representations of $\mathrm{SU}(2l + 1)$ with their particular symmetries. Similarly, for the spins we have tensors $\varphi^{(1/2)}_{\mu_1}(1), \ldots, \varphi^{(1/2)}_{\mu_r}(r)$ of rank r in the space of dimension 2. Again we must find the irreducible pieces. Finally, to satisfy the Pauli principle we must assure that the total wave function is antisymmetric. For the detailed treatment, see Hamermesh (1989), Chap. 11, or Elliott and Dauber (1979), Chap. 8.

3.2 Nuclear Structure

In the case of atoms, we know that the interactions are electromagnetic, and we have just one type of identical particle, electrons. In the case of the nucleus we have two constituents, neutrons and protons, which have approximately equal masses and can transform into one another in β-ray processes. The Coulomb force between protons is small compared with the specific nuclear force, so we use a model in which we disregard the differences between n and p, and deal with a single type of identical particles, nucleons, in two charge states. This looks like the case of two possible spin states for

electrons. Both n and p have an intrinsic angular momentum $s = \frac{1}{2}$, but now there is another intrinsic feature, the isospin $t = \frac{1}{2}$. The other important difference is that we have little basis for choice of a model. We start with no knowledge of the interaction and must try various shapes for its radial dependence and its dependence on spin and isospin.

The most successful model requires the use of j–j coupling. In such models we first couple (i.e., take the product of) the $\psi^{(l)}$ and $\psi^{(s)}$ for an individual nucleon to give a resultant $\psi^{(j)}$ and then couple the $\psi^{(j)}$s (i.e., take products of $\psi_m^{(j)}$ for the nucleons). Since the nuclear force is attractive, the state with lowest j for a given l,s will have the highest symmetry (since the particles can get closest to one another). The order of the levels will also depend on the choice of the radial dependence of the potential. One model is shown in the following diagram of single-particle levels and their degeneracies:

Nucleon states		Occupation	Total
$1g_{9/2}$		10	50
	$2p_{1/2}$	2	
$1f_{5/2}$		6	
	$2p_{3/2}$	4	
$1g_{7/2}$		8	28
$1d_{3/2}$		4	20
	$2s_{1/2}$	2	
$1d_{5/2}$		6	
$1p_{1/2}$		2	8
$1p_{3/2}$		4	
$1s_{1/2}$		2	2

The numbers in the right-hand column are the magic numbers corresponding to the closed shells.

As in the atomic case, we look only at the partially filled shells. If we consider a level $(j)^3$, the single-particle state has a wave function $\psi^{(j)}$ that is a vector in a space of dimension $2j+1$. Since the js are halves of odd integers, $2j+1$ is even. The wave function for $(j)^3$ is $\psi_{m_i}^{(j)}(1)\psi_{m_2}^{(j)}(2)\psi_{m_3}^{(j)}(3)$, a third-rank tensor in a space of dimension $2j+1$. Next we take the product of the three isospin functions which are vectors in a space of dimension 2. Just as in the atomic case, we must find the irreducible parts of each tensor. For the isospin tensor we get $T = \frac{3}{2}, \frac{1}{2}, \frac{1}{2}$. The completion of the problem would be the same as in the atomic case [see Hamermesh (1989), Chap. 11–9].

3.3 Solid State and Crystal-Field Theory

In discussing atomic structure we used a central-field model in which the Hamiltonian had the symmetry group SO(3). If we study the energy levels of electrons in a crystal, this is no longer the correct symmetry group. Now the electron moves in the periodic field of the crystal lattice. In the neighborhood of a lattice point the field is no longer spherically symmetric. It now has the symmetry of one of the crystal point groups. For example, the wave function of an electron in a cubic crystal will belong to an irreducible representation of the cubic point group O. Thus the possible degeneracy of a level will be 1, 2, or 3. (See the character table on p. 378.) If the crystal is distorted, the point symmetry group will be reduced from O to some lower symmetry and the levels may split. Or we may consider what happens to a level in a free atom when the atom is put at a crystal site. In the free atom the levels correspond to irreducible representations of the symmetry group SO(3). When the

atom is placed in the crystal, we must use the crystal-field symmetry. The level with a given l may split into levels belonging to the symmetry group of the crystal ("crystal-field theory"). We illustrate this for the octahedral group. A level belonging to the irreducible representation $D^{(L)}$ of SO(3) has the character

$$\chi^{(L)}_{(\theta)} = \sum_{m=-L}^{+L} e^{im\theta}.$$

In the crystal this representation will be reducible. The cubic field has only rotations through angles $\theta = 0, \pi/2, \pi$, and $2\pi/3$. In the crystal we need to record only the $\chi^{(L)}(\theta)$ for these values of θ. We repeat the character table of O from p. 376 and enter below the characters for $L = 0$, 1, etc. Then we decompose using Eq. (29):

	E	C_3 $C_3^2(8)$	$C_4^2(3)$	$C_2(6)$	C_4 $C_4^3(6)$
Γ_1	1	1	1	1	1
Γ_2	1	1	1	−1	−1
Γ_3	2	−1	2	0	0
Γ_4	3	0	−1	1	−1
Γ_5	3	0	−1	−1	1
$L = 0$	1	1	1	1	1
$L = 1$	3	0	−1	−1	1
$L = 2$	5	−1	1	1	−1
$L = 3$	7	1	−1	−1	−1
$L = 4$	9	0	1	1	1

For example, we see that $L = 0$ gives Γ_1. $L = 1$ is just Γ_5 so the vector does not split in a cubic field. For $L = 2$ we find $\Gamma_3 + \Gamma_4$, so the level splits into a doublet and a triplet. $L = 3$ splits into $\Gamma_2 + \Gamma_4 + \Gamma_5$, while $L = 4$ gives $\Gamma_1 + \Gamma_3 + \Gamma_4 + \Gamma_5$.

For details and applications to various point groups and the construction of wave-functions, see Hamermesh (1989), Chap. 9; Elliott and Dauber (1979), Chap. 14; or Burns (1977), Chaps. 8, 11, and 12.

4 Summary

Group theory has developed in the last 60 years to become an essential tool in many branches of physics and chemistry. After a mathematical introduction to the subject, we have given examples of its application to atomic and nuclear structure, and to crystal-field theory.

List of Works Cited

Burns, G. (1977), *Introduction to Group Theory with Applications*, New York: Academic.
Elliott, J. P., Dauber, P. G. (1979), Symmetry in Physics, Vols. 1 and 2, New York: Oxford.
Hamermesh, M. (1989), *Group Theory and Its Application to Physical Problems*, New York: Dover. [Original publication (1962), Reading, MA: Addison-Wesley.]

Further Reading

Burns, G. (1977), *Introduction to Group Theory with Applications*, New York: Academic.
Elliott, J. P., Dauber, P. G. (1979), Symmetry in Physics, Vols. 1 and 2, New York: Oxford.
Hamermesh, M. (1962), *Group Theory and Its Application to Physical Problems*, Reading, MA: Addison-Wesley; reprint (1989), New York: Dover.
Perkins, D. H. (1987), *Introduction to High Energy Physics*, Reading, MA: Addison-Wesley.
Tung, W.-K. (1985), *Group Theory in Physics*, Philadelphia: World Scientific.

Mathematical Modeling

Kenneth Hartt
Physics Department, University of Rhode Island, Kingston, Rhode Island, USA

Introduction ***214***
1 About Models: Types and Trends ***215***
1.1 Chaos as a New Paradigm in Science *216*
1.2 Complex Systems as a Focus of Mathematical Modeling *217*
1.3 Computer Simulation Experiments as a Third Branch of Science *217*
2 Digital Computing Developments ***217***
2.1 Hardware *217*
2.2 Software *218*
3 Selected Models and Modeling Tools ***219***
3.1 Fractals *219*
3.2 Nonlinear Dynamics *221*
3.2.1 Deterministic Models *221*
3.2.2 Stochastic Models *221*
3.2.3 Formal and Experimental Dynamics *222*
3.2.4 Time-Delay Embedding *222*
3.2.5 Cellular Automata *223*
3.3 Rational Functions *223*
3.4 Monte Carlo Methods *224*
3.5 Numerical Modeling *226*
3.5.1 Chaotic Systems *227*
3.5.2 Finite Elements *228*
3.5.3 General Circulation Models *229*
3.5.4 Data-Assimilation Models *229*
3.6 Wavelets *230*
3.7 Systems Analysis *231*
3.7.1 Control Theory *231*
3.7.2 Self-Organization *233*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

4 Information Processing Models *233*
4.1 Principal-Component Analysis *233*
4.2 Neural Networks *235*
4.3 Time-Series Analysis *235*
4.3.1 Signal Processing *237*
4.4 Statistical Inference *238*
4.5 Information Theory *239*
5 Applications *240*
5.1 El Niño *240*
5.2 Chaotic Ocean Heat Transport *242*
5.3 Controlling Chaos *243*
5.4 Solitons *243*
Glossary *244*
List of Works Cited *246*
Further Reading *248*

Introduction

A mathematical model is a set of equations and algorithms, together with a collection of interpretive rules, designed to represent something. Mathematical models are invariably approximations. They include physical laws that, like Newtonian mechanics, the special and general theories of relativity, or quantum mechanics, are guiding precepts at the moment, within their specified limits. They also include exponential growth models that are Band-Aids, meant to be replaced. Mathematical models selectively magnify some parts of a system while ignoring or severely approximating other parts. The art of modeling is to distinguish essential from superfluous detail and to be able to "consider a spherical cow" (Harte, 1988). To learn of the goals, philosophy, and methods of mathematical modeling within any field of physics, one can do nothing better than read the article in this Encyclopedia pertaining to the field. The purpose here is to sample the breadth of emerging tools and techniques.

Change through computer power is the dominant feature in the world of mathematical modeling. During the brief history of electronic computation, barriers to formal mathematics eroded slowly. In 1976, the four-color problem was solved: It was proved that only four colors are needed for maps to assure that no two countries sharing a common border have the same color (Appel and Haken, 1978). This is a combinatoric problem, solved by making use of a computer in essential ways. Since then, computer-assisted researchers have won acceptance as not being children of a lesser God.

Although analytic methods and rigorous mathematics theorems continue to be pillars, they are enhanced by modern-day computers, sometimes through the use of symbolic computer software. Computer technology at present can give us commercially available teraflop speeds (10^{12} floating point operations per second), and will perhaps give 10^{15} flops in ten years (Pool, 1992). Massively parallel processing already exists. Every two years produces

workstations and personal computers using more sophisticated graphics and software, making the previous generation obsolete. A researcher must carefully choose the appropriate hardware and software system, because of overhead costs in financial resources (measurable in units of $10 000) and learning time (units of 6 months). The reward is the power to explore new models of greater complexity and old models to greater depth and to share rapidly advancing computer visualization.

Just as a treatise on mathematical modeling 180 years ago would have been remiss to omit mention of Fourier analysis, so would it be remiss today to omit mention of wavelets, chaos, fractals, nonlinear dynamics, numerical modeling, rational functions, symbolic computation, time series, Monte Carlo and inverse methods, systems analysis, and principal-component analysis. In their present form, they all are children of the computer age, with a scope extending beyond the context of their inception. We concentrate on these and add statistical inference, information theory, cellular automata, self-organization, and data assimilation, which are comparable in breadth of interest. We also mention software development.

Our discussion is illustrative rather than comprehensive and is not rigorously parallel or hierarchical. The recent treatise on mathematical modeling with which this article most overlaps is the two-volume work of Casti (1992). Each work contains things the other does not. Whereas Casti tends to emphasize techniques more than the models, and the journal *Applied Mathematical Modelling* emphasizes models more than techniques, we attempt to strike a median between the two. Our physicist's viewpoint relegates – perhaps unfairly – numerical modeling to less than 10% of the whole. Consequently, some important numerical analysis developments such as splines, singular integrals, and large matrix inversion are given little or no space. Whether one uses commercial numerical software systems or does one's own numerical programming, we advise caution and careful testing. We refer the reader to the article NUMERICAL METHODS for more detailed advice. By mathematicians' standards, this work is decidedly qualitative. We omit or relegate to the incidental almost all the standard fare of a first course on mathematical methods of physics, such as tensor analysis, finite- and infinite-dimensional linear spaces, distribution theory, Fourier and Hilbert transforms, special functions, methods of solving integral and differential equations, and group theory, for which excellent textbooks exist (Hassani, 1991; Arfken, 1985). What remains are topics we believe practicing scientific modelers can ignore only at their own peril!

1
About Models: Types and Trends

We follow the classification of climate modelers (Schneider and Dickinson, 1974) in identifying two types of models: *mechanistic models*, which emphasize mechanisms; and *simulation models*, which introduce explicitly as many relevant degrees of freedom as possible. Most models are a mixture of the two. Consider, for example, models designed to simulate circulations in the atmosphere and the oceans. They are called general circulation models (GCM's). The GCM's do not possess a grid scale sufficiently fine to permit simulation of clouds. However, insight gained from cloud simulation models within which environments of clouds are mechanistically

described has led to a better mechanistic description of cloudiness within GCM's. The interplay between mechanistic and simulation models also sheds light on feedbacks.

A mechanistic description of a subphenomenon within a larger simulational system is called *parametrization*. In GCM's, both cloudiness and precipitation are parametrized. The more mechanistic a model, the more directly observable are its variables. For example, pressure as an operationally defined variable in thermodynamics, a mechanistic model, contrasts with pressure as an end product of a lengthy computation in kinetic theory, a simulation model. The ideal gas law is an often-used parametrization. The validations of parametrizations within GCM's, particularly for climatic regimes other than the present and, hence, only indirectly accessible to observation, have been perhaps the major issue in their development and are crucial in the study of global warming.

Models are further classified according to their relation to experimental observations. *Direct models* make predictions for comparison with experiment, using given parameters. *Inverse models* make predictions of parameters of direct models, using experimental data. *Data-assimilation models* make direct predictions of the near future or more complete descriptions of the present, while using inverse methods for the purpose of updating and improving parameter values. Models that show temporal evolution are called *dynamic models*. Dynamic models are further classified as *deterministic* if the rule by which the next successive state in time computed is unique or *stochastic* if that rule can lead to more than one next state, with some probability.

We comment on two environmental systems: El Niño and the Southern Oscillation, in which various models are discussed; and North Atlantic Ocean circulation, where mechanistic models reveal a possibility of low-dimensional chaos. Also, we touch on two major successes: controlling chaos and soliton propagation in optical fibers. Three dominant motifs recur throughout the present discussion. They are centered on chaos, complex systems, and computer simulation, as we now explain.

1.1 Chaos as a New Paradigm in Science

Chaos is the apparently random behavior seen in some deterministic nonlinear dynamical systems. Noisy processes once thought to require stochastic modeling may be chaotic (Tufillaro et al., 1992). Even in planetary motions, chaos has a long-time destabilizing effect. Chaotic systems, intrinsically classical, have a poorly understood relationship with quantum mechanics. Therefore, chaos is sometimes advanced as a fundamentally new point of view, a new paradigm (Ford and Martica, 1992).

Distinguished from randomness by measures of order, chaotic evolution always has a fractal orbital structure, and there is always sensitivity to initial conditions (SIC). Both conservative and dissipative systems can be chaotic. Dissipative chaotic systems evolve to a low-dimensional asymptotic orbit, called an *attractor*. The attractor is reached from initial conditions in a region of phase space called the *basin of attraction*. Since lower dimensions are easier to work with, the existence of an attractor helps analyze dissipative chaos. There is information loss, measurable in bits per time step. Weather, with its long-range unpredictability, is probably chaotic.

1.2 Complex Systems as a Focus of Mathematical Modeling

As prime examples of complex systems, GCM's are heavily parametrized. Systems analysis, emphasizing the relationship among the components of a system, is relevant here; mechanistic models of complex systems use systems concepts of feedback and control. In other developments, phase-transition models for aggregate matter are used in the theory of strongly interacting elementary particles, self-organization, and chaos. Adaptive learning systems such as neural networks, themselves models of complex systems, are used heuristically for model building (Zurada, 1992).

1.3 Computer Simulation Experiments as a Third Branch of Science

Simulation of complex or chaotic systems generates large data sets. Complex systems may be deterministic, such as with molecular dynamics (Yonezawa, 1993), stochastic, as in Monte Carlo simulations (Creutz, 1992), or a mixture of the two. Interesting, complicated data are also produced from simulation of a chaotic one-dimensional nonlinear driven oscillator. Investigations using such computer-generated data have become a dominant effort in scientific work and are sometimes referred to as *experimental theoretical physics* (ETP). Improved digital computer power will make possible ETP simulations of aerodynamic flows and many important systems, which can also include hypothetical structures such as exotic chemical compounds not yet seen in the laboratory.

2 Digital Computing Developments

Trends in digital computing today are

1. the explosion of computer power, both serial and parallel;
2. a rapid increase in connectivity ("networking" and "clustering") making stand-alone computing obsolete;
3. a rapid increase in specialized software; and
4. the broadening scope of software systems.

2.1 Hardware

The order of magnitude of current workstation speed is 100 megaflops. This resembles supercomputer speed of just ten years ago. It is slower by many orders of magnitude than current linear supercomputers, now in the gigaflop range; however, through parallel links of the slower workstation processors, competitive parallel computers are obtained. Massively parallel computer designs are now at a teraflop level. Pressure to increase computer speeds continues from industrial, defense, environmental, medical, and academic demands.

Massively parallel computer designs currently couple 10^3 to 10^4 individual processors with their own individual memories (Pool, 1992). In a large class of designs, each processor is only interconnected with its nearest neighbors. Current major efforts are devoted to solving the problem of communications among processors and the problem of programming for their efficient use. Modeling of some, but not all, complex systems lends itself to massively parallel computing. Modern supercomputers, such as those produced by Cray, excel

because of their sequential power, even though some can possess up to 16 parallel processors. A sequential problem run on the current generation of massively parallel computers will go slowly, because their individual processors are slow by supercomputer standards. There are diverse concepts of parallel processors. They in turn can be modeled in broad categories such as non-shared memory, shared memory, and Boolean circuits (Lakshmivarahan and Dhall, 1990). It is useful for mathematical modelers whose work challenges modern computer capabilities to become familiar with computer architecture at this level. Consequently, modelers collaborate with computer manufacturer in-house scientists (Butler et al., 1993). Highly connected parallel computers being built show promise for problems that are not essentially sequential and that do contain synchronous parts.

2.2 Software

The past 30 years have seen the development of numerous higher-level programming languages and the increased use of software systems. FORTRAN, long a standard bearer for numerical programming, has progressed to FORTRAN 90 with parallel processing capabilities. Many other supported programming languages are available. Programming languages are accompanied by increasingly sophisticated debugging and structuring software.

Commercial software systems exist for practically every purpose and are heavily advertised in computer journals. The journal *Computers in Physics* also has sections on research and education, visualization, techniques and technology, and refereed articles. Technical emphasis in universities is on learning to use a progression of specialized software systems, often starting with spread-sheets and including computer algebra. MATLAB is a commercial numerical software system for interactive use. Accessible graphics, facile handling of matrix problems, some limited symbol manipulation, and continual broadening of scope are features of MATLAB. For numerical work, Numerical Recipes and NAG (Numerical Algorithms Group) are well-known commercial groups that also continue to diversify. Several commercial software systems exist for neural networks. In the public domain are many specialized software packages such as FERRET, used by geophysicists to do interactive processing of large data sets.

Historically, computer algebra has been developed for analytic problems of celestial mechanics, general relativity, and quantum mechanics, saving months to years over pencil and paper. The oldest well-known symbolic software system, REDUCE, is open and accessible to independent developers. MATHEMATICA has strong advocates, and its graphics have been extolled in recent reviews. MAPLE continues its development through an academic consortium, and like MATHEMATICA has strong graphics, a users' group, and a newsletter. Many find the academic commitment to MAPLE, through Ph.D. dissertations and faculty interest at the University of Waterloo and Eidgenössische Technische Hochschule, Zurich, to be appealing. Users can affect work on MAPLE through direct electronic mail correspondence with faculty. These symbolic systems and others, including AXIOM, DERIVE, and MACSYMA, have recently been discussed (Fitch, 1993).

Computer algebra is playing an increasing role in the study of ordinary and partial differential equations (Tournier, 1991). This is because symmetries (characterized

through Lie algebras and Lie groups) and analytic transformations among classes of differential equations (such as Bäcklund and auto-Bäcklund transforms) are sometimes far more easily explored with computer algebra than by hand.

Another modeling enhancement is telecommunications. Electronic mail and networking, remote and local, are currently estimated to serve more than 10^7 individual workers. For many researchers, telecommunications provides ready access to remote installations offering vastly more powerful machines when needed, with more versatile and more appropriate software. Collaborations of groups spread across the globe can be carried out on just one computer, in some cases with almost instant response to comments and queries.

3 Selected Models and Modeling Tools

3.1 Fractals

A "fractal" is an infinite set of points described in terms of a noninteger scaling parameter. This common thread encompasses an increasing variety. One broad class, *finite-size fractals*, was first analyzed by Mandelbrot, who originated the modern study of fractals, coining the word fractal (Mandelbrot, 1983). Some finite-size fractals are described in terms of a self-affine scaling behavior: if one looks at them through a magnifier, the images will have clear similarities to the originals, although there may be distortions in different directions. If there are no distortions at all, these fractals are called self-similar. The Hausdorff dimension is the nonintegral scaling parameter for many finite-size fractals. *Fat fractals*, an exception, have integral Hausdorff dimension, though they still possess a nonintegral scaling parameter (Vicsek, 1992). *Growth fractals* form another major class of fractals. They are not self-affine except in some statistical sense. Prototypes are plane dendritic growths (Vicsek, 1992). A union of fractals with differing scaling behaviors is called a *multifractal*. It is easy to construct fractals (including multifractals), using either deterministic or stochastic rules. It is harder to identify naturally occurring fractal structures.

Fractals are ubiquitous in spatial structures and are even associated with modern art. A chaotic system always possesses a fractal orbit. For example, fractal structures have been observed in preturbulent flow (Brandstater and Swinney, 1987). Fully developed turbulence has been shown to have a well-defined multifractal spatial structure (Sreenivasan, 1991), which is not yet fully understood. The detection and characterization of fractals anywhere in nature is an important step in mathematical modeling. Models of fractal growth are being studied intensively (Vicsek, 1992). In Secs. 3.2.1, 3.5.1, 4.4, and 5.3, fractal structure is intrinsic to the discussions. It is therefore relevant to understand how, in a simpler context, noninteger scaling is inferred.

Here we introduce a fractal dimension D which is a standard approximation to the Hausdorff dimension. Dimension D is defined so that, intuitively, D is an integer for nonfractal sets. Consider first a line segment of length L. The number $N(l)$ of small segments of length l needed to cover the length L is Ll^{-1}. This is true for progressively smaller l. The negative of the exponent of l is the scaling parameter, D; the fractal dimension D is unity for this nonfractal set. Now consider a square of side L. The number $N(l)$ of small squares

of side l needed to cover the square of size L^2 is $L^2 l^{-2}$. Then again, $N(l) \sim l^{-D}$, and the scaling parameter D is the integer 2 for this nonfractal set. This relation can also be written

$$D = -\lim_{l \to 0} \ln \frac{N(l)}{\ln(l)}. \qquad (1)$$

We now construct a fractal, Smale's horseshoe map, perhaps the simplest exemplar of a chaotic dynamical system (Smale, 1967). We define an infinite sequence of iterations by first drawing a rectangle of width W and height H. Each iteration occurs in two steps. The first step, stretching, is to draw another rectangle of width $2W$ and height reduced by a factor $1/2\delta$ where δ is some fixed number greater than 1 – say, 2. The last step of the iteration, folding, is to bend the rectangle over into the shape of a horseshoe of width W again, whose sides do not touch. The new area inside the boundary of the stretched and folded rectangle is, of course, less than the old. If we draw a vertical line through the new shape, its intersection with the inner area is in two pieces, each of height $H(1/2\delta) = H/4$. When the stretching and folding process is iterated exactly the same way k times, a layered structure develops whose inner area makes $2k$ intersections with a vertical line, each of height $H(1/2\delta)^k$. We now use Eq. (1) to obtain $D = (\ln 2)/(\ln 2\delta) = 0.5$. The limiting internal area is zero. The fractal dimension of the whole layered figure is the sum of the vertical and horizontal fractal dimensions, or $1 + (\ln 2)/(\ln 2\delta) = 1.5$.

Shown in Fig. 1 are the first few iterations that in the limit $k \to \infty$ define two other fractals. They are the Cantor middle-thirds set in (a) finite-size and (b) growth versions. We compute the fractal dimension in (a) as before. Case (a) corresponds to a chaotic orbit of the logistic map (see Sec. 3.5.1). After k iterations, there are 2^k pieces, each of length 3^{-k}, leading to a value $D = (\ln 2/\ln 3) = 0.6309\ldots$. Figure 1(a) is converted to a fat fractal by removing the middle $(1/3^k)$th piece rather than the middle third. The fractal dimension D for this fat fractal equals unity.

For growth fractals [case (b)], D is calculated differently. Consider a growth fractal of maximum linear extent L. Here, there is some minimum number $N(L)$ of covering segments of *fixed* edge length l. As $k \to \infty$, we have $L \to \infty$, and D is defined through the relation $N(L) \sim L^D$. The growth rule is iterated three times in Fig. 1(b). The result for $k \to \infty$ is the same as before, $D = (\ln 2)/(\ln 3)$. The union of all iterations in Fig. 1(b) also has the same fractal dimension. In nature, the infinite limit never occurs, but scale invariance

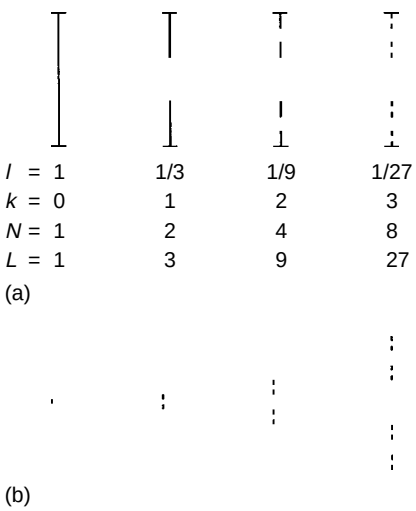


Fig. 1 Construction of a Cantor middle-thirds fractal, for two cases: (a) finite-size and (b) growth fractals. Fractals are defined iteratively in the limit where $k \to \infty$. In both cases, the fractal dimension D equals $\ln(2)/\ln(3)$ (see text for details). Figs. 1 and 3 were quickly drawn using "paint" facility on a DEC-5000 workstation

can be found where L or l varies over many powers of ten (Mandelbrot, 1983).

It is simple to generalize the conceptual process for finding D to a higher-dimensional space, where covering balls (or hypercubes) of volume $V \propto l^d$ are used, with d representing the usual dimension of the underlying space, frequently as large as 10. However, when d exceeds 2, a method of computing D based upon its definition, known as a box-counting method, must be replaced by a more computation-efficient method (Schuster, 1988). Finally, multifractals are observed by using scaling parameters sensitive to the density distribution within a fractal. The examples given here are not multifractal because the distribution of points within the fractals is uniform. Detecting nonuniformity can be accomplished through introduction of generalized dimensions, D_q, defined in terms of the qth moments of this density distribution. The values of q for which D_q is measured can be $q = 0, \pm 1, \pm 2, \ldots, \pm\infty$, and D_0 is identified with the previously defined D. For multifractals, D_q decreases rather than remaining constant as q increases. The art of measuring multifractals is discussed by Sreenivasan (1991). Major efforts to refine the experimental determination of all these non-integral scaling parameters continue (Abraham et al., 1990).

3.2 Nonlinear Dynamics

3.2.1 Deterministic Models

Dynamics is the study of how systems change in time. If time changes in discrete intervals, the evolution is described as an iterative map, or simply, a map. A map can be written

$$\mathbf{X}_{M+1} = \mathbf{F}(\mathbf{X}_M, \mu). \tag{2}$$

Here, $\mathbf{X}$ can lie in an N-dimensional vector space, $\mathbf{F}(\mathbf{X})$ is a vector-valued function in that space, and μ is a set of parameters. The space is called the *phase space*. If the time changes continuously, the evolution is referred to as a flow. The dynamical equations for a flow can be written

$$\frac{d}{dt}\mathbf{X} = \mathbf{F}(\mathbf{X}, \mu). \tag{3}$$

Here, $\mathbf{X}$, $\mathbf{F}(\mathbf{X}, \mu)$, and μ are defined as before, except that now $\mathbf{X}$ depends upon continuous time. Solutions of both kinds of systems are referred to as orbits. Flows are numerically approximated as maps. Hence, the connection between these two versions of dynamics is important.

Nonlinear dynamics is a rich field for modelers. In controlled experiments, a change of control parameters sometimes leads to topologically distinguishable states. The non-linear dynamics theory of such transitions, called bifurcations, is well developed (Schuster, 1988). Bifurcations in lasers can lead to chaotic states, and circuits have been designed that bifurcate into chaos.

3.2.2 Stochastic Models

What have just been defined are deterministic dynamical systems. Other systems, referred to as stochastic, have functions describing time evolution, such as $\mathbf{F}$, that are random variables, wholly or partially. A class of stochastic dynamical systems of considerable interest (Abraham et al., 1990) are the kth order Markov processes, defined in the discrete case by

$$\begin{aligned} &P(\mathbf{X}_L|\mathbf{X}_1, \ldots, \mathbf{X}_{L-1}) \\ &\quad = P(\mathbf{X}_L|\mathbf{X}_{L-k}, \ldots, \mathbf{X}_{L-1}), \end{aligned} \tag{4}$$

representing the conditional probability of observing $\mathbf{X}_L$ as the Lth state in the sequence $\mathbf{X}_1, \ldots, \mathbf{X}_L$. When $k = 1$, this is

simply called a Markov process. Stochastic noise is usually mixed in with chaos.

Analysis of chaotic data attempts to distinguish between stochastic noise and true chaotic variability. Stochastic resonance also can occur, in which noise plays an essential role (Moss, 1991). In stochastic resonance, transitions occur among different stable states of a deterministic nonlinear dynamical system driven by a periodic function plus noise. Stochastic resonance has been observed in lasers, and provides one conceptual model for climatic change. Nonlinear dynamics – deterministic, stochastic, or both together – is the framework for modeling of weather and climate.

3.2.3 Formal and Experimental Dynamics

The literature on nonlinear dynamics is blossoming. In describing this work, we use the word "formal" to denote an analytic mathematical approach. Formal work is sometimes concerned with questions pertaining to effects of finite numerical precision; however, any results reached in a formal approach are characterized by a logical precision that meets the standards of pure mathematics. We employ the broadened meaning of the word "experimental" to include experiments with computer-generated data. Monographs tend to contain a mixture of formal and experimental work (Schuster, 1988). Lorenz's computation of chaotic behavior in 1963, discussed below, is a premier example of experimental dynamical theory. Incisive formal results buttress the field non-linear dynamics. They include those of Sharkovsky in 1964, discussed here in Sec. 3.5.1, and Gardner et al. in 1967, discussed here in Sec. 5.4. Also, in 1967 Smale investigated classes of smooth maps and developed an important exemplar, the horseshoe map (Tufillaro et al., 1992). Formal, too, was the 1971 proposal by Ruelle and Takens of a new route to chaotic turbulence *via* a few bifurcations, seen experimentally in 1975 (Lichtenberg and Lieberman, 1992).

The formal and the experimental are combined in shadowing theorems. A shadowing theorem seeks to answer an important question in chaotic dynamics. All chaotic systems show SIC, seen in divergence of nearby orbits and loss of memory of initial conditions when finite precision calculations are made. Then can any finite precision calculation approximate an asymptotic orbit? Formal computer-assisted answers in the affirmative have been found in special cases – for example, for the Hénon and the Ikeda maps – through the introduction of machine-independent procedures, both in defining truncation and in finding overlapping regions, i.e., a shadow, of a large succession of mapped neighborhoods (Hammel et al., 1988).

3.2.4 Time-Delay Embedding

Time-delay embedding is a technique for reconstructing an orbit in a nonlinear dynamical system from a time series of a single scalar observable $y(t)$. If d is the topological dimension of the manifold in which the orbit lies, there is now considerable evidence that the vector whose components are time-delayed values of $y(t)$, namely,

$$\mathbf{X}(t) = (y(t), y(t+\tau), y(t+2\tau), \ldots, y(t+\{d_E-1\}\tau))^{\mathrm{T}} \qquad (5)$$

(where T is matrix transposition), often adequately mimics the topological and metric behavior of that orbit. If so, the orbit is said to be embedded in the reconstructed finite-dimensional space. This is an indispensable insight: Without

it, one would be required to identify and measure all the observables relevant to that orbit for an adequate description. Time-delay embedding was articulated ca 1980 independently by physicists and a mathematician (Tufillaro et al., 1992). In the absence of an infinite set of noise-free data, it becomes important to optimize the delay parameter τ and to find a practical embedding dimension d_E smaller than the value $2d + 1$, which has been conditionally proved to be sufficient. Information theory has been of practical value in finding τ (Fraser and Swinney, 1986), while numerical experiments show that a good d_E is often smaller than $2d + 1$. Time-delay embedding is a major experimental tool of chaos research.

3.2.5 Cellular Automata

Cellular automaton (CA) models are simplified dynamical systems having discrete spatial and temporal degrees of freedom (Wolfram, 1986). The spatially distinct points, called "cells," can be interpreted to represent arrays filling different dimensions, according to how they are made to interact with their defined "neighborhoods." The CA models have seen use in describing excitable media, both biological and chemical, lattice-gas automata and turbulence, and deterministic evolution of Ising lattices. In three-dimensional excitable media, CA models of propagating scroll-shaped waves are two orders of magnitude faster than partial differential equation (PDE) models (Gerhardt et al., 1991); however, the validation of CA models is relatively more problematic. Although straightforward computationally, allowing rapid checks of various hypotheses, CAs need development of their formal foundations. For example, formal criteria for determining conserved quantities are not generally known (Hattori and Takesue, 1991).

3.3 Rational Functions

A rational function is any ratio of two polynomials, written $P_L(x)/Q_M(x)$, where the degrees of the polynomials are L and M. Taking a function $f(x)$ to be rational is a potent working assumption. Use can often be made of the meromorphic extension of $f(x)$ to the complex plane. Differing from a smoothing technique like a spline, the rational function *Ansatz* is a global statement about the nature of the analyticity as well as the asymptotic nature of $f(x)$. Rational approximants are successfully applied in solving integral and differential equations, but still, much about them remains to be understood (McInnes and Marshall, 1993).

High-degree rational functions sometimes require ultrahigh-precision arithmetic to distinguish between noncanceling poles and zeros. Hence, restriction to a rational function is seminumerical modeling in that analytic structure is maintained, while ultimately numerical evaluation is needed. Sometimes the function to be approximated is entire, having no singularities in the finite complex plane, but, as Padé discovered in his doctoral dissertation in 1892, even the exponential can be represented as a rational function in some domains. Baker rediscovered the power of Padé approximants in 1960 while solving for critical exponents in the theory of phase transitions. Recognition of the power of this tool continues to grow. Even functions with cuts in the complex plane are sometimes representable as rational functions, with arrays of poles and zeros tracing the cuts.

As reintroduced by Baker (Baker and Graves-Morris, 1981), Padé approximants are easily obtained by solving linear equations or linear recursion relations that lead to numerical values of the $\{a_i\}$ and the $\{b_j\}$ in the expression

$$f(z) = \frac{a_0 + a_1 z + \cdots + a_L z^L}{b_0 + b_1 z + \cdots + b_M z^M} + O(z^{L+M+1}) \qquad (6)$$

when the Taylor series expansion

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots \qquad (7)$$

is known. The Baker convention is to make the additional constraint $b_0 = 1$.

One extension of rational-function approximants useful for data modeling and inverse models is the "statistical Padé approximant" (SPA). The SPA has the same form as the PA but is obtained by fitting numerical data by a linear iterative method (Yidana and Hartt, 1988) that frequently converges well even in the presence of significant noise. The SPA has been used for fitting nuclear-scattering data, for inverse-scattering problems, for direct solutions of the Schrödinger equation, and for finding analytic solutions to Riemann boundary-value problems.

Rational functions are fully described by enumeration of their zeros and poles and a leading numerical coefficient and are highly accessible to symbolic computer software and iteration dynamics studies, including chaos. Rational-function iterations form nonlinear dynamical systems through which Fatou, Julia, and Mandelbrot sets are definable (Mandelbrot, 1983) and which occur in fractal image compression methods. Finally, rational functions play a central role in linear circuit theory and in signal processing. Because of their computability and usefulness, they have an interdisciplinary following and are the subject of frequent conferences:

3.4 Monte Carlo Methods

The Monte Carlo (MC) method (*q.v.*) employs algorithms that use random numbers for making convergent statistical estimates. These estimates can be solutions of deterministic problems, evaluations of high-dimensional integrals, or averages over random processes. This method vastly increases the range of problems that can be dealt with theoretically. Although analytic approaches are continually being proposed for complex systems, there are many problems for which no known analytic procedure yet works. Massively parallel computer architecture is useful in MC calculations.

Historically, the MC approach to a modeling problem has had three aspects:

1. generation of random numbers;
2. development of algorithms to use random numbers for solving deterministic problems; and
3. development of models for the use of the other two aspects.

Currently, there is activity in all three. Even the first has its hazards. All methods using computers to generate number sequences involve finite algorithms, which must fail some test for randomness, if for no other reason than that computers are finite-state systems. Such computer-generated sequences are therefore called "pseudorandom numbers." All finite tests for randomness are incomplete (Knuth, 1981), and even the concept of randomness is debated (Lichtenberg and Lieberman, 1992). Linear congruence methods, which perform multiplications modulo a properly chosen large number, receive heavy

use in the constructions of uniformly distributed sequences in [0,1]. Recent calculations solving problems with known solutions show some methods to be better than others (Ferrenberg et al., 1992). It behooves MC users to validate their methods separately for all algorithms.

For MC modeling, algorithms are needed that generate a pseudorandom number sequence according to an arbitrary probability distribution $p(x)$, given a pseudorandom sequence $\{y_j\}$ distributed in [0,1]. This is sometimes done analytically in one dimension (Gould and Tobochnik, 1988), by direct inversion of $P(x) = \int_{-\infty}^{x} dxp(x)$, since $dP(x) = dy = p(x)\,dx$. Direct inversion generalizes to a small dimension m and a rapid MC approximation of $\int d^m x F(x_1, \ldots, x_m)$ only if $F(x_1, \ldots, x_m)$ resembles a tractable probability density function $p(x_1, \ldots, x_m)$. Efficient numerical inversion methods are necessary for higher dimensions.

The powerful Metropolis-Hastings (MH) procedure (Creutz, 1992) was first used 40 years ago. The MH introduces an artificial first-order Markov process and is known as a dynamic MC method. Like direct inversion, MH produces "importance sampling" through which a nonuniform distribution emphasizes regions where the integrand is large (Binder, 1987). It converges by selecting a long string of steps through successive configurations of the random variables, in accordance with general constraints such as the detailed-balance condition. The steps are highly correlated. If an uncorrelated sequence of N configurations is used, as in the simplest MC calculations, estimated errors diminish notoriously slowly, by $N^{-1/2}$. As the sequence becomes more correlated, convergence is slower, because smaller regions tend to be sampled by a fixed number of configurations. The evaluation of when convergence has occurred must consider these correlations.

Models are often formulated from the start for use of the MC method. Subjects of separate monographs are the MC approach to boundary value problems (Sabelfeld, 1991), atmospheric radiative transfer (Marchuk et al., 1980), small quantum mechanical systems (Kalos and Whitlock, 1986), quantum fields (Creutz, 1992), and statistical physics (Binder, 1987). Each subfield has its own set of computational algorithms. We illustrate current activities with two recent accomplishments.

The quantum mechanical ground-state energy E_g of a few-body system satisfies the inequality $E_g \leq E_t$, where E_t is the expectation value of the energy operator using a trial wave function $\mathbf{\Psi}_t(\mathbf{X}, \mu)$. The function $\mathbf{\Psi}_t(\mathbf{X}, \mu)$ is an approximate solution of the time-independent Schrödinger equation. For an n-particle system, $\mathbf{X}$ has up to $4n$ space and spin components, and μ is a parameter vector whose value determines an analytic $\mathbf{\Psi}_t$. For electrons in a molecule, $\mathbf{\Psi}_t$ can contain 10^2 parameters. The parameters are not directly associated with the electronic degrees of freedom. Rather, they give the relative weights of analytically defined Slater determinants and the shapes of state-dependent correlation functions. Evaluation of E_t requires importance sampling of large numbers of values of $\mathbf{X}$, referred to as configurations. This whole process must be repeated many times with a minimization algorithm. Even super-computer power can be too slow to reach convergence. However, the variance of the local energy mean over all configurations of an exact solution is zero. In practice, that variance is adequately approximated using far fewer configurations (by a factor of 10^3) than needed for energy

evaluation. Little accuracy is lost in using the same configurations for successive steps of the parameter vector μ (Umrigar, 1989), and a positive benefit accrues through noise reduction among successive values of the variance, which must be compared. Immense time savings occur from minimizing the variance with infrequent calculations of the energy. As a result, molecular-structure problems can sometimes be done on workstations.

Recent MC calculations of eight particle masses were good to 6% (Butler et al., 1993). They took one year on a special-purpose massively parallel computer, the IBM GF11 computer, where integrals on a space-time lattice were evaluated. Successively larger lattices up to $30 \times 32^2 \times 40$ in size approximated the limit of zero lattice spacing and infinite volume, requiring 10^{17} floating point operations. This result is a breakthrough for computer modeling, for the QCD (quantum chromodynamics) particle model, and for parallel processing. A similar program, the QCD Teraflops Project, is underway. It will use a $128^3 \times 256$ lattice on a platform derived from the Thinking Machines Corporation's CM5 massively parallel, highly connected computer.

3.5 Numerical Modeling

A numerical model is a set of computational algorithms associated with a mathematical model. In order to be complete, all but the simplest mathematical models must contain a numerical model. Just a few examples illustrate this need. The logistic equation,

$$x_{N+1} = 4\lambda x_N(1 - x_N), \qquad (8)$$

where x is real and lies in the interval [0,1], provides the prime textbook example of the period-doubling approach to chaos, as λ increases from 0. Feigenbaum's discovery in 1978 of universality in the bifurcation sequence of this map has been a cornerstone in the development of chaos theory (Schuster, 1988). Yet this iterative equation is also an approximate numerical scheme to solve the differential equation, $dx/dt = ax - bx^2$, with a and b suitably chosen, which has only analytic, nonchaotic solutions. However, the logistic map also approximates a delay-differential equation, which can have chaotic solutions. Delay-difference and delay-differential equations are frequently used in biological modeling. Clearly, studies of their global behavior are important.

Discretization is the approximation of a continuum by points with some minimum spacing, as can be carried out formally to obtain the logistic equation. Alternatively, *spectral models* maintain a partial continuum description in introducing expansions of continuous functions of some space-time degrees of freedom. The expansions may use infinite complete orthonormal sets. The approximation in using them lies in keeping only a finite number of terms. As a byproduct, such truncation reduces the spectral width of the observables, a form of filtering – important both in spectral and discrete methods (Ghil and Malanotte-Rizzoli, 1991; O'Brien, 1986).

Smoothing is a process by which numerical data are either replaced or interpolated by means of other data that display a minimum of short-range fluctuations. Smoothing is standard experimental practice as a preliminary to applying the finite Fourier transform (see Sec. 4.3). This is sometimes accomplished by averaging over a finite subset of nearby data. The data can be results of experimental measurements or results of computations. The smoothing assumptions are that

1. local data fluctuations represent an imperfection best eliminated; or
2. analyticity requirements need to be enforced (as opposed to requiring a specific functional form).

In chaotic systems with sparse data, smoothing is sometimes unwarranted and leads to the inference of spurious low fractal dimensions (Grassberger, 1986). Therefore, the issue as to whether to smooth needs to be settled first. From a systems point of view, smoothing is a special kind of filtering (see Sec. 3.7).

If smoothing is warranted, *splines* represent a computer intensive, yet conceptually simple, approach. The spline process introduces a basis set of functions in one or more dimensions that possess smoothness often defined in terms of existence of higher-order derivatives. The data are partitioned, and the spline basis is fitted within each separate partition so as to satisfy interpartition smoothness criteria (Schumaker, 1980). A cubic spline, in which the cubic polynomial is the basis function, has long been a common option available to modelers (Press et al., 1987). Yet cubic interpolation continues to be refined (Huynh, 1993). Spline smoothing also provides a foundation for nonparametric regression analysis, which in some contexts is an alternative to autoregressive/moving average (ARMA) (Eubank, 1988; also see Sec. 4.3).

3.5.1 Chaotic Systems

An approach sometimes used in numerical modeling of chaotic systems is to create a kind of coarsegraining of the possible states associated with a mathematical model. Each coarse-grained state is assigned a different letter, and the totality of letters is called the alphabet. This approach is referred to as symbolic dynamics. There are other valid ways than that mentioned to create an alphabet, such as, for example, a symbolic classification of the unstable periodic orbits. Here, the impact of the Sharkovsky theorem (Sharkovsky, 1964) becomes evident, as it introduces symbolic dynamics to make fundamental topological statements suggesting a certain set of preconditions for chaos (coexistence of infinite numbers of kinds of periodic solutions). It holds for *all* one-dimensional continuous, unimodal maps on the interval [0,1]. The continuous function $f(x)$ which maps the interval [0,1] into itself is termed unimodal if $f(0) = f(1) = 0$ and $f(x)$ possesses exactly one critical point. Sharkovsky's theorem does not necessarily imply a positive Liapunov exponent (Schuster, 1988). It establishes the following ordering of the natural numbers:

$$3 \succ 5 \succ 7 \succ \cdots \succ 2 \cdot 3 \succ 2 \cdot 5 \succ \cdots \succ$$
$$2^2 \cdot 3 \succ 2^2 \cdot 5 \succ \cdots \succ 2^3 \cdot 3 \succ 2^3 \cdot 5 \succ \cdots$$
$$\cdots \succ 2^3 \succ 2^2 \succ 2 \succ 1,$$

where the symbol $\succ$ is used for "precedes." If $f(x)$ has a cycle of period p then there exist points x' in [0,1] that are members of cycles of period q, for all q that satisfy $p \succ q$. In particular, if one numerically determines the presence of a cycle of period three, then the map $f(x)$ contains cycles of all integral periods. The search goes on for similar insights using symbolic dynamics in higher dimensions (Tufillaro et al., 1992).

The Lorenz equations, which ushered in the modern era of chaos theory (Lorenz, 1963), are themselves a drastically truncated spectral model of a dissipative system undergoing Rayleigh-Bénard convection. Rayleigh-Bénard convection is essentially the motion of a closed fluid system in response to heating from the bottom. It is

known experimentally to exhibit chaotic behavior in regions of its parameter space. Yet the Lorenz equations do not properly describe the real physical system when it is in its chaotic state: this is a case where a model takes on a life of its own, independently of its original motivation. In particular, identification of chaotic behavior as characterized by one-dimensional maps, both through computation (Lorenz, 1963) and experiment in similar systems (Brandstater and Swinney, 1987), reveals how low dimensionality of the fractal attractor provides a simplifying insight into complex phenomena. The Lorenz equations, a set of three coupled first-order differential equations, require further numerical modeling (discretization). There is continued interest in the dependence of chaotic solutions upon numerical algorithms.

The local evolution of chaotic systems is extremely sensitive to discretization errors; however, the global properties, such as the Liapunov exponents $\{\lambda_i\}$, the Kolmogoroff-Sinai entropy K_{KS}, and the generalized fractal dimensions $\{D_q\}$, are not (Schuster, 1988; see also Secs. 3.1 and 4.5). All chaotic states have fractal orbits. Therefore, determining that at least one measure of dimensionality of the orbit, such as the Hausdorff dimension, is nonintegral and hence fractal is suggestive that chaos is present. For the logistic map, a wide range of values of λ gives chaos (Schuster, 1988). The asymptotic chaotic orbit for the value $\lambda = \frac{9}{8}$ is just the Cantor middle-thirds set discussed in Sec. 3.1.

The maximum Liapunov exponent λ_0 is the real litmus test for chaos, because only the inequality $\lambda_0 > 0$ yields exponential divergence of nearby orbital points, assuring SIC. Smale's horseshoe map discussed in Sec. 3.1 has two Liapunov exponents. They are both computed from a relation derivable from the general definition (Schuster, 1988) in this special case: $\Delta x' = \exp(\lambda \Delta t)\Delta x$. We take $\Delta t = 1$ for one iteration step, and $\Delta x'$ and Δx are the new and the old distances between nearby points after one iteration measured

1. along the direction of stretching in the asymptotic chaotic orbit and
2. perpendicular to that direction.

The Liapunov exponents are $\lambda_0 = \ln 2$ and $\lambda_1 = \ln(1/2\delta)$. *Dissipative chaos* corresponds to the present case in which the sum of Liapunov exponents is negative, which implies that an elementary volume element in the space of the dynamical system evolves to zero. *Conservative chaos*, in which the sum of all the Liapunov exponents is zero, denotes systems in which the size of a volume element evolves unchanged. Conservative chaos occurs in Hamiltonian systems and is heavily investigated for its relevance to the foundations of classical and quantum mechanics (Ford and Mantica, 1992).

The 1980s saw successful numerical modeling, much of it using time-delay embedding (Abraham et al., 1990; Schuster, 1988). Improvement is sought in the study of experimental data. A typical exercise is to generate small, noisy data sets from known chaotic attractors and to try to recover their chaotic properties. There are procedures to identify spurious Liapunov exponents generated by time-delay embedding (Parlitz, 1992).

3.5.2 Finite Elements

Finite elements is a form of numerical analysis devoted to continuum systems such as solid structures and electromagnetic fields (Cook et al., 1989), in which continua are discretized into highly symmetric elements. For example, these might

be triangles, tetrahedra, or hexahedra. Algorithms are then developed for interpolating between nodes. Finite elements is also applied in simple quantum mechanical systems of low symmetry for which analytic solutions do not exist (Shertzer, 1989). Large generalized eigenvalue problems sometimes result, with variational bounds.

3.5.3 General Circulation Models

The GCM's coupling the atmosphere, ocean, land, biosphere, and cryosphere on global scales are faced with Herculean tasks. This is because space-time discretization must be so coarse-grained. Consider the global atmosphere. Even 10^6 cells, counting 20 vertical levels, make a sparse approximation to a continuum and corresponds to somewhere in a mid-mesoscale size, far too large to simulate the physics of clouds, precipitation, turbulence, or local storms or eddies. Therefore, the dynamical and thermodynamical laws must be averaged. If the horizontal and vertical scales are reduced by a factor of 100 and if the order of 10 quantities are calculated at each time, a microscale description of the atmosphere could result in which a teraflop computer would require 10 s to simulate one time step. Within such a discretization, eddies might be resolved, but clouds, precipitation, and turbulent dissipation still need to be parametrized. For oceanic phenomena, characteristic time scales are orders of magnitude larger and length scales orders of magnitude smaller.

Decrease in GCM computation times has been achieved with hybrid spectral and discretization methods (Peixoto and Oort, 1992). The horizontal fluxes are expressed in terms of a truncated spherical-harmonics expansion; the vertical and the time coordinates are discretized. Spherical harmonics make a complete set on a sphere. Figure 2 shows a perturbation of a sphere by a higher-order spherical harmonic. It was produced by a few simple MAPLE graphics commands. When discretizing, modelers work within required upper bounds to products of $\Delta t \Delta x$, the space and time intervals, and employ filters to remove fluctuations corresponding to unwanted sound and gravity waves (O'Brien, 1986). Different numerical modeling schemes are regularly compared.

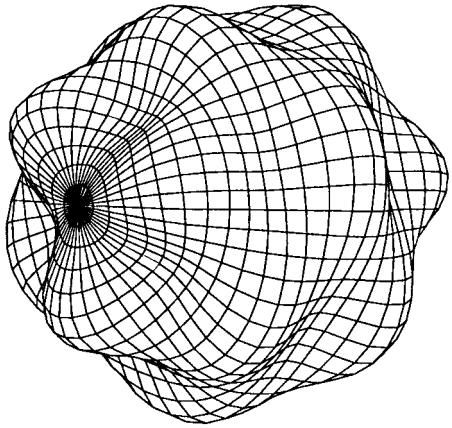


Fig. 2 Three-dimensional plot of a small perturbation of a sphere by the surface spherical harmonic Y_7^3. Requiring a few short commands in MAPLE, this illustrates computer visualization in modeling

3.5.4 Data-Assimilation Models

The goal of several-day weather prediction has been achieved through supercomputer calculations of atmospheric circulation fields on fine grids, combined with updates, i.e., corrections, from measurements. Sophisticated assimilation models have been developed for atmospheric and oceanographic purposes (Ghil and Malanotte-Rizzoli, 1991). Assimilation models are needed because of lack of knowledge of a complete set of initial conditions, inaccuracies in the numerical

models, and atmospheric chaos. Atmospheric sampling and modeling are good enough for the National Meteorological Center (NMC) to create ensembles of computer runs that, combined, make possible predictions and estimates of reliability. In contrast, experimental sampling of oceanic temperature, salinity, and velocity fields is poor and will continue to be so in the foreseeable future. Its ill-posed nature makes oceanic circulation the greater challenge. Many techniques are used, including two discussed in this article – singular-value decomposition and Kalman filters (Ghil and Malanotte-Rizzoli, 1991; Carter, 1989).

3.6 Wavelets

A wavelet is an integral transform designed to perform a localized time-frequency analysis. It provides a microscope with which to look at details of a function's spectrum within a small time interval centered about some arbitrary time. Taking its early motivation from geophysical applications, the wavelet has undergone robust development mathematically and in application (Daubechies, 1992). The wavelet transform provides a useful tool to study signals that, like speech, impart information that evolves in time. There is a discrete wavelet transform, and it is important, but we shall concentrate here on salient defining characteristics of the continuous transform. Wavelets and windowed Fourier transforms are sometimes similar. Remarkably, the slight generalization of the windowed Fourier transform in the wavelet transform can make a vast difference.

Consider a one-dimensional Fourier transform:

$$g(\omega) = \int dt e^{-i\omega t} f(t). \tag{9}$$

The Fourier transform takes one from the time (t) domain to the frequency domain (ω is the angular frequency). The windowed Fourier transform replaces $f(t)e^{-i\omega t}$ with

$$f_{\text{win}}(t)e^{-i\omega t} = f(t)h(t-s)e^{-i\omega t}, \tag{10}$$

to produce $g_{\text{win}}(\omega)$. Here, h is a simple function, such as a Gaussian, centered around $t - s = 0$, giving a smooth cutoff for large values of its argument. This transformation does not yet provide the microscope – just the localization. In the wavelet transform, the function $he^{-i\omega t}$ is replaced by a function $\mathbf{\Psi}$, referred to as the "mother wavelet," which there is considerable leeway to define:

$$|a|^{-1/2}\mathbf{\Psi}\left(\frac{t-b}{a}\right) \tag{11}$$

subject to the condition

$$\int dt \mathbf{\Psi}(t) = 0. \tag{12}$$

For acceptable $\mathbf{\Psi}$, a serves as a microscope parameter and b a localization parameter. As a decreases, the wavelet transform $g_{\text{wav}}(\omega)$ zooms in to a decreasing width of times, centered around b. The function $\mathbf{\Psi}$ is reminiscent of, but not equivalent to, the spectral window introduced in the Fourier transform (Priestley, 1988).

The form $\mathbf{\Psi}(t) = (1 - t^2)\ \exp(-t^2/2)$, sometimes called the Mexican hat function, is a typical choice for $\mathbf{\Psi}$. Significant progress has been made in developing orthonormal bases for the mother wavelet. A mother wavelet and associated orthonormal bases of compact support have also seen considerable attention. In analogy with Fourier analysis, the wavelet transform can often be inverted to recover the original function, which can thereby be formally expressed in terms of the wavelet basis (Daubechies, 1992).

Wavelets promote systematic investigations of data. Turbulence is an application of wavelet transforms, because theories of turbulence point to self-similarity, which wavelets elegantly characterize (Vergassola and Frisch, 1991). Wavelets also provide a tool for studying multifractals.

3.7 Systems Analysis

A system is a process or a structure or both combined. In *systems analysis* (SA), the relationships among a system's parts are studied. SA is concerned with explanation and prediction, when applied to natural systems, and achieving a goal, when applied to constructed systems. Natural systems include climate, living organisms, neural networks, a flow into a pile of grains of sand, and much more. Constructed systems include manufacturing processes, servomechanisms, circuit networks, communication devices, and artificial neural networks. Social hierarchies could be considered either natural or constructed. The same concepts from SA can be universally applied. In natural systems, *synergetics* is a SA that seeks extensions of physical principles required for explanation of self-organization. In climate and weather, SA is used for mechanistic modeling and for understanding simulation models. Life processes such as self-replication, evolution, and cell specification are another endeavor of SA. SA is a framework for mechanistic modeling.

Frequently encountered concepts are feedback, control, filter, and optimality. Unfortunately, working definitions vary. The concepts are precise in engineering. In signal processing, a system is any process that transforms signals, a signal being a function of one or more independent variables. Linear systems, feedbacks, filters, and control theory are standard in engineering curricula. Unless formal, SA heavily employs flow diagrams. An overview of applications and modern research, including the theory of nonlinear systems, is in a Kalman *Festschrift* (Antoulas, 1991).

Filter is a name for a dynamical system that performs a task such as removing some unwanted variance, such as noise, in a signal. As viewed in time-series analysis (Priestley, 1988), a linear and time-invariant system which relates input U_t to output V_t through the equation

$$V_t = \sum_{s=0}^{\infty} a_s U_{t-s} \tag{13}$$

is a filter. Filters can be more generally defined and are not necessarily linear. A Kalman filter, in its simpler form, is a linear finite-state process that optimally estimates the state of a linear dynamical system in the presence of noise and deterministic control input. Optimality is defined as minimization of variance. The Kalman filter becomes a unique, recursive algorithm and can be adapted to smoothing, prediction, or control. Unlike earlier filters, the Kalman filter works for nonstationary time series. The Kalman filter has been extended now for use with nonlinear systems. A textbook example is to estimate the orbit of a satellite (Chui and Chen, 1987).

3.7.1 Control Theory

Viewed dynamically, a system is describable in terms of an n-dimensional state vector, $\mathbf{X}(t, \mathbf{\Theta})$, where $\mathbf{\Theta}$ can be an m-dimensional control vector. Much of control theory is a subfield of nonlinear dynamics and can be stated in terms of either maps or flows. When the system is nonlinear, bifurcations can sometimes occur

and be relevant. One problem in control theory is to find an optimal control vector $\mathbf{\Theta}(t)$ such that some state $\mathbf{Z}(t)$ is reached. Optimality can be defined in terms of minimization of cost or of time taken to reach $\mathbf{Z}(t)$, within defined constraints on $\mathbf{\Theta}(t)$.

The concept of a *closed-loop* or *feedback control* arises simply in a linear stochastic system as represented by a vector flow equation

$$\frac{d}{dt}\mathbf{X}(t, \mathbf{\Theta}) = F(t)\mathbf{X}(t, \mathbf{\Theta}) + G(t)\mathbf{w}(t) + L(t)\mathbf{\Theta}(t) \qquad (14)$$

where $\mathbf{w}(t)$ is a noise process, and $F(t)$, $G(t)$, and $L(t)$ are gain operators. In particular, $L(t)$ contains the influence of the control vector $\mathbf{\Theta}(t)$ upon the state vector $\mathbf{X}(t, \mathbf{\Theta})$. If $\mathbf{\Theta}(t)$ is a function of time alone, then this is *open-loop control*, while if $\mathbf{\Theta}(t)$ is explicitly a function of $\mathbf{X}(t, \mathbf{\Theta})$ also, then this is called a closed-loop or feedback control.

Stringent necessary conditions for attainability, reachability within a fixed interval, and controllability are more easily obtained for linear than nonlinear systems. Such conditions, sometimes for open-loop control, can be useful to mathematical modeling. Consider, for example, the growth of several competing species controlled by a single essential growth-limiting nutrient. There is a well-known principle of competitive exclusion: if the rate of input and washout are constant, then at most one species survives. It turns out that extinctions do not always have to occur. Input and washout rates can be allowed to vary with time. Under rather general conditions, a nonlinear control-theoretic mathematical model establishes necessary and sufficient conditions for admissible controls [i.e., an admissible nutrient concentration $s(t)$] such that all n species can coexist, and it gives insight into producing controlled selective extinctions (Rao and Roxin, 1990).

Atmospheric and oceanographic modeling often focuses on natural feedback loops and can use a SA approach. In unraveling the impact of atmospheric CO_2 doubling, for example, modelers are concerned about many feedback loops; the radiation – cloudiness – temperature feedback, for example, has only recently been understood to be globally stabilizing (Levi et al., 1992). A more explicit SA model was for large-scale climate oscillations in the global glacier-ocean-atmosphere system (Sergin, 1980). Schlesinger (1989) seeks to unify the climate modelers' nomenclature by employing standard electrical engineering definitions of feedback. Figure 3 shows a block diagram for the climate system adapted and redrawn from his work. In it, the "feedback factor" of the climate system is $f = GF$, where G is the gain, F is a measure of the feedback, and the input forcing, caused for example by increase of CO_2, is ΔQ. When no feedback is present,

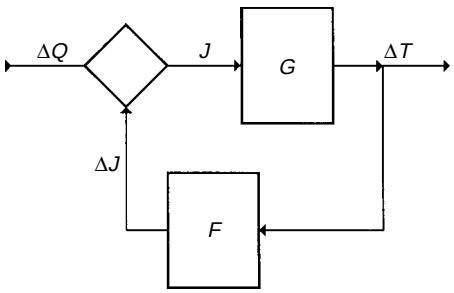


Fig. 3 Block diagram for natural climate feedback control. In the absence of feedback ($F = 0$), an amount of input forcing heat ΔQ produces an output J, which multiplied by the gain G equals the temperature change response ΔT. In the presence of feedback ($F \neq 0$), the output is changed by $\Delta J = F\Delta T$, causing ΔT to be multiplied by $1/(1 - f)$, with a feedback factor $f = GF$

$J = \Delta Q$ and the surface temperature response is just $\Delta T = GJ = G\Delta Q$. When there is feedback, the output J is increased by $\Delta J = F\Delta T$ to become $J = \Delta Q + F\Delta T$. Then

$$\Delta T = \frac{G}{1-f}\Delta Q. \tag{15}$$

A positive feedback factor destabilizes, while a negative factor diminishes the response (the case $f \geq 1$ is excluded). Numerical simulation experiments and ingenious physical experimentation are ingredients in SA modeling like this. Low-dimensional climate models viewed as spatial averages, with few competing feedbacks, have also produced insights (Peixoto and Oort, 1992).

3.7.2 Self-Organization

Self-organization is modeled in many ways. These include specializations and extensions of known physical models such as thermodynamics and dynamical systems theory, as well as control theory. Neural networks have been employed. Further, a relatively new field, synergetics, pioneered by H. Haken, is dedicated to the study of structures arising within far-from-equilibrium systems that manifest self-organization (Yates, 1987). Synergetics has enunciated a slaving principle, supported by various examples both from animate and inanimate phenomena. Essentially, slaving occurs when long-lasting quantities serve as order parameters in transient states and dominate over short-lasting quantities. For example, when pump power in a laser goes from weak to strong, the laser's fluctuating dipoles become coherently enslaved by the external field. At just enough pumping, a bifurcation occurs, symmetry is broken, and the motion of each dipole is enslaved by the external electric field. In such a manner, it is posited that enslaving creates a hierarchic structure that reduces the number of degrees of freedom of complex systems. Synergetics attempts to describe the development and operation of living structures in this way.

Fractal structures abound in large systems. The suggestion of a fractal structure is an inverse power-law (or $1/f$) behavior of a temporal or spatial power spectrum $S(f)$, taken to mean $S(f) \propto f^{-\alpha}$, with $\alpha > 0$ (West and Shlesinger, 1990). Some models predict $1/f$ phenomena associated with critical states toward which large systems naturally evolve. Such a self-organized criticality stems from work initiated by Bak and his collaborators (Bak et al., 1987). Using CAs to simulate grain-by-grain formation of a sandpile, they reached a critical state. In the critical state, temporal self-similarity (yielding flicker noise) arises, and spatial self-similarity is also observed in the comparable probabilities of large and small avalanches. Self-organized critical states are only weakly chaotic, neighboring points showing power-law rather than exponential divergence. This modeling has stimulated theoretical and experimental work; a notable success is a power-law prediction for energies and spatial distributions of earthquakes (Lichtenberg and Lieberman, 1992).

4 Information Processing Models

4.1 Principal-Component Analysis

Principal-component analysis (PCA) is a method of analyzing the variance within a data stream for the purpose of finding a reduced set of variables. Ideally, these variables should be

1. statistically uncorrelated and
2. responsible for most of the variance.

Property **1** is automatically achieved, but property **2** is just a possible outcome. In analogy with Fourier series, an orthonormal basis set is obtained, but with PCA, the basis is empirically defined. After the advent of the digital computer, outstanding successes in meteorology and oceanography (Preisendorfer, 1988) and signal processing have marked the history of this 90-year-old method. The PCA has many applications, sometimes used as a low-pass filter in conjunction with statistical regression. Another major use of PCA is in chaos research, to help separate extraneous noise from chaotic variability.

A standard first step is to construct a symmetric covariance matrix of space-time events. We use bold face for matrices but not their components, superscript T for transpose, and matrix notation for vectors. Consider a time series of p-dimensional vectors $\mathbf{x}_i$ of some observable scalar field, such as temperatures at p different places, measured at n times. Assume that these are all zero-mean measurements, i.e., they are all given relative to their mean values. The sample covariance matrix $\mathbf{S}$ for the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ has as its (j, k)th element

$$\frac{1}{(n-1)} \sum_{i=1}^{n} x_{ij} x_{ik}, \tag{16}$$

where x_{ij} is the jth component of the ith vector. This is a symmetric matrix diagonalizable by a matrix whose columns are its orthonormal eigenvectors. Here we adopt geophysical terminology and refer to an eigenvector $\mathbf{a_r}$ as an empirical orthogonal function (EOF). In statistics literature, it is referred to as a principal component. The matrix $\mathbf{S}$ is non-negative; the eigenvalue l_r for each nontrivial solution is positive and is the variance associated with its eigenvector, $\mathbf{SA_r} = l_r \mathbf{A_r}$. This eigenvalue problem satisfies a variance maximization principle. Hence, the first EOF, having the largest eigenvalue, is the normalized linear combination of the p variables with maximum variance. The second eigenvector is the combination of largest variance such that it is orthogonal to the first, and so on. Each of the n data vectors is expandable in terms of the p EOF's. In geophysical terminology, the time-dependent expansion coefficients of the n data vectors are the principal components (PCs). Criteria for dropping terms associated with small variances, called selection rules, are developed in the context of the applications. In one of the first computer-assisted PCA analyses, E. N. Lorenz at MIT made prediction studies in 1956 of a 500-mbar-height anomaly field over North America, using five years of data obtained at 64 stations. The first nine eigenvectors obtained accounted for 88% of the variance for a one-day prediction (Preisendorfer, 1988).

Singular-value decomposition (SVD) is a different but equivalent formulation of this eigenvalue problem, based on an SVD theorem showing that an arbitrary $n \times p$ matrix is expressible in the form $\mathbf{X} = \mathbf{ULA}^{\mathrm{T}}$. Here, $\mathbf{L}$ is an $r \times r$ diagonal matrix, where r is the rank of $\mathbf{X}$, and $\mathbf{U}$ and $\mathbf{A}$ are matrices with orthonormal columns. In a PCA application, the elements of $\mathbf{L}$ are proportional to $l_r^{1/2}$, where l_r is the rth eigenvalue of $\mathbf{S}$. SVD has an algorithm for finding EOFs.

PCA is also adaptable to data vectors $\{\mathbf{x}_i\}$ having components that represent different kinds of physical quantities, such as pressures and temperatures. A PCA is then usually defined in terms of a correlation matrix, formed by dividing each element x_{ij} of the n data vectors by

$\sigma_{jj}^{1/2}$, where σ_{jj} is the variance among the jth components, the variance being taken over the columns in x_{ij} defined by fixed values of j.

4.2 Neural Networks

A *neural network (q.v.)* is a system containing a highly interconnected set of elements inspired by characteristics of the brain. In one simple formulation, the elements (called nodes or neurons) fire (i.e., are placed in a binary "up" state of $+1$ as opposed to 0) when the summed weighted inputs S, both external and from all other elements, reach some threshold condition, $S \geq T$. A node thereby receives multiple inputs but produces a single output. A typical nonlinear feedback to a node is the sigmoid function $1/[1 + \exp(-Gz)]$, with G the gain parameter and z the sum of binary inputs from other nodes, each input taking values ± 1. Then $+1$ acts as an excitation and -1 an inhibition. Generally, a neural network as a whole accepts multiple inputs and delivers multiple outputs. There is a wealth of significantly differing neural networks using these simple ideas, along with feedforward, feedback, nonlinear dynamics, and hidden layers. A hidden layer is a collection of nodes not directly connected to input or output. The weights can generally be made to change in the training mode of operation. *Back-propagation* is the name for algorithms by which output errors are sent back to a hidden layer to process changes of the weights (Zurada, 1992).

Neural networks have had a slow beginning from their earliest inceptions 50 years ago, which antedated Rosenblatt's feedforward "perceptron" of 1962. But now the effort devoted to neural network research and implementation has become staggering. The reason is partly that neural networks can be trained to perform easily and quickly valuable pattern recognition tasks. Such tasks as optical character recognition (OCR), financial forecasting (through function estimation), and manufacturing process control are among present major uses of neural networks (Hammerstrom, 1993). Neural networks are parallel structures, and recently developed parallel integrated circuit devices have increased performance speeds by factors of 10^3. Training times tend to be long, on the order of months, but the trained neural network can often perform in real time. A drawback is that although the nonlinearity and feedback structures of neural networks mirror complexities of real systems, their results, even when correct, are difficult to explain.

Some implementations have a capacity for unsupervised learning. Some of the Hopfield type, with symmetric connections, achieve associative memory, and neural networks over a wide range of complexity can become chaotic. Neural networks are used in meteorology, where their pattern recognition and predictive skills are competitive with older methods (Elsner and Tsonis, 1993).

4.3 Time-Series Analysis

A *time series* is a record of the values of any set of fluctuating quantities measured at different times. In the analysis of time series, regularities are sought, sometimes for making predictions. If there is more than one fluctuating quantity, the time series is called *multivariate*; otherwise, it is called *univariate* or scalar. A time

series can be continuous or discrete. Time-series analysis is used in information theory, dynamical systems theory, systems analysis, and data assimilation.

Recent work is considerably beyond the pioneering 1956 success of Lorenz, for using PCA to uncover dynamical processes in multivariate time series (Vautard et al., 1992) and to forecast. A related El Niño forecast is discussed in Sec. 5.1 (Keppenne and Ghil, 1992). The analysis of noise is often crucial: for example, whether determinism in paleoclimate variability could be construed from the present climatic time series data hangs in the balance of the inference of noise (Vautard and Ghil, 1989). Retreating from criteria for chaotic determinism, one can settle for less and test for the existence of nonlinearity in short, noisy series (Theiler et al., 1992). Testing for nonlinearity uses advanced statistical inference techniques, to be discussed. In spoken communication, where the temporal behavior of the sound pulse is vital, signal processing has made use of SVD and PCA applied to the discretized data stream.

The modeling of time series is generally begun with the definition of a *strict white noise* process, denoted by $\{e_t\}$. For a stationary (i.e., exhibiting time-independent statistical behavior) strict white noise process, the mean μ and variance σ^2 are independent of time t. In any case, the autocovariance function, $\mathrm{cov}\{e_t, e_{t'}\}$, is zero for all $t \neq t'$. Here we specialize our discussion to the covariance function for a univariate continuous random variable X_t, given by

$$\mathrm{cov}\{X_t, X_{t'}\} = E[(X_t - \mu)(X_{t'} - \mu)]. \quad (17)$$

The expectation, or mean, is just $E[X_t] = \mu$, a constant whenever X_t is stationary.

Modeling in terms of a random variable such as X_t implies the existence of an associated probability distribution p_t, which is time dependent unless X_t is stationary. If X_t can take a continuous range of values, $x \in [a, b]$, then p_t is called a *probability density function*. One then defines the *expectation* E of X_t at time t by

$$E[X_t] = \mu_t = \int_a^b x p_t(x)\, dx. \quad (18)$$

Finally, the *variance* of X_t is $\sigma^2 = \mathrm{cov}\{X_t, X_t\}$. Such statistical quantities are estimated from the data stream. A computational method is often a basis for choosing a form of time-series analysis.

Although for brevity we explicitly refer to scalar time series, the extensions to the multivariate case are generally necessary and straightforward. Even scalar series from a dynamical system become multivariate through time-delay embedding. Consider now a discretized time series $\{X_t\}$ with t taking on only integer values. A model for this series is an equation

$$F\{\ldots X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, \ldots\} = e_t, \quad (19)$$

where e_t is a zero-mean strict white noise process and $F\{\cdot\}$ is the model function. Finding the model function is equivalent to reducing the unexplained part of the series to a strict white noise process. Well-known linear model functions include AR (autoregressive), MA (moving average), and ARMA (autoregressive/moving average). The ARMA(k,l) model includes the AR(k) and MA(l) models as special cases and is expressible as a linear operator: $X_t = G(B)e_t$, where $G(u)$ is the rational function $G(u) = \beta_l(u)/\alpha_k(u)$, and u is replaced by the linear back-shift operator, $B : Be_t = e_{t-1}$. Also

$\beta_l(z) = 1 + b_1 z + \cdots + b_l z^l$ and $\alpha_k(u) = 1 + a_1 u + \cdots + a_k u^k$. This process can be pictured as representing the action of a linear filter. The estimation of the unknown parameters $k, l, a_1, \ldots, a_k, b_1, \ldots, b_t$ is a model-fitting problem. For convergence with this model, it is necessary that the time series be stationary; also, consistency requires $G(u)$ to have no poles for $|u| \leq 1$. An ARMA model works only if the system is stationary. Optimal simulation and parameter estimation both for stationary and nonstationary time series can be done through modeling of filters.

An explicit representation of nonlinear series arises from inverting and expanding the function F. Just up to the first nonlinear term, the resulting discrete-time Volterra series is

$$X_t = \mu + \sum_{i=0}^{\infty} h_i e_{t-i} + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} h_{ij} e_{t-i} e_{t-j} + \cdots. \quad (20)$$

A way to look for nonlinearity is to compute polyspectra. Polyspectra are multiple Fourier transforms of third- and higher-order moments of the series about its mean. If $\{X_t\}$ is a Gaussian process, the bispectrum and all higher-order polyspectra vanish (Priestly, 1988).

4.3.1 Signal Processing

With digital technology come new and powerful modeling techniques following quantization. A signal is quantized by sampling it at a time interval T, the sampling period, and rounding to a certain fixed number of significant figures. The quantized signal (and also the unquantized one) can be subjected to *spectral analysis*, which is the determination of its frequency content. The Fourier transform, in either its discrete or its windowed version, permits useful analysis in the frequency domain. Finding the finite Fourier transform is a common step in spectral analysis. It can now be performed in real time by use of the fast Fourier-transform (FFT) algorithm proposed by Cooley and Tukey in 1965 (see FOURIER AND OTHER MATHEMATICAL TRANSFORMS). The power spectrum $S(f)$, which is the square of the magnitude of the Fourier transform, provides the most accessible vital characterization of a signal. For a chaotic system, the windowed power spectrum must be continuous and not just contain a large, finite collection of peaks. Quasiperiodicity and complicated periodicity can be identified but not usually distinguished in the power spectrum. The Fourier transform creates a frequency-domain representation sometimes used in ARMA, other parameter-estimation algorithms, and SVD (Kumaresan, 1993). Meteorological and oceanographic PCA can also be profitably recast through Fourier transforms of the time or the azimuth variables, to focus upon wave structures (Bernardet et al., 1990).

Discrete-time analysis uses the z transform, a generalization of the Fourier transform. The z transform of a signal expressible as a sum of exponentials can be written as a rational function in the variable $z = \exp(i\omega)$. This allows a "pole–zero" approach to signal analysis and the use of non-numerical, symbol manipulation modeling of signals. In the context of digital signal processing, it is easy to understand various pitfalls of spectral analysis such as *aliasing*, the misidentification of power spectrum frequencies because of low sampling rates (Oppenheim and Schafer, 1989).

4.4 Statistical Inference

Statistical inference in the computer age has dramatically changed (Efron and Tibshirani, 1991) and is developing rapidly. Powerful formal results such as the central limit theorem, which establishes asymptotic normal distributions (Priestly, 1982), are more useful than ever, because digital computers help reach the asymptotic region in a statistical analysis. More important is that computer experiments make it possible to learn some of the limitations of statistical inference methods. Estimation errors are more easily evaluated from finite sampling of infinite populations, noise, and finite sampling of finite populations. With the use of current computer-intensive data processing techniques, the original data can be

1. sampled through partitioning or through successive computer-assisted selection models;
2. altered in definable ways to generate surrogate data; or
3. reorganized along paths linking common characteristics.

The jackknife and bootstrap methods exemplify **1**, randomizing phases of Fourier-transformed data exemplify **2**, and the CART (classification and regression trees) method exemplifies **3**. Given small and imperfect historical data sets of natural phenomena such as sunspot activity, global temperatures, or various proxy data for global climate (Peixoto and Oort, 1992), processing techniques give new meaning to calculations of fractal dimensions (Vautard and Ghil, 1989; Theiler et al., 1992) in searches for underlying nonlinear dynamics and possibly chaos. Theories of processing of samples of finite populations are developed along similar lines by Särndal et al. (1992).

Simplicity serves as a guiding principle in model development (Harte, 1988). The Bayesian approach to scientific inference, though controversial, gives support to simplicity. The Bayesian approach asserts that probability can be associated with plausibility in evaluation of theories. In particular, a theory that gives the higher probability to what is actually observed is the more likely to be correct. Bayes's theorem, proved by the English clergyman in 1761, is suggestive of such an approach (Jefferys and Berger, 1992): Let $P(X|Y)$ be the conditional probability that hypothesis X is true, given information Y. Then Bayes's theorem states

$$P(H_i|D\&I) = \frac{P(D|H_i\&I)P(H_i|I)}{P(D|I)}. \quad (21)$$

This theorem relates the contingent probability of the truth of hypothesis H_i to the prior information I and to the new data D. In this way, Bayesian inference supports simplicity: Consider an alternative, more complicated hypothesis H_j. If the new data are more sharply predicted by H_i than by H_j, then this can increase $P(H_i|D\&I)$. But by their nature, simpler theories tend to have fewer possible outcomes over which to spread the probability, tending to make sharper predictions.

Less controversial is another plausible inference method, the method of maximum likelihood, which also has an analogy with Bayes's theorem. Maximum likelihood is used in the same contexts as least-squares estimation: Assume that the probability density for a measured vector $\mathbf{X}$ is $f(\mathbf{X}, \mathbf{\Theta})$, where the parameter vector $\mathbf{\Theta}$ is to be estimated from the data. One reinterprets $f(\mathbf{X}, \mathbf{\Theta})$ to signify the likelihood for a particular choice of $\mathbf{\Theta}$, given

the measurement $\mathbf{X}$. The parameter estimate is reduced to finding that parameter vector $\hat{\mathbf{\Theta}}$ that maximizes f or $\ln(f)$. Under general conditions, the estimator $\hat{\mathbf{\Theta}}$ can be proved to have a normal distribution asymptotically when the number n of measurements gets large (Priestly, 1982). With computing power available, the details of the large-n limit become more tractable, increasing the significance of such formal results.

4.5 Information Theory

Information theory was developed as a mathematical model of communication by Shannon in 1948. It considers transmission rates in the presence of noise (Gallager, 1968). Defined by analogy with physical entropy, the information H is given by

$$H(Q) = -\sum_i p_i \log_2(p_i), \tag{22}$$

where p_i is the probability of a possible message i. Here, Q represents the space of all possible messages, and $H(Q)$ is the information gained, on the average, from measuring which message occurs. All messages are treated as statistically independent, and $\Sigma_i p_i = 1$. A continuum formulation is also often used. An important quantity in communication theory is *mutual information*, $I(Q, S)$, defined in the space of ordered pairs of outcomes from Q and S, and given by

$$I(Q, S) = H(Q) + H(S) - H(S, Q). \tag{23}$$

Given a measurement in S, $I(Q, S)$ gives the number of bits of information, on the average, that can be predicted about a subsequent measurement in Q. Here, I can represent a rate with which information crosses a channel in the presence of noise, computable from the conditional distribution function $P_{q|s}(q_i, s_j)$, where $q_i \in Q$ and $s_j \in S$. Shannon's channel capacity theorem employs $I(Q, S)$.

Information is an important concept for dynamical systems. So are extensions of information, following work in the late 1950s of Kolmogoroff and Sinai (Schuster, 1988). The Kolmogoroff-Sinai (or metric) entropy K_{KS} and an infinite sequence of generalized entropies are computable from time series, where K_{KS} is a measure of the rate at which information is generated in a map or a flow. Chaotic systems constantly generate information; consequently, the information existing at any one moment is insufficient to make long-range predictions. Hence, the long-range unpredictability of the weather (viewed as a deterministic chaotic system) is explained in information-theoretic terms. A random sequence has an infinite K_{KS}, because the same finite amount of information ΔH is generated at successive times, no matter how small the time interval Δt, and hence $dH/dt = \infty$. Mutual information is also definable for a map or flow, giving the number of bits of information, on the average, that can be predicted about the measurement of the state at time t from having measured the state at a time $t - T$. When time-delay embedding is used, the delay time τ taken as the first minimum of the mutual information of the vectorized series is more optimal than a τ chosen as the first zero of the correlation function (Fraser and Swinney, 1986).

Ergodic theory is related to information theory. By the ergodic property, the time average of an observable equals its phase-space average. Only when a dynamical system is ergodic are most of the procedures described in this article

for analyzing dynamical systems valid. Formal theorems exist using K_{KS} to compare ergodic systems (Schuster, 1988; Lichtenberg and Lieberman, 1992).

5 Applications

5.1 El Niño

El Niño is an anomalous persistent warming of the eastern equatorial Pacific that occurs irregularly, generally every two to five years. Invariably, this sea-surface temperature (SST) anomaly is accompanied by a reversal of the usual low atmospheric pressure in the western equatorial Pacific and high pressure in the eastern equatorial Pacific. One measure of this is the southern oscillation index (SOI). The SOI is a scalar time series that can be constructed as a normalized and smoothed sea-level pressure difference between Darwin, Australia, and Tahiti (Keppenne and Ghil, 1992).

A mechanism for El Niño seems clear: positive feedback between the equatorial easterly trade winds and the SST gradient. When the pressures reverse, the mean wind forcing of the Pacific surface waters toward the west stops, and as a consequence, the warmer surface waters accumulate off the western South American coast. Upwelling of cold eastern Pacific water stops. This promotes a convergence of warm moist air, cloudiness, precipitation, and low-surface atmospheric pressure. This positive feedback works in the other direction, as well, to produce a cold SST episode (La Niña). The two coupled phenomena are referred to as ENSO (El Niño–Southern Oscillation). It is ENSO that is modeled (Philander, 1990). Predicting ENSO would have practical importance to the South American fishing industry, as ENSO episodes are destructive to west coastal fish abundance. Various globally distributed largescale atmospheric and oceanic phenomena are correlated with ENSO, referred to as "teleconnections."

Modeling ENSO poses a major challenge. So far, mechanistic models can reproduce the phenomenon qualitatively, without necessarily being reliable predictors either of the occurrence or the structure of episodes. The ENSO occurs in a large variety of ways with respect to frequency, intensity, duration, and spatial distribution (Philander, 1990). The extent to which GCM's will be able to describe and/or predict ENSO and its teleconnections is an open question. To what extent tropically based mechanisms are sufficient to trigger ENSO episodes is an important theoretical question. On a more technical level, the role played by the annual weather cycle and the relationship of the quasibiennial oscillation (QBO) to ENSO, if any, require clarification. The QBO is a tropical lower stratospheric fluctuation of zonal winds of approximately a 2-yr period.

One simple mechanistic model of ENSO is Vallis' one-dimensional, two-point description, the equations of which resemble the Lorenz system. In his model, a mean Pacific equatorial wind creates a mean westward ocean flow, with feedbacks. The equations are

$$\frac{du}{dt} = \frac{B(T_e - T_w)}{2\Delta x} - C(u - u^*); \quad (24)$$

$$\frac{dT_w}{dt} = \frac{u}{2\Delta x}(\bar{T} - T_e) - A(T_w - T^*); \quad (25)$$

$$\frac{dT_e}{dt} = \frac{u}{2\Delta x}(T_w - \bar{T}) - A(T_e - T^*). \quad (26)$$

Here, T_e and T_w are east and west upper-ocean temperatures separated at a distance Δx, $\bar{T}$ is the deep-ocean constant temperature, and u is eastward oceanic current. The terms A, B, C, u^*, and T^* are constants. The mechanism is simply that an eastward (westward) wind field driven in part by a temperature difference $T_e > T_w (T_e < T_w)$ would produce an oceanic downwelling (upwelling) in the east and upwelling (downwelling) in the west. The equations conserve mass. Low-dimensional chaos is produced, as can be inferred by evaluation of the random-appearing model data.

Vallis has extended this mechanism to a one-dimensional continuum and integrated the resulting equations, making the chaos disappear and fluctuations be modulated (Vallis, 1988). One-dimensional models allow easy exploration of ideas. With stochastic forcing, the Vallis continuum model can produce stochastic resonance oscillations. When natural seasonal variability is introduced, ENSO appears as the occasional enlargement of an annual signal. In the spirit of Lorenz's original modeling effort, the prediction itself of chaos is interesting. Because of the paucity of long-term, frequently sampled time series for ENSO, chaos is difficult to corroborate experimentally.

A recent model has used simplified equatorial ocean and atmosphere dynamics, computed on an equatorial grid. In turn, Münnich et al. (1991) have constructed an interpretive mechanistic model of that model. Included is oceanic thermal reservoir feedback, expressed as a nonlinear relation between the depth of the eastern Pacific ocean thermocline, h_e, and the amplitude A of the zonal forcing wind stress τ. The thermocline depth is the mean depth of the thin oceanic layer where the temperature changes rapidly between the warm surface and the cold deeper region. Hence, h_e is a measure of the thermal energy of the western equatorial Pacific ocean. When $A(h_e)$ is expressed as a nonlinear function, bifurcations occur as the strength parameter of the function $A(h_e)$ increases. Successively more realistic parameterizations, starting with a simple cubic nonlinearity

$$A(h_e) = \kappa(h_e - h_e^3), \tag{27}$$

have all produced three bifurcations, the last being into a nonperiodic regime as the strength parameter κ is increased. This pattern of bifurcations suggests that a chaotic state has been reached *via* the Ruelle-Takens-Newhouse route (Schuster, 1988). For their model of ENSO, Münnich et al. conclude that stochastic forcing is not needed, various mechanisms for chaos are possible within their simplified framework, and annual forcing enhances ENSO episodes.

Keppenne and Ghil (1992) have done a PCA of a time series of the SOI. A temporal PCA is referred to as a singular spectral analysis (SSA). By keeping just the first four principal components in a reconstruction of the time series, they filter out higher-frequency oscillations. This smooths the reconstructed series, filtering some possible noise perturbations of a deterministic system. The agreement between the 50-yr time series (smoothed by taking a 5-month running mean) and the recombined first four PCs is remarkable, with El Niño and La Niña episodes identified with all the maxima and minima of greatest amplitude. Yet these four principal components only account for 21.7% of the SOI variance! When Keppenne and Ghil combine the two high-frequency components and the two low-frequency components separately, they obtain smooth oscillatory behavior amenable to autoregressive

linear prediction. The 36-month predictions agree with El Niño and La Niña events over the last ten years, and suggest a La Niña event in the winter of 1993–1994. This prediction uses no SST data at all and a minimal procedure – time-delay embedding and temporal PCA analysis – for filtering out external noise.

5.2 Chaotic Ocean Heat Transport

Climatic data show fluctuations on all time scales. In the long paleoclimatological time scales, the fluctuations are partly governed by stochastic mixing in data storaging and accessing processes (Crowley and North, 1991). However, climate also shows a significant variability associated with Earth's orbital variations on these same time scales. On decadal and interdecadal time scales, some observed fluctuations may be associated with the important poleward ocean transport of heat driven by thermohaline circulation (THC) (Covey, 1991). THC is the largely vertical oceanic motion caused by differing densities. It is an Archimedes-principle buoyancy that lifts warmer and less saline water. In the present climatic regime, THC produces an equator-to-poles flow of warm surface water and a deep cold-water return flow.

Modeling searches for a significant climatic impact of THC. Just as with ENSO, a major problem is to model the air-sea interaction. Here we discuss models where the ocean physics (momentum, heat, and salinity transport) is discretized while boundary conditions describe the atmospheric effects. Models differ on the amount of natural variability and stability. Recent numerical experiments with oceanic GCM's (OGCM's) have produced a phenomenon known as the polar halocline catastrophe. This is a blocking of the polar downwelling component of oceanic circulation, to alter dramatically oceanic salinity and heat fluxes.

A recent series of numerical OGCM experiments (Weaver et al., 1993) has parametrized the freshwater flux forcing of THC in an OGCM computed with a coarse grid and other features allowing for rapid numerical modeling. One of these features is use of a linearized numerical model for horizontal momentum transport. Subsystem time steps are allowed to differ, and they range between 2 h and 5 d. The ocean system is a flat-bottomed basin 60° wide, with latitude ranges from 0° to between 64°N and 72°N. Different runs of this OGCM employ 15 to 33 vertical levels and horizontal grids ranging from $2^\circ \times 2^\circ$ to $4^\circ \times 3.75^\circ$. This "spherical cow" is not quite complete: boundary conditions are needed, both to "wind up" the system, i.e., to start it up and put it into some climatologically reasonable steady state, and to maintain a reasonable air-sea interaction. A temporally constant wind field and windstress forcing are set up. The air-sea interaction involves only negligible transport of salts between air and sea. However, the freshwater flux, $P - E$, where P is precipitation and E is evaporation (in zonally averaged units of $\mathrm{m} \cdot \mathrm{yr}^{-1}$) clearly affects salinity (measured in grams of salts per kilogram of seawater). During the windup, the salinity and heat fluxes are subjected to restoring damping laws:

$$Q_T = C_T(T_O - T_A), \tag{28}$$

$$Q_S = C_S(S_O - S_A), \tag{29}$$

where Q_T and Q_S are heat and salinity fluxes, C_T and C_S are positive constants, while T_O and T_A are temperatures and S_O and S_A are salinities of ocean and atmosphere. The forcing temperature and

salinity fields T_A and S_A are prescribed from data.

After windups to experimental zonally averaged salinity and temperature fields, the systems are cut loose from the restoring salinity forcing, but are subjected to a temporally constant value of $P - E$, inferred from the model in its wound-up steady state. This is called mixed boundary conditions, because the thermal forcing is still maintained. Grounds for mixed boundary conditions are the absence of any physical basis for forcing through salinity damping.

Here is where chaos can set in. If the salinity boundary condition used in spin-ups equals the zonally averaged salinity field, a $P - E$ minimum occurs at 54°N, and an unstable downwelling takes place at that point, a reverse circulation cell occurring at high latitudes. An irregular flush takes place on the average every eight years, restoring to normal downwelling at the Pole. A stochastic forcing component changes the frequency and intensity of the flushes. To what extent the reported results here are artifacts is a matter of ongoing debate. Again, in the spirit of the Lorenz equations, what is important here is perhaps not this model's validation and its detailed predictions so much as the possibility of chaos or stochastic resonance it presents.

5.3 Controlling Chaos

Controlling chaos is a process of forcing a chaotic system to reach and remain in a state or a sequence of states. Control theory uses the high leverage of chaotic behavior to advantage: SIC is exploited both to rapidly target a state and to remain there. The formal insight that chaotic orbits consist of unstable periodic orbits and their limit points can be used here. A full range of technical tools such as embedding and numerical determination of the stability matrix (with or without knowledge of the underlying dynamics) has already been employed since some practical methods of controlling chaos were introduced in 1990 by Ott, Grebogi, and Yorke (Auerbach et al., 1992). Technological applications have been rapid, and to date they include controlling the following chaotic systems: vibrations of a magnetoelastic ribbon, lasers operating at ultrahigh intensities, diode circuits, and cardiac arrhythmia (Garfinkel et al., 1992). An associated message-encoding algorithm has been developed that appears robust in the presence of noise and advantageous for information transmission (Hayes et al., 1993).

5.4 Solitons

Solitons are solutions of nonlinear, dispersive wave equations with particle-like properties. They are spatially localized and do not change shape either when propagating or after colliding with other solitons. Solitons can only be solutions to equations for which the effects of dispersion and nonlinearity somehow cancel. Theoretical interest has been high since demonstration of their mathematical properties in a numerical computer experiment by Zabusky and Kruskal in 1965, although a physical soliton had been observed over a century before (in 1834) as a wave in a Scottish canal. Solitons of widely differing topological characteristics arise in many mathematical systems, including nuclear interactions, elementary particles, and plasmas. Now, engineering interest is high, too. Solitonic pulses on optical fibers may soon become the mechanism

for transoceanic communications, at rates exceeding 10^{10} bits per second (10 Gb/s) (Haus, 1993).

The history of solitons reveals interplay among different approaches to modeling. In 1895, Kortweg and deVries succeeded in deriving a shallow-water equation for small-amplitude gravity waves. Their (KdV) equation takes the form

$$u_t - 6uu_x + u_{xxx} = 0. \qquad (30)$$

Here the subscripts refer to time and space differentiation. Dissipation effects were ignored in the derivation. Analytic solutions were known in 1895, but the general stability and behavior of multiple-soliton solutions were first obtained numerically in 1965. Then, with another breakthrough, Gardner et al. in 1967 discovered powerful formal solutions by reducing the problem to inverse scattering in a linear Schrödinger-type equation. The developments of *inverse scattering theory* (IST), originally motivated by nuclear physics, saw their first major application in this soliton context. Nonlinear equations, if cast into IST form, are shown to be integrable – as is the one-dimensional nonlinear Schrödinger equation, relevant to optical fibers.

There is no general recipe for determining whether a nonlinear equation is reducible to IST (Drazin and Johnson, 1990). However, a determined effort is underway to develop better tests, and with them, appropriate reductions to IST (Ames and Rogers, 1992). One key would be to find sufficiently simple equivalent characterizations of complete integrability (Ablowitz and Segur, 1981). This can be succinctly expressed in the language of Hamiltonian dynamics. Qualitatively, completely integrable systems of differential equations when cast into the Hamiltonian form must have a sufficient number of properly constrained functionals whose Poisson brackets with each other vanish. Not all initial conditions lead to solitonic solutions. If a solitonic solution occurs, the IST generates the soliton or solitons through a process equivalent to generating action-angle variables. Such action-angle variables contain constants of the motion required to exist from complete integrability. Some of the relationships among some possible alternative approaches to determining complete integrability are discussed by Ablowitz and Segur (1981), and newer formulations and updates are given by Ames and Rogers (1992).

The IST is concerned with finding a linear Schrödinger-type equation, given the data in the form of asymptotic behavior of its solution. What remains to be determined is a potential energy operator. The integral equations for that potential, the Marchenko equations, are linear; hence, IST simplifies the solution of the original nonlinear equation. The data include scattering phase shifts (or in one dimension, reflection coefficients), and bound-state poles and residues. Solitonic solutions are found after setting the reflection coefficient equal to zero. Relevant to optical fiber design, the nonzero reflection-coefficient case is algebraically challenging, even when data are provided in terms of rational functions (Sabatier, 1988). The IST work should benefit from computer algebra. As yet, only numerical methods are used to study solitonic stability in the presence of dissipation.

Glossary

Cellular Automaton: Dynamical system temporally and spatially discrete (in cells).

Chaotic System: Deterministic dynamical system that exhibits noisy, random-appearing behavior. Technically, must show sensitivity to initial conditions (SIC).

Deterministic Dynamical System: Dynamical system having a rule that specifies its state at the next instant of time uniquely, given the state at any specific time.

Discretization: The approximation of a continuum by a discrete set of points or a discrete set of elements. These elements can be finite physical segments, or they can be sets of functions. Invariably performed when there are space-time dimensions to be modeled.

Dynamical System: Mathematical model of time evolution.

El Niño: Sporadic anomalous warming of the waters of the eastern equatorial Pacific.

Filter: General term for a dynamical system that performs a task such as tuning or removing some unwanted variance, such as noise, in a signal.

Finite Elements: A form of discretization of a continuum that treats the system as composed of small, highly symmetric, interacting parts, such as polyhedra.

Flow: Name for continuous time evolution of a dynamical system.

General Circulation Model (GCM): Parametrized simulation model of the circulation within Earth's oceans, the global atmosphere, or both combined.

Importance Sampling: A method for generating a sequence of configurations for performing Monte Carlo (MC) calculations, in which successive configurations are correlated.

Inverse Scattering Theory (IST): A formulation of scattering theory, the solution of which is a potential operator in a Schrödinger-type equation, given sufficient data.

La Niña: Sporadic anomalous cooling of the waters of the eastern equatorial Pacific.

Map (Or Iterative Map): Name for discrete-time evolution of a dynamical system.

Mechanistic Model: Mathematical model that attempts to describe the processes, or mechanisms, within a system that are relevant to the problem at hand.

Markov Process of Order *k*: Dynamical process that is accurately described in terms of transition probabilities between states, or configurations, of the system, $P(\mathbf{X} \to \mathbf{X}')$, that are functions only of the last k states. When $k = 1$, it is simply referred to as Markov process.

Monte Carlo Method: Computational method of great scope based upon the use of random-number sequences and stochastic algorithms.

Neural Network: A system possessing elements, called nodes or neurons, with large numbers of interconnections. Patterned after real neural systems and capable of learning to perform complex tasks such as pattern-matching.

Padé Approximant: Rational approximation matching a function's partial Taylor series.

Paleoclimatology: The study of climate on time scales that can be observed using historical and geological data ranging, approximately, from 10^2 to 10^8 years.

Parametrization: Introduction of a mechanistic description of an important subsystem within a simulation model.

Power Spectrum: The magnitude squared of a Fourier transform. An important

tool for distinguishing periodicity and quasiperiodicity from noise and $1/f$ phenomena.

Pseudorandom Numbers: A sequence of numbers generated by a computer algorithm, which passes a set of tests for randomness.

Rational Function: Function expressed as a ratio of two polynomials.

Self-Organized Criticality: State toward which some large systems naturally evolve, characterized by spatial and temporal self-similarity. Weakly chaotic.

Simulation Model: Model containing as many of the known relevant degrees of freedom of the system as possible. In the ideal limiting case, there is no parametrization.

Soliton: A particle-like solution to a nonlinear, dispersive wave equation. It is localized and retains its shape while propagating or upon colliding with another soliton.

Statistical Padé Approximant: Rational function that is fitted to numerical data by minimizing errors in some way.

Stochastic Dynamical System: Dynamical system having a rule that specifies its state at the next instant of time only as one of several possible states with some probability.

Synergetics: An approach to the study of large, complex systems, which seeks explanations in terms of the slaving principle, bifurcations, order parameters, and physical principles.

Time-Delay Embedding: Technique used to reconstruct a vector orbit in a nonlinear dynamical system, using time-delayed values of a scalar observable $y(t)$.

List of Works Cited

Abraham, N., Albano, A. M., Passamante, A., Rapp, P. E. (Eds.) (1990), *Measures of Complexity and Chaos*, New York: Plenum.

Ablowitz, M. J., Segur, H. (1981), *Solitons and the Inverse Scattering Transform*, Philadelphia: SIAM.

Ames, W. F., Rogers, C. (Eds.) (1992), *Nonlinear Equations in the Applied Sciences*, New York: Academic Press.

Antoulas, A. A. (Ed.) (1991), *Mathematical System Theory*, Berlin: Springer-Verlag.

Appel, K., Haken, W. (1978), in: L. A. Steen (Ed.), *Mathematics Today*, New York: Springer-Verlag.

Arfken, G. (1985), *Mathematical Methods for Physicists*, 3rd ed., Orlando: Academic Press.

Auerbach, D., Grebogi, C., Ott, E., Yorke, J. A. (1992), *Phys. Rev. Lett.* **69**, 3479–3482.

Bak, P., Tang, C., Wiesenfeld, K. (1987), *Phys. Rev. Lett.* **59**, 381–384.

Baker, G. A., Jr., Graves-Morris, P. (1981), *Padé Approximants Part I & Part II*, Reading, MA: Addison-Wesley.

Bernardet, P., Butet, A., Déqué, M., Ghil, M., Pfeffer, R. L. (1990), *J. Atmos. Sci.* **47**, 3023–3043.

Binder, K. (Ed.) (1987), *Applications of the Monte Carlo Method in Statistical Physics*, 2nd ed., Berlin: Springer-Verlag.

Brandstater, A., Swinney, H. L. (1987), *Phys. Rev. A* **35**, 2207–2220.

Butler, F., Chen, H., Sexton, J., Vaccarino, A., Weingarten, D. (1993), *Phys. Rev. Lett.* **70**, 2849–2852.

Carter, E. F. (1989), *Dynamics of Atmospheres and Oceans* **13**, 335–348.

Casti, J. L. (1992), *Reality Rules: I & II, Picturing The World In Mathematics*, New York: Wiley-Interscience.

Chui, C. K., Chen, G. (1987), *Kalman Filtering*, Berlin: Springer-Verlag.

Cook, R. D., Malkus, D. S., Plesha, M. E. (1989), *Concepts and Applications of Finite Element Analysis*, New York: Wiley.

Covey, C. (1991), *Nature* **353**, 796–797.

Creutz, M. (1992) (Ed.), *Quantum Fields on the Computer*, Singapore: World Scientific.

Crowley, T. J., North, G. R. (1991), *Paleoclimatology*, Oxford: Oxford University Press.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Philadelphia: Society for Industrial and Applied Mathematics.
Drazin, P. G., Johnson, R. S. (1990), *Solitons: An Introduction*, Cambridge, U.K.: Cambridge University Press.
Efron, B., Tibshirani, R. (1991), *Science* **253**, 390–395.
Elsner, J. B., Tsonis, A. A. (1993), *Bull. Am. Meteor. Soc.* **74**, 243.
Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.
Ferrenberg, A. M., Landau, D. P., Wong, Y. J. (1992), *Phys. Rev. Lett.* **69**, 3382–3384.
Fitch, J. (1993), *Phys. World* **6**, 48–52.
Ford, J., Mantica, G. (1992), *Am. J. Phys.* **60**, 1086–1097.
Fraser, A. M., Swinney, H. L. (1986), *Phys. Rev. A* **33**, 1134–1140.
Gallager, R. G. (1968), *Information Theory and Reliable Communication*, New York: Wiley.
Garfinkel, A., Spano, M. L., Ditto, W. L., Weiss, J. N. (1992), *Science* **257**, 1230–1235.
Gerhardt, M., Schuster, H., Tyson, J. J. (1991), *Phys. D* **50**, 189–206.
Ghil, M., Malanotte-Rizzoli, P. (1991), in: R. Dmowska, B. Saltzman (Eds.), Advances in Geophysics, Vol. 33, San Diego: Academic Press, pp. 141–266.
Gould, H., Tobochnik, J. (1988), *Computer Simulation Methods, Part I and Part II*, Reading, MA: Addison-Wesley.
Grassberger, P. (1986), *Nature* **323**, 609–612.
Hammel, S. M., Yorke, J. A., Grebogi, C. (1988), *Bull. Am. Math. Soc.* **19**, 465–469.
Hammerstrom, D. (1993), *IEEE Spectrum*, June, 26–32.
Harte, J. (1988), *Consider a Spherical Cow*, Mill Valley, CA: University Science Books.
Hassani, S. (1991), *Foundations of Mathematical Physics*, Boston: Allyn and Bacon.
Hayes, S., Grebogi, C., Ott, E. (1993), *Phys. Rev. Lett.* **70**, 3031–3034.
Hattori, T., Takesue, S. (1991), *Phys. D* 295–322.
Haus, H. A. (1993), *IEEE Spectrum*, March, 48–53.
Huynh, H. T. (1993), *SIAM J. Numer. Anal.* **30**, 57–102.
Jefferys, W. H., Berger, J. O. (1992), *Am. Sci.* **80**, 64–72.
Kalos, M. H., Whitlock, P. A. (1986), Monte Carlo Methods, Vol. 1, New York: Wiley-Interscience.
Keppenne, C. L., Ghil, M. (1992), *J. Geophys. Res.* **97**, 20449–20454.
Knuth, D. E. (1981), The Art of Computer Programming, Vol. 2, Reading, MA: Addison-Wesley.
Kumaresan, R. (1993), in: S. K. Mitra, J. F. Kaiser (Eds.), *Handbook for Digital Signal Processing*, New York: Wiley, Chap. 16.
Lakshmivarahan, S., Dhall, S. K. (1990), *Analysis and Design of Parallel Algorithms: Arithmetic and Matrix Problems*, New York: McGraw Hill.
Levi, B. G., Hafemeister, D., Scribner, R. (Eds.) (1992), *Global Warming: Physics and Facts*, New York: American Institute of Physics.
Lichtenberg, A. J., Lieberman, M. A. (1992), *Regular and Chaotic Dynamics*, 2nd ed., New York: Springer-Verlag.
Lorenz, E. N. (1963), *J. Atmos. Sci.* **20**, 130–141.
Mandelbrot, B. B. (1983), *The Fractal Geometry of Nature*, New York: Freeman.
Marchuk, G. I., Mikhailov, G. A., Nazarliev, M. A., Darbinjan, R. A., Kargin, B. A., Elepov, B. S. (1980), *The Monte Carlo Methods in Atmospheric Optics*, Berlin: Springer-Verlag.
McInnes, A. W., Marshall, T. H. (1993), *J. Approx. Theory* **75**, 85–106.
Moss, F. (1991), *Ber. Bunsenges. Phys. Chem.* **95**, 303–311.
Münnich, M., Cane, M. A., Zebiak, S. E. (1991), *J. Atmos. Sci.* **47**, 1562–1577.
O'Brien, J. J. (Ed.) (1986), *Advanced Physical Oceanographic Numerical Modeling*, Dordrecht: Reidel.
Oppenheim, A. V., Schafer, R. W. (1989), *Discrete-Time Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall.
Parlitz, U. (1992), *Int. J. Bifurcation and Chaos* **2**, 155–165.
Peixoto, J. P., Oort, A. H. (1992), *Physics of Climate*, New York: American Institute of Physics.
Philander, S. G. (1990), *El Niño, La Niña, and the Southern Oscillation*, San Diego: Academic.
Pool, R. (1992), *Science* **256**, 44–62.
Preisendorfer, R. W. (1988), *Principal Component Analysis in Meteorology and Oceanography*, Amsterdam: Elsevier.
Press, W. H., Flannery, B. P., Teukolsky, S. A., Vettering, W. T. (1987), *Numerical Recipes*, Cambridge, U.K.: Cambridge University Press.
Priestly, M. B. (1982), Spectral Analysis and Time Series, 2 vols., London: Academic Press.

Priestly, M. B. (1988), *Non-Linear and Non-Stationary Time Series Analysis*, San Diego: Academic.
Rao, N. S., Roxin, E. O. (1990), *SIAM J. Appl. Math.* **50**, 853–864.
Sabatier, P. C. (Ed.) (1988), *Some Topics on Inverse Problems*, Singapore: World Scientific.
Sabelfeld, K. K. (1991), *Monte Carlo Methods in Boundary Value Problems*, Berlin: Springer-Verlag.
Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.
Schlesinger, M. E. (1989), in A. Berger, R. E. Dickinson, J. W. Kidson (Eds.), *Understanding Climate Change*, Washington, DC: American Geophysical Union, pp. 177–187.
Schneider, S. H., Dickinson, R. E. (1974), *Rev. Geophys. Space Phys.* **12**, 447–493.
Schumaker, L. L. (1980), *Spline Functions: Basic Theory*, New York: Wiley-Interscience.
Schuster, H. G. (1988), *Deterministic Chaos*, 2nd ed., Weinheim, Germany: VCH.
Sergin, V. Ya. (1980), *Science* **209**, 1477–1482.
Sharkovsky, A. N. (1964), *Ukr. Math. Z.* **16**, 61–72.
Shertzer, J. (1989), *Phys. Rev. A* **39**, 3833–3835.
Smale, S. (1967), *Bull. Am. Math. Soc.* **73**, 747–817.
Sreenivasan, K. R. (1991), *Annu. Rev. Fluid Mech.* **23**, 539–600.
Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J. D. (1992), *Physica D* **58**, 77–94.
Tournier, E. (Ed.) (1991), *Computer Algebra and Differential Equations*, New York: Academic Press.
Tufillaro, N. B., Abbott, T., Reilly, J. (1992), *An Experimental Approach to Nonlinear Dynamics and Chaos*, Reading, MA: Addison-Wesley.
Umrigar, C. J. (1989), *Int. J. Quantum Chem. Symp.* **23**, 217–230.
Vallis, G. K. (1988), *J. Geophys. Res.* **93**, 13979–13991.
Vautard, R., Ghil, M. (1989), *Physica D* **35**, 395–424.
Vautard, R., Yiou, P., Ghil, M. (1992), *Physica D* **58**, 95–126.
Vergassola, M., Frisch, U. (1991), *Physica D* **54**, 58–64.
Vicsek, T. (1992), *Fractal Growth Phenomena*, Singapore: World Scientific.
Weaver, A. J., Marotzke, J., Cummins, P. F., Sarachik, E. S. (1993), *J. Phys. Oceanogr.* **23**, 39–60.
West, B. J., Shlesinger, M. (1990), *Am. Sci.* **78**, 40–45.
Wolfram, S. (Ed.) (1986), *Theory and Applications of Cellular Automata*, Singapore: World Scientific.
Yidana, P. V. A., Hartt, K. (1988), *J. Comp. Phys.* **78**, 481–492.
Yonezawa, F. (1993), *Science* **260**, 635–640.
Yates, F. E. (1987), *Self-Organizing Systems*, New York: Plenum Press.
Zurada, J. M. (1992), *Artificial Neural Systems*, St. Paul: West.

Further Reading

Borchers, R. R. (Ed.) (1990), *Computers in Physics – Special Issue: Modeling the Environment*, Vol. 3, (4).
Casti, J. L. (1992), *Reality Rules: I and II*, New York: Wiley.
Gould, H., Tobochnik, J. (1988), *Computer Simulation Methods, Parts 1 and 2*, Reading, MA: Addison-Wesley.
Schroeder, M. (1991), *Fractals, Chaos, Power Laws*, New York: Freeman.

Monte-Carlo Methods

K. Binder
Institut für Physik, Johannes-Gutenberg-Universität Mainz, Mainz, Germany

Introduction and Overview ***250***
1 Random-Number Generation ***251***
1.1 General Introduction *251*
1.2 Properties That a Random-Number Generator Should Have *252*
1.3 Comments about a Few Frequently Used Generators *252*
2 Simple Sampling of Probability Distributions Using Random Numbers ***253***
2.1 Numerical Estimation of Known Probability Distributions *253*
2.2 "Importance Sampling" versus "Simple Sampling" *253*
2.3 Monte Carlo as a Method of Integration *254*
2.1 Infinite Integration Space *254*
2.5 Random Selection of Lattice Sites *255*
2.6 The Self-Avoiding Walk Problem *255*
2.7 Simple Sampling versus Biased Sampling: the Example of SAWs Continued *256*
3 Survey of Applications to Simulation of Transport Processes ***257***
3.1 The "Shielding Problem" *257*
3.2 Diffusion-Limited Aggregation (DLA) *258*
4 Monte-Carlo Methods in Statistical Thermodynamics: Importance Sampling ***258***
4.1 The General Idea of the Metropolis Importance Sampling Method *258*
4.2 Comments on the Formulation of a Monte-Carlo Algorithm *259*
4.3 The Dynamic Interpretation of the Monte-Carlo Method *261*
4.4 Monte-Carlo Study of the Dynamics of Fluctuations near Equilibrium and of the Approach toward Equilibrium *262*
4.5 The Choice of Statistical Ensembles *263*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

5 Accuracy Problems: Finite-Size Problems, Dynamic Correlation of Errors, Boundary Conditions *265*
5.1 Finite-Size-Induced Rounding and Shifting of Phase Transitions *265*
5.2 Different Boundary Conditions: Simulation of Surfaces and Interfaces *266*
5.3 Estimation of Statistical Errors *267*
6 Quantum Monte-Carlo Techniques *268*
6.1 General Remarks *268*
6.2 Path-Integral Monte-Carlo Methods *268*
6.3 An Application Example: the Momentum Distribution of Fluid ^{4}He *269*
6.4 A Few qualitative Comments on Fermion Problems *270*
7 Lattice Gauge Theory *271*
7.1 Some Basic Ideas of Lattice Gauge Theory *271*
7.2 A Recent Application *272*
8 Selected Comments on Applications in Classical Statistical Mechanics of Condensed-Matter Systems *273*
8.1 Metallurgy and Materials Science *273*
8.2 Polymer Science *274*
8.3 Surface Physics *276*
9 Concluding Remarks *277*
Glossary *277*
List of Works Cited *278*
Further Reading *280*

Introduction and Overview

Many problems in science are very complex: e.g., statistical thermodynamics considers thermal properties of matter resulting from the interplay of a large number of elementary particles. A deterministic description in terms of the equation of motion of all these particles would make no sense, and a probabilistic description is required. A probabilistic description may even be intrinsically implied by the quantum-mechanical nature of the basic processes (e.g., emission of neutrons in radioactive decay) or because the problem is incompletely characterized, only some degrees of freedom being considered explicitly while the others act as a kind of background causing random noise. While thus the concept of probability distributions is ubiquitous in physics, often it is not possible to compute these probability distribution functions analytically in explicit form, because of the complexity of the problem. For example, interactions between atoms in a fluid produce strong and nontrivial correlations between atomic positions, and, hence, it is not possible to calculate these correlations analytically.

Monte-Carlo methods now aim at a numerical estimation of probability distributions (as well as of averages, that can be calculated from them), making use of (pseudo) random numbers. By "pseudo-random numbers" one means a sequence of numbers produced on a computer with a deterministic procedure from a suitable "seed." This sequence, hence, is not truly

random – see Sec. 1 for a discussion of this problem.

The outline of the present article is as follows. Since all Monte-Carlo methods heavily rely on the use of random numbers, we briefly review random-number generation in Sec. 1. In Sec. 2, we then elaborate on the discussion of "simple sampling," i.e., problems where a straightforward generation of probability distributions using random numbers is possible. Section 3 briefly mentions some applications to transport problems, such as radiation shielding, and growth phenomena, such as "diffusion-limited aggregation" (DLA).

Section 4 then considers the importance sampling methods of statistical thermodynamics, including the use of different thermodynamic ensembles. This article emphasizes applications in statistical mechanics of condensed matter, since this is the field where most activity with Monte-Carlo methods occurs.

Some more practical aspects important for the implementation of algorithms and the judgment of the tractability of simulation approaches to physical problems are then considered in Sec. 5: effects resulting from the finite size of simulation boxes, effects of choosing various boundary conditions, dynamic correlation of errors, and the application to studies of the dynamics of thermal fluctuations.

The extension to quantum-mechanical problems is mentioned in Sec. 6 and the application to elementary particle theory (lattice gauge theory) in Sec. 7. Section 8 then illustrates some of the general concepts with a variety of applications taken from condensed matter physics, while Sec. 9 contains concluding remarks.

We do not discuss problems of applied mathematics such as applications to the solution of linear operator equations (Fredholm integral equations, the Dirichlet boundary-value problem, eigenvalue problems, etc.); for a concise discussion of such problems, see, e.g., Hammersley and Handscomb (1964). Nor do we discuss simulations of chemical kinetics, such as polymerization processes (see, e.g., Bruns et al. 1981).

1 Random-Number Generation

1.1 General Introduction

The precise definition of "randomness" is a problem in itself (see, e.g., Compagner, 1991) and is outside the scope of the present article. Truly random numbers are unpredictable in advance and must be produced by an appropriate physical process such as radioactive decay. Series of such numbers have been documented but would be very inconvenient to use for Monte-Carlo simulations, and usually their total number is too limited anyhow. Thus, we do not discuss them here any further.

Pseudorandom numbers are produced in the computer by one of several simple algorithms, some of which will be discussed below, and thus are predictable, as their sequence is exactly reproducible. (This reproducibility, of course, is a desirable property, as it allows detailed checks of Monte-Carlo simulation programs.) They are thus not truly random, but they have statistical properties (nearly uniform distribution, nearly vanishing correlation coefficients, etc.) that are very similar to the statistical properties of truly random numbers. Thus, a given sequence of (pseudo)random numbers appears "random" for many practical purposes. In the

following, the prefix "pseudo" will be omitted throughout.

1.2 Properties That a Random-Number Generator Should Have

What one needs are numbers that are uniformly distributed in the interval [0,1] and that are uncorrelated. By "uncorrelated" we not only mean vanishing pair correlations for arbitrary distances along the random-number sequence, but also vanishing triplet and higher correlations. No algorithm exists that satisfies these desirable requirements fully, of course; the extent to which the remaining correlations between the generated random numbers lead to erroneous results of Monte-Carlo simulations has been a matter of long-standing concern (Knuth, 1969; James, 1990); even random-number generators that have passed all common statistical tests and have been used successfully for years may fail for a new application, in particular if it involves a new type of Monte-Carlo algorithm (see, e.g., Ferrenberg et al., 1992, for a recent example). Therefore, the testing of random-number generators is a field of research in itself (see, e.g., Marsaglia, 1985; Compagner and Hoogland, 1987).

A limitation due to the finite word length of computers is the *finite period*: Every generator begins after a long but finite period to produce exactly the same sequence again. For example, simple generators for 32-bit computers have a maximum period of $2^{30}(\approx 10^9)$ numbers only. This is not enough for recent high-quality applications! Of course, one can get around this problem (Knuth, 1969; James, 1990), but, at the same time, one likes the code representing the random-number generator to be "portable" (i.e., in a high-level programing language like FORTRAN usable for computers from different manufactures) and "efficient" (i.e., extremely fast so it does not unduly slow down the simulation program as a whole). Thus, inventing new random-number generators that are in certain respects a better compromise between these partially conflicting requirements is still of interest (e.g., Marsaglia et al., 1990).

1.3 Comments about a Few Frequently Used Generators

Best known is the linear multiplicative algorithm (Lehmer, 1951), which produces random integers X_i recursively from the formula

$$X_i = aX_{i-1} + c(\text{modulo } m), \qquad (1)$$

which means that m is added when the results otherwise were negative. For 32-bit computers, $m = 2^{31} - 1$ (the largest integer that can be used for that computer). The integer constants a, c, and X_0 (the starting value of the recursion, the so-called "seed") need to be appropriately chosen {e.g., $a = 16\,807, c = 0, X_0$ odd}. Obviously, the "randomness" of the X_i results since, after a few multiplications with a, the result would exceed m and hence is truncated, and so the leading digits of X_i are more or less random. But there are severe correlations: If d-tuples of such numbers are used to represent points in d-dimensional space having a lattice structure, they lie on a certain number of hyperplanes (Marsaglia, 1968).

Equation (1) produces random numbers between 0 and m. Converting them into real numbers and dividing by m yields random numbers in the interval [0,1], as desired.

More popular now are shift-register generators (Tausworthe, 1965; Kirkpatrick and Stoll, 1981), based on the formula

$$X_i = X_{i-p} \cdot \mathrm{XOR} \cdot X_{i-q}, \qquad (2)$$

where ·XOR· is the bitwise "exclusive or" operation, and the "lags" p, q have to be properly chosen [the popular "R250" (Kirkpatrick and Stoll, 1981) uses $p = 109$, $q = 250$ and thus needs 250 initializing integers]. "Good" generators based on Eq. (2) have fewer correlations between random numbers than those resulting from Eq. (1), and much larger period.

A third type of generators is based on Fibonacci series and also recommended in the literature (Knuth, 1969; Ahrens and Dieter, 1979; James, 1990). But a general recommendation is that every user of random numbers should not rely on their quality blindly, and should perform his own tests in the context of his application.

2 Simple Sampling of Probability Distributions Using Random Numbers

In this section, we give a few nearly trivial examples of the use of Monte-Carlo methods, which will be useful for the understanding of later sections. More material on this subject can be found in standard textbooks like Koonin (1981) and Gould and Tobochnik (1988).

2.1 Numerical Estimation of Known Probability Distributions

A known probability distribution p_i that a (discrete) state i occurs with $1 \le i \le n$, with $\Sigma_{i=1}^{n} p_i = 1$, is numerically realized using random numbers uniformly distributed in the interval from zero to unity: defining $P_i = \Sigma_{j=1}^{i} p_i$, we choose a state i if the random number ζ satisfies $P_{i-1} \le \zeta < P_i$, with $P_0 = 0$. In the limit of a large number (M) of trials, the generated distribution approximates p_i, with errors of order $1/\sqrt{M}$.

Monte-Carlo methods in statistical mechanics can be viewed as an extension of this simple concept to the probability that a point $\mathbf{X}$ in phase space occurs,

$$P_{\mathrm{eq}}(\mathbf{X}) = \left(\frac{1}{Z}\right) \exp\left\{\frac{-\mathscr{H}(\mathbf{X})}{k_{\mathrm{B}}T}\right\},$$

k_{B} being Boltzmann's constant, T the absolute temperature, and $Z = \Sigma_{\mathbf{X}} \exp\{-\mathscr{H}(\mathbf{X}) / k_{\mathrm{B}}T\}$ the partition function, although in general neither Z nor $P_{\mathrm{eq}}(\mathbf{X})$ can be written explicitly (as function of the variables of interest, such as T, particle number N, volume V, etc.). The term $\mathscr{H}(\mathbf{X})$ denotes the Hamiltonian of the (classical) system.

2.2 "Importance Sampling" versus "Simple Sampling"

The sampling of the Boltzmann probability $P_{\mathrm{eq}}(\mathbf{X})$ by Monte-Carlo methods is not completely straightforward: One must not choose the points $\mathbf{X}$ in phase space completely at random, since $P_{\mathrm{eq}}(\mathbf{X})$ is extremely sharply peaked. Thus, one needs "importance sampling" methods which generate points $\mathbf{X}$ preferably from the "important" region of space where this narrow peak occurs.

Before we treat this problem of statistical mechanics in more detail, we emphasize the more straightforward applications of "simple sampling" techniques. In the following, we list a few problems for which simple sampling is useful. Suppose one wishes to generate a configuration of a

randomly mixed crystal of a given lattice structure, e.g., a binary mixture of composition A_xB_{1-x}. Again, one uses pseudorandom numbers ζ uniformly distributed in [0,1] to choose the occupancy of lattice sites $\{j\}$: If $\zeta_i < x$, the site j is taken by an A atom, and else by a B atom. Such configurations now can be used as starting point for a numerical study of the dynamical matrix, if one is interested in the phonon spectrum of mixed crystals. One can study the distribution of sizes of "clusters" formed by neighboring A atoms if one is interested in the "site percolation problem" (Stauffer, 1985), etc.

If one is interested in simulating transport processes such as diffusion, a basic approach is the generation of simple random walks. Such random walks, resulting from addition of vectors whose orientation is random, can be generated both on lattices and in the continuum. Such simulations are desirable if one wishes to consider complicated geometries or boundary conditions of the medium where the diffusion takes place. Also, it is straightforward to include competing processes (e.g., in a reactor, diffusion of neutrons in the moderator competes with loss of neutrons due to nuclear reactions, radiation going to the outside, etc., or gain due to fission events). Actually, this problem of reactor criticality (and related problems for nuclear weapons!) was the starting point for the first large-scale applications of Monte-Carlo methods by Fermi, von Neumann, Ulam, and their coworkers (Hammersley and Handscomb, 1964).

2.3 Monte Carlo as a Method of Integration

Many Monte-Carlo computations may be viewed as attempts to estimate the value of a (multiple) integral. To give the flavor of this idea, let us discuss the one-dimensional integral

$$I = \int_0^1 f(x)dx \equiv \int_0^1 \int_0^1 g(x,y)dxdy,$$
$$\text{with } g(x,y) = \begin{cases} 0 & \text{if } f(x) < y, \\ 1 & \text{if } f(x) \geq y, \end{cases} \tag{3}$$

as an example (suppose, for simplicity, that also $0 \leq f(x) \leq 1$ for $0 \leq x \leq 1$). Then I simply is interpreted as the fraction of the unit square $0 \leq x, y \leq 1$ lying underneath the curve $y = f(x)$. Now a straightforward (though often not very efficient) Monte-Carlo estimation of Eq. (3) is the "hit or miss" method: We take n points (ζ_x, ζ_y) uniformly distributed in the unit square $0 \leq \zeta_x \leq 1, 0 \leq \zeta_y \leq 1$. Then I is estimated by

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g(\zeta_{xi}, \zeta_{yi}) = \frac{n^*}{n}, \tag{4}$$

n^* being the number of points for which $f(\zeta_{xi}) \geq \zeta_{yi}$. Thus, we count the fraction of points that lie underneath the curve $y = f(x)$. Of course, such Monte-Carlo integration methods are inferior to many other techniques of numerical integration, if the integration space is low dimensional, but the situation is worse for high-dimensional integration spaces: For any method using a regular grid of points for which the integrand needs to be evaluated, the number of points sampled along each coordinate is $M^{1/d}$ in d dimensions, which is small for any reasonable sample size M if d is very large.

2.1 Infinite Integration Space

Not always is the integration space limited to a bounded interval in space. For

example, the ϕ^4 model of field theory considers a field variable $\phi(\mathbf{x})$, where $\mathbf{x}$ is drawn from a d-dimensional space and $\phi(\mathbf{x})$ is a real variable with distribution

$$P(\phi) \propto \exp\left[-\alpha\left(-\tfrac{1}{2}\phi^2 + \tfrac{1}{4}\phi^4\right)\right]; \quad \alpha > 0. \tag{5a}$$

While $-\infty < \phi < +\infty$, the distribution $P'(y)$

$$P'(y) = \frac{\int_{-\infty}^{y} P(\phi)d\phi}{\int_{-\infty}^{+\infty} P(\phi)d\phi} \tag{5b}$$

varies in the unit interval, $0 \leq P' \leq 1$. Hence, defining $Y = Y(P')$ as the inverse function of $P'(y)$, we can choose a random number ζ uniformly distributed between zero and one, to obtain $\phi = Y(\zeta)$ distributed according to the chosen distribution $P(\phi)$. Of course, this method works not only for the example chosen in Eq. (5) but for any distribution of interest. Often it will not be possible to obtain $Y(P')$ analytically, but then one can compute numerically a table before the start of the sampling (Heermann, 1986).

2.5 Random Selection of Lattice Sites

A problem that occurs very frequently (e.g., in solid-state physics) is that one considers a large lattice (e.g., a model of a simple cubic crystal with $N = L_x \times L_y \times L_z$ sites), and one wishes to select a lattice site (n_x, n_y, n_z) at random. This is trivially done using the integer arithmetics of standard computers, converting a uniformly distributed random number $\zeta_x (0 \leq \zeta_x < 1)$ to an integer n_x with $1 \leq n_x \leq L_x$ via the statement $n_x = \mathrm{int}(\zeta_x L_x + 1)$. This is already an example where one must be careful, however, when three successive pseudorandom numbers drawn from a random-number generator (RNG) are used for this purpose: If one uses a RNG with bad statistical qualities, the frequency with which individual sites are visited may deviate distinctly from a truly random choice. In unfavorable cases, successive pseudorandom numbers are so strongly correlated that certain lattice sites would be never visited!

2.6 The Self-Avoiding Walk Problem

As an example of the straightforward use of simple sampling techniques, we now discuss the study of self-avoiding walks (SAWs) on lattices (which may be considered as a simple model for polymer chains in good solvents; see Kremer and Binder, 1988). Suppose one considers a square or simple cubic lattice with coordination number (number of nearest neighbors) z. Then, for a random walk (RW) with N steps, we would have $Z_{RW} = z^N$ configurations, but many of these random walks intersect themselves and thus would, not be self-avoiding. For SAWs, one only expects of the order of $Z_{SAW} = \text{const.} \times N^{\gamma-1} z_{\text{eff}}^N$ configurations, where $\gamma > 1$ is a characteristic exponent (which is not known exactly for $d = 3$ dimensions), and $z_{\text{eff}} \leq z - 1$ is an effective coordination number, which also is not known exactly. But it is already obvious that an exact enumeration of all configurations would be possible for rather small N only, while most questions of interest refer to the behavior for large N; e.g., one wishes to study the end-to-end distance of the SAW,

$$\langle R^2 \rangle_{\text{SAW}} = \frac{1}{Z_{\text{SAW}}} \sum_{\mathbf{X}} [\mathbf{R}(\mathbf{X})]^2, \tag{6}$$

the sum being extended over all configurations of SAWs which we denote formally as points $\mathbf{X}$ in phase space. One expects that $\langle R^2 \rangle_{\mathrm{SAW}} \propto N^{2\nu}$, where ν is another characteristic exponent. A Monte-Carlo estimation of $\langle R^2 \rangle_{\mathrm{SAW}}$ now is based on generating a sample of only $M \ll Z_{\mathrm{SAW}}$ configurations $\mathbf{X}_l$, i.e.:

$$\overline{R^2} = \frac{1}{M} \sum_{l=1}^{M} [\mathbf{R}(\mathbf{X}_l)]^2 \approx \langle R^2 \rangle_{\mathrm{SAW}}. \qquad (7)$$

If the M configurations are statistically independent, standard error analysis applies, and we expect that the relative error behaves as

$$\frac{\overline{(\delta R^2)^2}}{(\overline{R^2})^2} \approx \frac{1}{M-1} \left[\frac{\langle R^4 \rangle_{\mathrm{SAW}}}{\langle R^2 \rangle_{\mathrm{SAW}}^2} - 1 \right]. \qquad (8)$$

While the law of large numbers then implies that $\overline{R^2}$ is Gaussian distributed around $\langle R^2 \rangle_{\mathrm{SAW}}$ with a variance determined by Eq. (8), one should note that the variance does not decrease with increasing N. Statistical mechanics tells us that fluctuations decrease with increasing number N of degrees of freedom; i.e., one equilibrium configuration differs in its energy $E(\mathbf{x})$ from the average $\langle E \rangle$ only by an amount of order $1/\sqrt{N}$. This property is called "self-averaging." Obviously, such a property is not true for $\langle R^2 \rangle_{\mathrm{SAW}}$. This "lack of self-averaging" is easy to show already for ordinary random walks (Binder and Heermann, 1988).

2.7 Simple Sampling versus Biased Sampling: the Example of SAWs Continued

Apart from this problem, that the accuracy of the estimation of R^2 does not increase with the number of steps of the walk, it is also not easy to generate a large sample of configurations of SAWs for large N. Suppose we do this at each step by choosing one of $z-1$ neighbors at random (eliminating from the start immediate reversals, which would violate the SAW condition). Whenever the chosen lattice site is already taken, we also would violate the SAW condition, and the attempted walk is terminated. Now the fraction of walks that will continue successfully for N steps will only be of the order of $Z_{\mathrm{SAW}}/(z-1)^N \propto [z_{\mathrm{eff}}/(z-1)]^N N^{\gamma-1}$, which decreases to zero exponentially $\{\propto \exp(-N\mu)$ with $\mu = \ln[(z-1)/z_{\mathrm{eff}}]$ for large N$\}$; this failure of success in generating long SAWs is called the "attrition problem."

The obvious recipe, to select at each step only from among the lattice sites that do not violate the SAW restriction, does not give equal statistical weight for each configuration generated, of course, and so the average would not be the averaging that one needs in Eq. (6). One finds that this method would create a "bias" toward more compact configurations of the walk. But one can calculate the weights of configurations $w(\mathbf{X})$ that result in this so-called "inversely restricted sampling" (Rosenbluth and Rosenbluth, 1955), and in this way correct for the bias and estimate the SAW averages as

$$\overline{R^2} = \left\{ \sum_{l=1}^{M} [w(\mathbf{X}_l)]^{-1} \right\}^{-1} \times \sum_{l=1}^{M} [w(\mathbf{X}_l)]^{-1} [\mathbf{R}(\mathbf{X}_l)]^2. \qquad (9)$$

However, error analysis of this biased sampling is rather delicate (Kremer and Binder, 1988).

A popular alternative to overcome the above attrition problem is the "enrichment technique," founded on the principle "Hold fast to that which is good." Namely, whenever a walk attains a length that is a multiple of s steps without intersecting itself, n independent attempts to continue it (rather than a single attempt) are made. The numbers n, s are fixed, and, if we choose $n \approx \exp(\mu s)$, the numbers of walks of various lengths generated will be approximately equal. Enrichment has the advantage over inversely restricted sampling that all walks of a given length have equal weights, while the weights in Eq. (9) vary over many orders of magnitude for large N. But the disadvantage is, on the other hand, that the linear dimensions of the walks are highly correlated, since some of them have many steps in common! For these reasons, simple sampling and its extensions are useful only for a small fraction of problems in polymer science, and now importance sampling (Sec. 4) is much more used. But we emphasize that related problems are encountered for the sampling of "random surfaces" (this problem arises in the field theory of quantum gravity), in path-integral Monte-Carlo treatments of quantum problems, and in many other contexts.

3 Survey of Applications to Simulation of Transport Processes

The possibilities to simulate the random motions of particles are extremely widespread. Therefore, it is difficult to comment about such problems in general. Thus, we rather prefer again to proceed by briefly discussing a few examples that illustrate the spirit of the approach.

3.1 The "Shielding Problem"

A thick shield of absorbing material is exposed to γ radiation (energetic photons), of specified distribution of energy and angle of incidence. We want to know the intensity and energy distribution of the radiation that penetrates that shield.

The level of description is here that one may generate a lot of "histories" of those particles traveling through the medium. The paths of these γ particles between scattering events are straight lines, and different γ particles do not interact with each other. A particle with energy E, instantaneously at the point $\mathbf{r}$ and traveling in the direction of the unit vector $\mathbf{w}$, continues to travel in the same direction with the same energy, until a scattering event with an atom of the medium occurs. The standard assumption is that the atoms of the medium are distributed randomly in space. Then the total probability that the particle will collide with an atom while traveling a length δs of its path is $\sigma_c(E)\delta s$, $\sigma_c(E)$ being the cross section. In a region of space where $\sigma_c(E)$ is constant, the probability that a particle travels without collision a distance s is $F_c(s) = 1 - \exp[-\sigma_c(E)s]$. If a collision occurs, it may lead to absorption or scattering, and the cross sections for these types of events are assumed to be known.

A Monte-Carlo solution now simply involves the tracking of simulated particles from collision to collision, generating the distances s that the particles travel without collision from the exponential distribution quoted above. Particles leaving a collision point are sampled from the appropriate conditional probabilities as determined from the respective differential cross sections. For increasing sampling

efficiency, many obvious tricks are known. For example, one may avoid losing particles by absorption events: If the absorption probability (i.e., the conditional probability that absorption occurs given that a collision has occurred) is α, one may replace $\sigma_c(E)$ by $\sigma_c(E)(1-\alpha)$, and allow only scattering to take place with the appropriate relative probability. Special methods for the shielding problem have been extensively developed and already have been reviewed by Hammersley and Handscomb (1964).

3.2 Diffusion-Limited Aggregation (DLA)

Diffusion-limited aggregation is a model for the irreversible formation of random aggregates by diffusion of particles, which get stuck at random positions on the already formed object if they hit its surface in the course of their diffusion [sec Vicsek (1989), Meakin (1988), and Herrmann (1986, 1992) for detailed reviews of this problem and related phenomena]. Many problems (shapes of snowflakes, size distribution of asteroids, roughness of crack surfaces, etc.) can be understood as the end product of similar random dynamical and irreversible growth processes. Diffusion-limited aggregation is just one example of them. It may be simulated by iterating the following steps: From a randomly selected position on a spherical surface of radius R_m that encloses the aggregate (that has already been grown in the previous steps, its center of gravity being in the center of the sphere), a particle of unit mass is launched to start a simple random-walk trajectory. If it touches the aggregate, it sticks irreversibly on its surface. After the particle has either stuck or moved a distance R_f from the center of the aggregate such that it is unlikely that it will hit in the future, a new particle is launched. Ideally one would like to have $R_f \to \infty$, but, in practice, $R_f = 2R_m$ is sufficient. By this irreversible aggregation of particles, one forms fractal clusters. That means that the dimension d_f characterizing the relation between the mass of the grown object and its (gyration) radius R, $M \propto R^{d_f}$, is less than the dimension d of space in which the growth takes place. Again there are some tricks to make such simulations more efficient: For example, one may allow the particles to jump over larger steps when they travel in empty regions. From such studies, researchers have found that $d_f = 1.715 \pm 0.004$ for DLA in $d = 2$, while $d_f = 2.485 \pm 0.005$ in $d = 3$ (Tolman and Meakin, 1989). Such exponents as yet cannot be analytically predicted.

4 Monte-Carlo Methods in Statistical Thermodynamics: Importance Sampling

4.1 The General Idea of the Metropolis Importance Sampling Method

In the canonical ensemble, the average of an observable $A(\mathbf{X})$ takes the form

$$\langle A \rangle = \frac{1}{Z} \int_\Omega d^k X A(\mathbf{X}) \exp\left[-\frac{\mathscr{H}(\mathbf{X})}{k_\mathrm{B} T}\right], \tag{10}$$

where Z is the partition function,

$$Z = \int_\Omega d^k X \exp\left[-\frac{\mathscr{H}(\mathbf{X})}{k_\mathrm{B} T}\right], \tag{11}$$

Ω denoting the (k-dimensional) volume of phase space $\{\mathbf{X}\}$ over which is integrated, $\mathscr{H}(\mathbf{X})$ being the (classical) Hamiltonian. For this problem, a simple sampling analog to Sec. 2 would not work: The probability distribution $p(\mathbf{X}) =$

$(1/Z)\exp[-\mathscr{H}(\mathbf{X})/k_B T]$ has a very sharp peak in phase space where all extensive variables $A(\mathbf{X})$ are close to their average values $\langle A\rangle$. This peak may be approximated by a Gaussian centered at $\langle A\rangle$, with a relative halfwidth of order $1/\sqrt{N}$ only, if we consider a system of N particles. Hence, for a practically useful method, one cannot sample the phase space uniformly, but the points $\mathbf{X}_\nu$ must be chosen preferentially from the important region of phase space, i.e., the vicinity of the peak of this probability distribution. This goal is achieved by the importance sampling method (Metropolis et al., 1953): Starting from some initial configuration $\mathbf{X}_1$, one constructs a sequence of configurations $\mathbf{X}_\nu$ defined in terms of a transition probability $W(\mathbf{X}_\nu \to \mathbf{X}'_\nu)$ that rules stochastic "moves" from an old state $\mathbf{X}_\nu$ to a new state $\mathbf{X}'_\nu$, and, hence, one creates a "random walk through phase space." The idea of that method is to choose $W(\mathbf{X} \to \mathbf{X}')$ such that the probability with which a point $\mathbf{X}$ is chosen in this process converges toward the canonical probability

$$P_{\mathrm{eq}}(\mathbf{X}) = \left(\frac{1}{Z}\right)\exp\left[-\frac{\mathscr{H}(\mathbf{X})}{k_B T}\right]$$

in the limit where the number M of states $\mathbf{X}$ generated goes to infinity. A condition sufficient to ensure this convergence is the so-called principle of detailed balance,

$$P_{\mathrm{eq}}(\mathbf{X})\,W(\mathbf{X} \to \mathbf{X}') = P_{\mathrm{eq}}(\mathbf{X}')\,W(\mathbf{X}' - \mathbf{X}). \tag{12}$$

For a justification that Eq. (12) actually yields this desired convergence, we refer to Hammersley and Handscomb (1964), Binder (1976), Heermann (1986), and Kalos and Whitlock (1986). In this importance sampling technique, the average Eq. (10) then is estimated in terms of a simple arithmetic average,

$$\bar{A} = \frac{1}{M - M_0}\sum_{\nu=M_0+1}^{M} A(\mathbf{X}_\nu). \tag{13}$$

Here it is anticipated that it is advantageous to eliminate the residual influence of the initial configuration $\mathbf{X}_1$ by eliminating a large enough number M_0 of states from the average. [The judgment of what is "large enough" is often difficult; see Binder (1976) and Sec. 5.3 below.] It should also be pointed out that this Metropolis method can be used for sampling any distribution $P(\mathbf{X})$: One simply must choose a transition probability $W(\mathbf{X} \to \mathbf{X}')$ that satisfies a detailed balance condition with $P(\mathbf{X})$ rather than with $P_{\mathrm{eq}}(\mathbf{X})$.

4.2 Comments on the Formulation of a Monte-Carlo Algorithm

What is now meant in practice by the transition $\mathbf{X} \to \mathbf{X}'$? Again there is no general answer to this question; the choice of the process may depend both on the model under consideration and the purpose of the simulation. Since Eq. (12) implies that $W(\mathbf{X} - \mathbf{X}')/W(\mathbf{X}' \to \mathbf{X}) = \exp(-\delta\mathscr{H}/k_B T)$, $\delta\mathscr{H}$ being the energy change caused by the move from $\mathbf{X} \to \mathbf{X}'$, typically it is necessary to consider small changes of the state $\mathbf{X}$ only. Otherwise the absolute value of the energy change $|\delta\mathscr{H}|$ would be rather large, and then either $W(\mathbf{X} \to \mathbf{X}')$ or $W(\mathbf{X}' \to \mathbf{X})$ would be very small. Then it would be almost always forbidden to carry out that move, and the procedure would be poorly convergent. For example, in the lattice gas model at constant particle number, a transition $\mathbf{X} \to \mathbf{X}'$ may consist of moving one particle to a randomly chosen neighboring

site. In the lattice gas at constant chemical potential, one removes (or adds) just one particle at a time, which is isomorphic to single flips in the Ising model of anisotropic magnets.

Another arbitrariness concerns the order in which the particles are selected for considering a move. Often one chooses to select them in the order of their labels (in the simulation of a fluid or lattice gas at constant particle number) or to go through the lattice in a regular typewriter-type fashion (in the case of spin models, for instance). For lattice systems, it may be convenient to use sublattices (e.g., the "checkerboard algorithm," where the white and black sublattices are updated alternatively, for the sake of an efficient "vectorization" of the program; see Landau, 1992). An alternative is to choose the lattice sites (or particle numbers) randomly. The latter procedure is somewhat more time consuming, but it is a more faithful representation of a dynamic time evolution of the model described by a master equation (see below).

It is also helpful to realize that often the transition probability $W(\mathbf{X} \to \mathbf{X}')$ can be written as a product of an "attempt frequency" times an "acceptance frequency." By clever choice of the attempt frequency, it is possible sometimes to attempt large moves and still have a high acceptance, and thus make the computations more efficient.

For spin models on lattices, such as Ising or Potts models, XY and Heisenberg ferromagnets, etc., algorithms have been devised where one does not update single spins in the move $\mathbf{X} \to \mathbf{X}'$, but, rather, one updates specially constructed clusters of spins (see Swendsen et al., 1992, for a review). These algorithms have the merit that they reduce critical slowing down, which hampers the efficiency of Monte-Carlo simulations near second-order phase transitions. "Critical slowing down" means a dramatic increase of relaxation times at the critical point of second-order phase transitions, and these relaxation times also control statistical errors in Monte-Carlo simulations, as we shall see in Sec. 5. Since these "cluster algorithms" work for rather special models only, they will not be discussed further here. But further development of such algorithms is an important area of current research (e.g., Barkema and Marko, 1993).

There is also some arbitrariness in the choice of the transition probability $W(\mathbf{X} \to \mathbf{X}')$ itself. The original choice of Metropolis et al. (1953) is

$$W(\mathbf{X} \to \mathbf{X}') = \begin{cases} \exp\left(\dfrac{-\delta\mathscr{H}}{k_{\mathrm{B}}T}\right) & \text{if } \delta\mathscr{H} > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

An alternative choice is the so-called "heat-bath method." There one assigns the new value α_i' of the ith local degree of freedom in the move from $\mathbf{X}$ to $\mathbf{X}'$ irrespective of what the old value α_i' was. One thereby considers the local energy $\mathscr{H}_i(\alpha_i')$ and chooses the state α_i' with probability

$$\frac{\exp[-\mathscr{H}_i(\alpha_i')/k_{\mathrm{B}}T]}{\sum_{\{\alpha_i''\}} \exp[-\mathscr{H}_i(\alpha_i'')/k_{\mathrm{B}}T]}.$$

We now outline the realization of the sequence of states $\mathbf{X}$ with chosen transition probability W. At each step of the procedure, one performs a *trial move* $\alpha_i \to \alpha_i'$, computes $W(\mathbf{X} \to \mathbf{X}')$ for this trial move, and compares it with a random number η, uniformly distributed in the interval $0 < \eta < 1$. If $W < \eta$, the trial move is rejected, and the old state (with

α_i) is counted once more in the average, Eq. (13). Then another trial is made. If $W > \eta$, on the other hand, the trial move is accepted, and the new configuration thus generated is taken into account in the average, Eq. (13). It serves then also as a starting point of the next step.

Since subsequent states $\mathbf{X}_\nu$ in this Markov chain differ by the coordinate α_i of one particle only (if they differ at all), they are highly correlated. Therefore, it is not straightforward to estimate the error of the average, Eq. (13). Let us assume for the moment that, after n steps, these correlations have died out. Then we may estimate the statistical error δA of the estimate $\bar{A}$ from the standard formula,

$$\overline{(\delta A)^2} = \frac{1}{m(m-1)} \sum_{\mu=\mu_0}^{m+\mu_0-1} [A(\mathbf{X}_\mu) - \bar{A}]^2, \quad m \gg 1, \tag{15}$$

where the integers μ_0, μ, m are defined by $m = (M - M_0)/n$, μ_0 labels the state $\nu = M_0 + 1$, $\mu = \mu_0 + 1$ labels the state $\nu = M_0 + 1 + n$, etc. Then also $\bar{A}$ for consistency should be taken as

$$\bar{A} = \frac{1}{m} \sum_{\mu=\mu_0}^{m+\mu_0-1} A(\mathbf{X}_\mu). \tag{16}$$

If the computational effort of carrying out the "measurement" of $A(\mathbf{X}_\mu)$ in the simulation is rather small, it is advantageous to keep taking measurements every Monte-Carlo step per degree of freedom but to construct block averages over n successive measurements, varying n until uncorrelated block averages are obtained.

4.3 The Dynamic Interpretation of the Monte-Carlo Method

It is not always easy to estimate the appropriate number of configurations M_0 after which the correlations to the initial state $\mathbf{X}_1$, which typically is a state far from equilibrium, have died out, nor is it easy to estimate the number n between steps after which correlations in equilibrium have died out. A formal answer to this problem, in terms of relaxation times of the associated master equation describing the Monte-Carlo process, is discussed in the next section. This interpretation of Monte-Carlo sampling in terms of master equations is also the basis for Monte-Carlo studies of the dynamics of fluctuations near thermal equilibrium, and is discussed now. One introduces the probability $P(\mathbf{X}, t)$ that a state $\mathbf{X}$ occurs at time t. This probability then decreases by all moves $\mathbf{X} \to \mathbf{X}'$, where the system reaches a neighboring state $\mathbf{X}'$; on the other hand, inverse processes $\mathbf{X}' \to \mathbf{X}$ lead to a gain of probability. Thus, one can write down a rate equation, similar to chemical kinetics, considering the balance of all gain and loss processes:

$$\frac{d}{dt} P(\mathbf{X}, t) = -\sum_{\mathbf{X}'} W(\mathbf{X} \to \mathbf{X}') P(\mathbf{X}, t) + \sum_{\mathbf{X}'} W(\mathbf{X}' \to \mathbf{X}) P(\mathbf{X}', t). \tag{17}$$

The Monte-Carlo sampling (i.e., the sequence of generated states $\mathbf{X}_1 \to \mathbf{X}_2 \to \ldots \to \mathbf{X}_\nu \to \ldots$) can hence be interpreted as a numerical realization of the master equation, Eq. (17), and then a "time" t is associated with the index ν of subsequent configurations. In a system with N particles, we may normalize the "time unit" such that N single-particle moves are attempted per unit time. This is often called a "sweep" or "1 Monte-Carlo step (MCS)."

For the thermal equilibrium distribution $P_{\mathrm{eq}}(\mathbf{X})$, because of the detailed balance principle, Eq. (12); there is no change of probability with time, $dP(\mathbf{X}, t)/dt = 0$;

thus, thermal equilibrium arises as the stationary solution of the master equation, Eq. (17). Thus, it is also plausible that Markov processes described by Eq. (17) describe a relaxation that always leads toward thermal equilibrium, as desired.

Now, for a physical system (whose trajectory in phase space, according to classical statistical mechanics, follows from Newton's laws of motion), it is clear that the stochastic trajectory through phase space that is described by Eq. (17) in general has nothing to do with the actual dynamics. For example, Eq. (17) never describes any propagating waves (such as spin waves in a magnet, or sound waves in a crystal or fluid, etc.).

In spite of this observation, the dynamics of the Monte-Carlo "trajectory" described by Eq. (17) sometimes does have physical significance. In many situations, one does not wish to consider the full set of dynamical variables of the system, but rather a subset only: For instance, in an interstitial alloy where one is interested in modeling the diffusion processes, the diffusion of the interstitials may be modeled by a stochastic hopping between the available lattice sites. Since the mean time between two successive jumps is orders of magnitude larger than the time scale of atomic vibrations in the solid, the phonons can be reasonably well approximated as a heat bath, as far as the diffusion is concerned.

There are many examples where such a separation of time scales for different degrees of freedom occurs: For example, for a description of the Brownian motion of polymer chains in polymer melts, the fast bond-angle and bond-length vibrations may be treated as heat bath, etc. As a rule of thumb, any very slow relaxation phenomena (kinetics of nucleation, decay of remanent magnetization in spin glasses, growth of ordered domains in adsorbed monolayers at surfaces, etc.) can be modeled by Monte-Carlo methods. Of course, one must pay attention to building in relevant conservation laws into the model properly (e.g., in an interstitial alloy, the overall concentration of interstitials is conserved; in a spin glass, the magnetization is not conserved) and to choosing microscopically reasonable elementary steps representing the move $\mathbf{X} \to \mathbf{X}'$. The great flexibility of the Monte-Carlo method, where one can choose the level of the modeling appropriately for the model at hand and identify the degrees of freedom that one wishes to consider, as well as the type and nature of transitions between them, is a great advantage and thus allows complementary applications to more atomistically realistic simulation approaches such as the molecular dynamics (*q.v.*) method where one numerically integrates Newton's equations of motion (Heermann, 1986; Ciccotti and Hoover, 1986; Hockney and Eastwood, 1988). By a clever combination with cluster-flip algorithms, one sometimes can construct very efficient algorithms and hence span a very broad range of time scales (Barkema and Marko, 1993).

4.4 Monte-Carlo Study of the Dynamics of Fluctuations near Equilibrium and of the Approach toward Equilibrium

Accepting Eq. (17), the average in Eq. (13) then is simply interpreted as a time average along the stochastic trajectory in phase space,

$$\bar{A} = \frac{1}{t_M - t_{M_0}} \int_{t_{M_0}}^{t_M} A(t)\, dt,$$
$$t_M = \frac{M}{N}, \quad t_{M_0} = \frac{M_0}{N}. \tag{18}$$

It is thus no surprise that, for the importance-sampling Monte-Carlo method, one needs to consider carefully the problem of ergodicity: Time averages need not agree with ensemble averages. For example, near first-order phase transitions there may be long-lived metastable states. Sometimes the considered moves do not allow one to reach all configurations (e.g., in dynamic Monte-Carlo methods for self-avoiding walks; see Kremer and Binder, 1988).

One can also define time-displaced correlation functions: $\langle A(t)B(0)\rangle$, where A, B stand symbolically for any physical observables, is estimated by

$$\overline{A(t)B(0)} = \frac{1}{t_M - t - t_{M_0}} \times \int_{t_{M_0}}^{t_M - t} A(t+t')B(t')\,dt', \quad t_M - t > t_{M_0}. \tag{19}$$

Equation (19) refers to a situation where t_{M_0} is chosen large enough such that the system has relaxed toward equilibrium during the time t_{M_0}; then the pair correlation depends on the difference t and not the two individual times t', $t'+t$ separately.

However, it is also interesting to study the nonequilibrium relaxation process by which equilibrium is approached. In this region, $A(t) - \bar{A}$ is systematically dependent on the observation time t, and an ensemble average $\langle A(t)\rangle_T - \langle A(\infty)\rangle_T$ [$\lim_{t\to\infty} \bar{A} = \langle A\rangle_T \equiv \langle A(\infty)\rangle_T$ if the system is ergodic] is nonzero. One may define

$$\langle A(t)\rangle_T = \sum_{\{\mathbf{X}\}} P(\mathbf{X}, t)A(\mathbf{X}) = \sum_{\{\mathbf{X}\}} P(\mathbf{X}, 0)A(\mathbf{X}(t)). \tag{20}$$

In the second step of this equation, the fact was used that the ensemble average involved is actually an average weighted by $P(\mathbf{X},0)$ over an ensemble of initial states $\mathbf{X}(t=0)$, which then evolve as described by the master equation, Eq. (17). In practice, Eq. (20) means an average over a large number $n_{\text{run}} \gg 1$ statistically independent runs,

$$[\bar{A}(t)]_{\text{av}} = \frac{1}{n_{\text{run}}} \sum_{l=1}^{n_{\text{run}}} A(t, l), \tag{21}$$

where $A(t,l)$ is the observable A observed at time t in the lth run of this nonequilibrium Monte-Carlo averaging.

Many concepts of nonequilibrium statistical mechanics can immediately be introduced in such simulations. For instance, one can introduce arbitrary "fields" that can be switched off to study the dynamic response functions, for both linear and nonlinear response (Binder, 1984).

4.5 The Choice of Statistical Ensembles

While so far the discussion has been (implicitly) restricted to the case of the canonical ensemble (NVT ensemble, for the case of a fluid), it is sometimes useful to use other statistical ensembles. Particularly useful is the grand canonical ensemble (μVT), where the chemical potential μ rather than the particle number N is fixed. In addition to moves where the configuration of particles in the box relaxes, one has moves where one attempts to add or remove a particle from the box.

In the case of binary (AB) mixtures, a useful variation is the "semi-grand canonical" ensemble, where $\Delta_\mu = \mu_A - \mu_B$ is held fixed and moves where an A particle is converted into a B particle (or

vice versa) are considered, in an otherwise identical system configuration.

The isothermal-isobaric (NpT) ensemble, on the other hand, fixes the pressure, and then volume changes $V \to V' = V + \Delta V$ need to be considered (rescaling properly the positions of the particles).

It also is possible to define artificial ensembles that are not in the textbooks on statistical mechanics. An example is

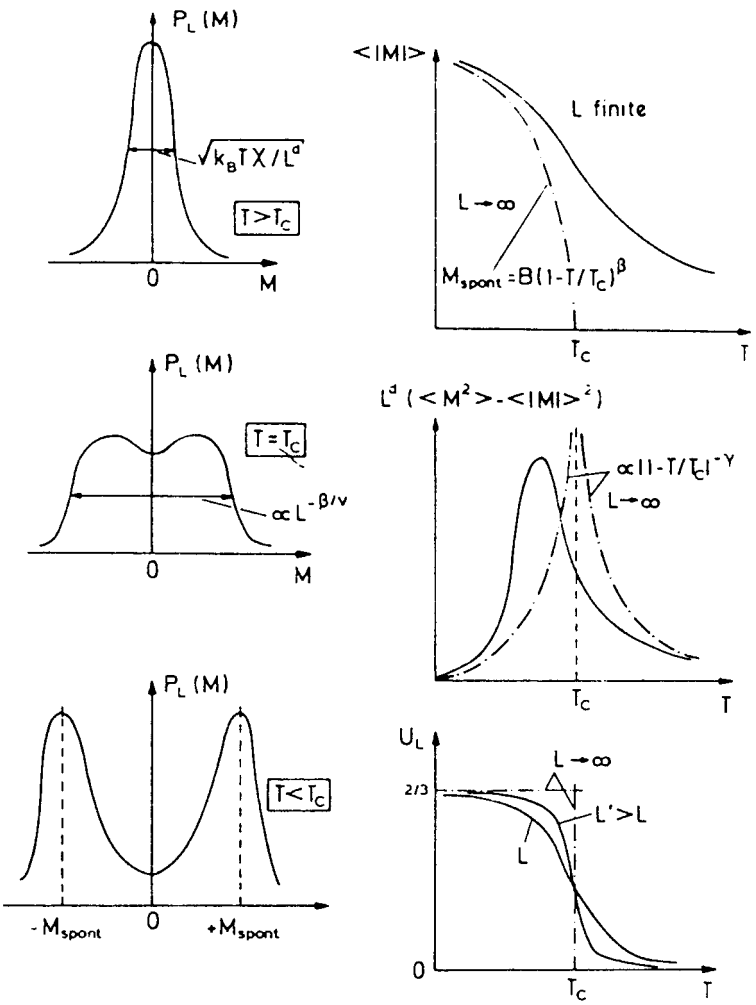


Fig. 1 Schematic evolution of the order-parameter probability distribution $P_L(M)$ from $T > T_c$ to $T < T_c$ (from above to below, left part), for an Ising ferromagnet (where M is the magnetization) in a box of volume $V = L^d$. The right part shows the corresponding temperature dependence of the mean order parameter $\langle |M| \rangle$, the susceptibility $k_B T \chi' = L^d(\langle M^2 \rangle - \langle |M| \rangle^2)$, and the reduced fourth-order cumulant $U_L = 1 - \langle M^4 \rangle / [3 \langle M^2 \rangle^2]$. Dash-dotted curves indicate the singular variation that results in the thermodynamic limit, $L \to \infty$

the so-called Gaussian ensemble (which interpolates between the canonical and microcanonical ensembles and is useful for the study of first-order phase transitions, as the so-called "multicanonical ensemble"). Particularly useful is the "Gibbs ensemble," where one considers the equilibrium between two simulation boxes (one containing liquid, the other gas), which can exchange both volume ΔV and particles (ΔN), while the total volume and total particle number contained in the two boxes are held fixed. The Gibbs ensemble is widely used for the simulation of gas-fluid coexistence, avoiding interfaces (Panagiotopoulos, 1992; Levesque and Weis, 1992).

A simulation at a given state point (NVT) contains information not only on averages at that state point but also on neighboring state points (NVT'), via suitable reweighting of the energy distribution $P_N(E)$ with a factor $\exp(E/k_B T)\exp(-E/k_B T')$. Such "histogram methods" are particularly useful near critical points (Swendsen et al., 1992).

5 Accuracy Problems: Finite-Size Problems, Dynamic Correlation of Errors, Boundary Conditions

5.1 Finite-Size-Induced Rounding and Shifting of Phase Transitions

A prominent application of Monte-Carlo simulation in statistical thermodynamics and lattice theory is the study of phase transitions. Now it is well known in statistical physics that sharp phase transitions can occur in the thermodynamic limit only, $N \to \infty$. Of course, this is no practical problem in everyday life – even a small water droplet freezing into a snowflake contains about $N = 10^{18}$ H_2O molecules, and, thus, the rounding and shifting of the freezing are on a relative scale of $1/\sqrt{N} = 10^{-9}$ and thus completely negligible. But the situation is different for simulations, which often consider extremely small systems (e.g., a hypercubic d-dimensional box with linear dimensions L, $V = L^d$, and periodic boundary conditions), where only $N \sim 10^2$ to 10^4 particles are involved.

In such small systems, phase transitions are strongly rounded and shifted (Barber, 1983; Binder, 1987, 1992a; Privman, 1990). Thus, care needs to be applied when simulated systems indicate phase changes. It turns out, however, that these finite-size effects can be used as a valuable tool to infer properties of the infinite system from the finite-size behavior. As a typical example, we discuss the phase transition of an Ising ferromagnet (Fig. 1), which has a second-order phase transition at a critical temperature T_c. For $L \to \infty$, the spontaneous magnetization M_{spont} vanishes according to a power law, $M_{\text{spont}} = B(1 - T/T_c)^{\beta}$, B being a critical amplitude and β a critical exponent (Stanley, 1971), and the susceptibility χ and correlation length ξ diverge,

$$\chi \propto \left|1 - \frac{T}{T_c}\right|^{-\gamma}, \quad \xi \propto \left|1 - \frac{T}{T_c}\right|^{-\nu}, \tag{22}$$

where γ, ν are the associated critical exponents. In a finite system, ξ cannot exceed L, and, hence, these singularities are smeared out.

Now finite-size scaling theory (Barber, 1983; Privman, 1990) implies that basically these finite-size effects can be understood from the principle that "L scales with ξ"; i.e., the order-parameter probability distribution $P_L(M)$ can be written (Binder,

1987, 1992a)

$$P_L = L^{\beta/\nu}\tilde{P}\left(\frac{L}{\xi}, ML^{\beta/\nu}\right), \qquad (23)$$

where $\tilde{P}$ is a "scaling function." From Eq. (23), one immediately obtains the finite-size scaling relations for order parameter $\langle|M|\rangle$ and the susceptibility (defined from a fluctuation relation) by taking the moments of the distribution P_L:

$$\langle|M|\rangle = L^{-\beta/\nu}\tilde{M}\left(\frac{L}{\xi}\right), \qquad (24)$$

$$k_B T'_\chi = L^d(\langle M^2\rangle - \langle|M|\rangle^2) = L^{\gamma/\nu}\tilde{X}\left(\frac{L}{\xi}\right), \qquad (25)$$

where $\tilde{M}$, $\tilde{\chi}$ are scaling functions that follow from $\tilde{P}$ in Eq. (23). At T_c where $\xi \to \infty$, we thus have $\chi' \propto L^{\gamma/\nu}$; from the variation of the peak height of χ' with system size, hence, the exponent γ/ν can be extracted.

The fourth-order cumulant U_L is a function of L/ξ only,

$$U_L \equiv 1 - \frac{\langle M^4\rangle}{3\langle M^2\rangle^2} = \tilde{U}\left(\frac{L}{\xi}\right). \qquad (26)$$

Here $U_L \to 0$ for a Gaussian centered at $\mathrm{M} = 0$, i.e., for $T > T_c$; $U_L \to \frac{2}{3}$ for the double-Gaussian distribution, i.e., for $T < T_c$; while $U_L = \tilde{U}(0)$ is a universal nontrivial constant for $T = T_c$. Cumulants for different system sizes hence intersect at T_c, and this can be used to locate T_c precisely (Binder, 1987, 1992a).

A simple discussion of finite-size effects at first-order transitions is similarly possible. There one describes the various phases that coexist at the first-order transition in terms of Gaussians if $L \gg \xi$ (note that ξ stays finite at the first-order transition). In a finite system, these phases can coexist not only right at the transition but over a finite parameter region. The weights of the respective peaks are given in terms of the free-energy difference of the various phases. From this phenomenological description, energy and order-parameter distributions and their moments can be worked out. Of course, this description applies only for long enough runs where the system jumps from one phase to the other many times, while for short runs where the systems stay in a single phase, one would observe hysteresis.

5.2 Different Boundary Conditions: Simulation of Surfaces and Interfaces

We now briefly mention the effect of various boundary conditions. Typically one uses periodic boundary conditions to study bulk properties of systems not obscured by surface effects. However, it also is possible to choose different boundary conditions to study surface effects deliberately; e.g., one may simulate thin films in a $L \times L \times D$ geometry with two free $L \times L$ surfaces and periodic boundary conditions otherwise. If the film thickness D is large enough, the two surfaces do not influence each other, and one can infer the properties of a semi-infinite system. One may choose special interactions near the free surfaces, apply surface "fields" (even if they cannot be applied in the laboratory, it may nevertheless be useful to study the response to them in the simulation), etc.

Sometimes the boundary conditions may stabilize interfaces in the system (e.g., in an Ising model for $T < T_c$ a domain wall between phases with opposite magnetization will be present, if we apply strong enough surface fields of opposite sign). Such interfaces also are often the object

of study in simulation. It may be desirable to simulate interfaces without having the systems disturbed by free surfaces. In an Ising system, this may simply be done by choosing antiperiodic boundary conditions. Combining antiperiodic and staggered periodic boundary conditions, even tilted interfaces may be stabilized in the system. In all such simulations of systems containing interfaces one must keep in mind, however, that because of capillary-wave excitations, interfaces usually are very slowly relaxing objects, and often a major effort in computing time is needed to equilibrate them. A further difficulty (when one is interested in interfacial profiles) is the fact that the center of the interface is typically delocalized.

5.3 Estimation of Statistical Errors

We now return to the problem of judging the time needed for having reasonably small errors in Monte-Carlo sampling. If the subsequent configurations used were uncorrelated, we simply could use Eq. (15), but in the case of correlations we have rather

$$\langle(\delta A)^2\rangle = \left\langle\left[\frac{1}{n}\sum_{\mu=1}^{n} A_\mu - \langle A\rangle\right]^2\right\rangle = \frac{1}{n}\left[\langle A^2\rangle - \langle A\rangle^2 + 2\sum_{\mu=1}^{n}\left(1-\frac{\mu}{n}\right) \times (\langle A_0 A_\mu\rangle - \langle A\rangle^2)\right]. \quad (27)$$

Now we remember that a time $t_\mu = \mu\delta t$ is associated with the Monte-Carlo process, δt being the time interval between two successive observations A_μ, $A_{\mu+1}$. Transforming the summation to a time integration yields

$$\langle(\delta A)^2\rangle = \frac{1}{n}(\langle A^2\rangle - \langle A\rangle^2) \times \left[1 + \frac{2}{\delta t}\int_0^{t_n}\left(1-\frac{t}{t_n}\right)\phi_A(t)\,dt\right], \quad (28)$$

where

$$\phi_A(t) \equiv \frac{\langle A(0)A(t)\rangle - \langle A\rangle^2}{\langle A^2\rangle - \langle A\rangle^2}.$$

Defining a relaxation time $\tau_A = \int_0^\infty dt\, \phi_A(t)$, one obtains for $\tau_A \ll n\delta t = \tau_{\text{obs}}$ (the observation time)

$$\langle(\delta A)^2\rangle = \frac{1}{n}[\langle A^2\rangle - \langle A\rangle^2]\left(1 + \frac{2\tau_A}{\delta t}\right) \approx 2\left(\frac{\tau_A}{\tau_{\text{obs}}}\right)[\langle A^2\rangle - \langle A\rangle^2]. \quad (29)$$

In comparison with Eq. (15), the dynamic correlations inherent in a Monte-Carlo sampling as described by the master equation, Eq. (17), lead to an enhancement of the expected statistical error $\langle(\delta A)^2\rangle$ by a "dynamic factor" $1 + 2\tau_A/\delta t$ (sometimes also called the "statistical inefficiency").

This dynamic factor is particularly cumbersome near second-order phase transitions (τ_A diverges: critical slowing down) and near first-order phase transitions (τ_A diverges at phase coexistence, because of the large life-time of metastable states). Thus, even if one is interested only in static quantities in a Monte-Carlo simulation, understanding the dynamics may be useful for estimating errors. Also the question of how many configurations (M_0) must be omitted at the start of the averaging for the sake of equilibrium [Eq. (18)] can be formally answered in terms of a nonlinear relaxation function

$$\phi^{(nl)}(t) = \frac{\langle A(t)\rangle_T - \langle A(\infty)\rangle_T}{\langle A(0)\rangle_T - \langle A(\infty)\rangle_T}$$

and its associated time $\tau_A^{(nl)} = \int_0^\infty \phi_A^{(nl)}(t)\, dt$ by the condition $t_{M_0} \gg \tau_A^{(nl)}$.

6 Quantum Monte-Carlo Techniques

6.1 General Remarks

Development of Monte-Carlo techniques to study ground-state and finite-temperature properties of interacting quantum many-body systems is an active area of research (for reviews see Ceperley and Kalos, 1979; Schmidt and Kalos, 1984; Kalos, 1984; Berne and Thirumalai, 1986; Suzuki, 1986; Schmidt and Ceperley, 1992; De Raedt and von der Linden, 1992). These methods are of interest for problems such as the structure of nuclei (Carlson, 1988) and elementary particles (De Grand, 1992), superfluidity of ^{3}He and ^{4}He (Schmidt and Ceperley, 1992), high-T_c superconductivity (e.g., Frick et al., 1990), magnetism (e.g., Reger and Young, 1988), surface physics (Marx et al., 1993), etc. Despite this widespread interest, much of this research has the character of "work in progress" and hence cannot feature more prominently in the present article. Besides, there is not just one quantum Monte-Carlo method, but many variants exist: variational Monte Carlo (VMC), Green's-function Monte Carlo (GFMC), projector Monte Carlo (PMC), path-integral Monte Carlo (PIMC), grand canonical quantum Monte Carlo (GCMC), world-line quantum Monte Carlo (WLQMC), etc. Some of these (like VMC, GFMC) address ground-state properties, others (like PIMC) finite temperatures. Here only the PIMC technique will be briefly sketched, following Gillan and Christodoulos (1993).

6.2 Path-Integral Monte-Carlo Methods

We wish to calculate thermal averages for a quantum system and thus rewrite Eqs. (10) and (11) appropriately,

$$\langle A \rangle = \frac{1}{\mathrm{Z}} \mathrm{Tr}\ \exp\left(-\frac{\hat{\mathscr{H}}}{k_\mathrm{B} T}\right) \hat{A},$$

$$\mathrm{Z} = \mathrm{Tr}\ \exp\left(-\frac{\hat{\mathscr{H}}}{k_\mathrm{B} T}\right), \qquad (30)$$

using a notation that emphasizes the operator character of the Hamiltonian $\hat{\mathscr{H}}$ and of the quantity $\hat{A}$ associated with the variable A that we consider. For simplicity, we consider first a system of a single particle in one dimension acted on by a potential $V(x)$. Its Hamiltonian is

$$\hat{\mathscr{H}} = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + V(x). \qquad (31)$$

Expressing the trace in the position representation, the partition function becomes

$$\mathrm{Z} = \int dx \left\langle x \left| \exp\left(\frac{-\hat{\mathscr{H}}}{k_\mathrm{B} T}\right) \right| x \right\rangle, \qquad (32)$$

where $|x\rangle$ is an eigenvector of the position operator. Writing $\exp(-\hat{\mathscr{H}}/k_\mathrm{B}T)$ formally as $[\exp(-\hat{\mathscr{H}}/k_\mathrm{B}TP)]^P$, where P is a positive integer, we can insert a complete set of states between the factors:

$$\mathrm{Z} = \int dx_1 \ldots \int dx_P \left\langle x_1 \left| \exp\left(\frac{-\hat{\mathscr{H}}}{k_\mathrm{B} TP}\right) \right| x_2 \right\rangle$$
$$\langle x_2 | \ldots | x_P \rangle \left\langle x_P \left| \exp\left(\frac{-\hat{\mathscr{H}}}{k_\mathrm{B} TP}\right) \right| x_1 \right\rangle. \qquad (33)$$

For large P, it is a good approximation to ignore the fact that kinetic and potential

energy do not commute. Hence, one gets

$$\left\langle x \left| \exp\left(\frac{-\hat{\mathscr{H}}}{k_B TP}\right) \right| x' \right\rangle \approx \left(\frac{k_B TmP}{2\pi\hbar^2}\right)^{1/2} \times \exp\left[\frac{-k_B TmP}{2\pi\hbar^2}(x-x')^2\right] \times \exp\left\{\frac{-1}{2k_B TP}[V(x)+V(x')]\right\}, \quad (34)$$

and

$$Z \approx \left(\frac{k_B TmP}{2\pi\hbar^2}\right)^{P/2} \times \int dx_1 \ldots \int dx_P \exp\left\{-\frac{1}{k_B T} \times \frac{1}{2}\sum_{s=1}^{P} \kappa (x_s - x_{s+1})^2 + P^{-1}\sum_{s=1}^{P} V(x_s)\right\}, \quad (35)$$

where

$$\kappa \equiv \left(\frac{k_B T}{\hbar}\right)^2 mP. \quad (36)$$

In the limit $P \to \infty$, Eq. (35) becomes exact. Apart from the prefactor, Eq. (35) is precisely the configurational partition function of a classical system of a ring polymer, consisting of P beads coupled by harmonic springs with spring constant κ. Each bead is under the action of a potential $V(x)/P$.

This approach is straightforwardly generated to a system of N interacting quantum particles – one ends up with a system of N classical cyclic "polymer" chains. As a result of this isomorphism, the Monte-Carlo method developed for simulating classical systems can be carried over to such quantum-mechanical problems, too. It is also easy to see that the system always behaves classically at high temperatures – κ gets very large, and then the cyclic chains contract essentially to a point, while at low temperatures they are spread out, representing zero-point motion. However, PIMC becomes increasingly difficult at low temperatures, since P has to be the larger the lower T: If σ is a characteristic distance over which the potential $V(x)$ changes, one must have $\hbar^2/m\sigma^2 \ll k_B TP$ in order that two neighbors along the "polymer chain" are at a distance much smaller than σ. In PIMC simulations, one empirically determines and uses that P beyond which the thermodynamic properties do not effectively change.

This approach can be generalized immediately to the density matrix $\rho(x-x') = \langle x| \exp(-\hat{\mathscr{H}}/k_B T)|x'\rangle$, while there are problems with the calculation of time-displaced correlation functions $\langle A(t)B(0)\rangle$, where t is now the true time (associated with the time evolution of states following from the Schrödinger equation, rather than the "time" of Sec. 4.3 related to the master equation).

The step leading to Eq. (34) can be viewed as a special case of the Suzuki–Trotter formula (Suzuki, 1986)

$$\exp(\hat{A}+\hat{B}) = \lim_{P\to\infty}\left[\exp\left(\frac{\hat{A}}{P}\right)\exp\left(\frac{\hat{B}}{P}\right)\right]^P, \quad (37)$$

which is also used for mapping d-dimensional quantum problems on lattices to equivalent classical problems (in $d+1$ dimensions, because of the additional "Trotter direction" corresponding to the imaginary time direction of the path-integral).

6.3 An Application Example: the Momentum Distribution of Fluid ^{4}He

We now consider the dynamic structure factor $S(\mathbf{k}, \omega)$, which is the Fourier transform of a time-displaced pair correlation

function of the density at a point $\mathbf{r}_1$ at time t_1 and the density at point $\mathbf{r}_2$ at time t_2 [$\hbar\mathbf{k}$ being the momentum transfer and $\hbar\omega$ the energy transfer of an inelastic scattering experiment by which one can measure $S(\mathbf{k}, \omega)$]. In the "impulse approximation," the dynamic structure factor $S(\mathbf{k}, \omega)$ can be related to the Fourier transform of the single-particle density matrix $\rho_1(\mathbf{r})$, which for ^{4}He can be written in terms of the wave function $\psi(\mathbf{r})$ as $\rho_1(\mathbf{r}) = \langle\psi^+(\mathbf{r}' + \mathbf{r})\psi(\mathbf{r})\rangle$. This relation is

$$S(k, \omega) \propto J(Y) = \frac{1}{\pi}\int_0^\infty \rho_1(r)\cos(Yr)\,dr,$$

where $Y \equiv m(\omega - k^2/2m)/k$ (West, 1975). Since $\mathbf{S}(k,\omega)$ has been measured via neutron scattering (Sokol et al., 1989), a comparison between experiment and simulation can be performed without adjustable parameters (Fig. 2). Thus, the PIMC method yields accurate data in good agreement with experiment.

The studies of ^{4}He have also yielded qualitative evidence for superfluidity (Ceperley and Pollock, 1987). For a quantitative analysis of the λ transition, a careful assessment of finite-size effects (Fig. 1) is needed since one works with very small particle numbers (of the order of 10^2 ^{4}He atoms only). This has not been possible so far.

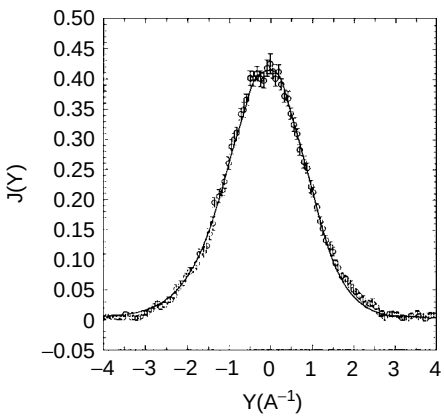


Fig. 2 The measured momentum distribution $J(Y)$ of ^{4}He at $T = 3.3$ K (circles, from Sokol et al., 1989) compared with the PIMC result of Ceperley and Pollock (1987) (solid line). From Schmidt and Ceperley (1992)

6.4 A Few qualitative Comments on Fermion Problems

Particles obeying Fermi–Dirac statistics (such as electrons or ^{3}He, for instance) pose particular challenges to Monte-Carlo simulation. If one tries to solve the Schrödinger equation of a many-body system by Monte-Carlo methods, one exploits its analogy with a diffusion equation (Ceperley and Kalos, 1979; Kalos, 1984). As mentioned at various places in this article, diffusion processes correspond to random walks and are hence accessible to Monte-Carlo simulation. However, while the diffusion equation (for one particle) considers the probability $P(\mathbf{r},t)$ that a particle starting at time $t = 0$ at the origin has reached the position $\mathbf{r}$ at time t, the wave function ψ in the Schrödinger equation is not positive definite. This fact creates severe problems for wave functions of many-fermion systems, since these wave functions must be antisymmetric, and the "nodal surface" in configuration space (where ψ changes sign) is unknown.

Formally the difficulty of applying importance-sampling techniques to distributions $\rho(\mathbf{r})$ that are not always positive can be overcome by splitting $\rho(\mathbf{r})$ into its sign, $s = \mathrm{sign}(\rho)$, and its absolute value, $\rho = s|\rho|$, and one can use $\tilde{\rho}(\mathbf{r}) = |\rho(\mathbf{r})|/\int|\rho(\mathbf{r})|d^3r$ as a probability density for importance sampling, and absorb the sign of $\rho(\mathbf{r})$ in the quantity to be measured (e.g., De Raedt and von der Linden, 1992). Symbolically, the average of

an observable A is obtained as

$$\langle A\rangle = \frac{\int d^3 r A(\mathbf{r}) s(\mathbf{r}) \tilde{\rho}(\mathbf{r})}{\int s(\mathbf{r}) \tilde{\rho}(\mathbf{r}) d^3 r} = \frac{\langle As\rangle_{\tilde{\rho}}}{\langle s\rangle_{\tilde{\rho}}},$$

where $\langle \ldots \rangle_{\tilde{\rho}}$ means averaging with $\tilde{\rho}$ as weight function. However, as is not unexpected, using $\tilde{\rho}$ one predominantly samples unimportant regions in phase space; therefore, in sampling the sign $\langle s\rangle_{\tilde{\rho}}$, one has large cancellations from regions where the sign is negative, and, for N degrees of freedom, one gets $\langle s\rangle_{\tilde{\rho}} \propto \exp(-\text{const.} \times N)$. This difficulty is known as the "minus-sign problem" and still hampers applications to fermion problems significantly!

Sometimes it is possible to start with a trial wave function where nodes are a reasonable first approximation to the actual nodes, and, starting with the population of random walks from this fixed-node approximation given by the trial function, one now admits walks that cross this nodal surface and sample the sign as indicated above. In this way, it has been possible to estimate the exchange-correlation energy of the homogeneous electron gas (Ceperley and Alder, 1980) over a wide range of densities very well.

7 Lattice Gauge Theory

Monte-Carlo simulation has become the primary tool for nonperturbative quantum chromodynamics, the field theory of quarks and hadrons and other elementary particles (e.g., Rebbi, 1984; De Grand, 1992). In this section, we first stress the basic problem, to make the analogy with the calculations of statistical mechanics clear. Then we very briefly highlight some of the results that have been obtained so far.

7.1 Some Basic Ideas of Lattice Gauge Theory

The theory of elementary particles is a field theory of gauge fields and matter fields. Choice of a lattice is useful to provide a cut-off that removes the ultraviolet divergences that would otherwise occur in these quantum field theories. The first step, hence, is the appropriate translation from the four-dimensional continuum (3 space + 1 time dimensions) to the lattice.

The generating functional (analogous to the partition function in statistical mechanics) is

$$Z = \int DAD\bar{\psi}\, D\psi \exp[-S_g(A, \bar{\psi}, \psi)], \tag{38}$$

where A represents the gauge fields, $\bar{\psi}$ and ψ represent the (fermionic) matter field, S_g is the action of the theory (containing a coupling constant g, which corresponds to inverse temperature in statistical mechanics as $\text{const.}/g^2 \to 1/k_B T$), and the symbols $\int D$ stand for functional integration. The action of the gauge field itself is, using the summation convention that indices that appear twice are summed over,

$$S_G = \tfrac{1}{4} \int d^4x F^{\alpha}_{\mu\nu}(x) F^{\mu\nu}_{\alpha}(x), \tag{39}$$

$F^{\alpha}_{\mu\nu}$ being the fields that derive from the vector potential $A^{\alpha}_{\mu}(x)$. These are

$$F^{\alpha}_{\mu\nu}(x) = \partial_\mu A^{\alpha}_{\nu}(x) - \partial_\nu A^{\alpha}_{\mu}(x) + g f^{\alpha}_{\beta\gamma} A^{\beta}_{\mu}(x) A^{\gamma}_{\nu}(x), \tag{40}$$

$f^{\alpha}_{\beta\gamma}$ being the structure constants of the gauge group, and g a coupling constant.

The fundamental variables that one then introduces are elements $U_\mu(x)$ of the gauge group G, which are associated with the links of the four-dimensional lattice,

connecting x and a nearest neighbor point $x + \mu$:

$$U_\mu(x) = \exp[igaT^\alpha A^\alpha_\mu(x)],$$
$$[U_m(x+m)]^\dagger = U_m(x), \qquad (41)$$

where a is the lattice spacing and T^α a group generator. Here $U^\dagger$ denotes the Hermitean conjugate of U. Wilson (1974) invented a lattice action that reduces in the continuum limit to Eq. (39), namely

$$\frac{S_U}{k_B T} = \frac{1}{g^2} \sum_n \sum_{\mu > \nu} \mathrm{Re\,Tr}\, U_\mu(n) \times U_\nu(n+\mu)\, U^\dagger_\mu(n+\nu)\, U^\dagger_\nu(n), \qquad (42)$$

where the links in Eq. (42) form a closed contour along an elementary plaquette of the lattice.

Using Eq. (42) in Eq. (38), which amounts to the study of a "pure" gauge theory (no matter fields), we recognize that the problem is equivalent to a statistical mechanics problem (such as spins on a lattice), the difference being that now the dynamical variables are the gauge group elements $U_\mu(n)$. Thus importance-sampling Monte-Carlo algorithms can be put to work, just as in statistical mechanics.

In order to include also matter fields, one starts from a partition function of the form

$$Z = \int DUD\bar{\psi}D\psi \times \exp\left\{-\frac{S_U}{k_{\bar{B}}T} + \sum_{i=1}^{n_f} \bar{\psi}\underline{M}\psi\right\} = \int DU(\det\underline{M})^{n_f} \exp\left(-\frac{S_U}{k_B T}\right), \qquad (43)$$

where we have assumed fermions with n_f degenerate "flavors." It has also been indicated that the fermion fields can be integrated out analytically, but the price is that one has to deal with the "fermion determinant" of the matrix $\underline{M}$. In principle, for any change of the U's this determinant needs to be recomputed; together with the fact that one needs to work on rather large lattices in four dimensions, in order to reproduce the continuum limit, this problem is responsible for the huge requirement of computing resources in this field.

It is clear that lattice gauge theory cannot be explained in depth on two pages – we only intend to give a vague idea of what these calculations are about to a reader who is not familiar with this subject.

7.2 A Recent Application

Among the many Monte-Carlo studies of various problems (which include problems in cosmology, like the phase transition from the quark-gluon plasma to hadronic matter in the early universe), we focus here on the problem of predicting the masses of elementary particles. Butler et al. (1993) have used a new massively parallel supercomputer with 480 processors ("GF11") exclusively for one year to run lattice sizes ranging from $8^3 \times 32$ to $24^3 \times 32$, $24^3 \times 36$, and $30 \times 32^2 \times 40$. Their program executes at a speed of more than 5 Gflops (Giga floating point operations per second), and the rather good statistics reached allowed a meaningful elimination of finite-size effects by an extrapolation to the infinite-volume limit. This problem is important, since the wave function of a hadron is spread out over many lattice sites.

Even with this impressive effort, several approximations are necessary:

1. The fermion determinant mentioned above is neglected (this is called "quenched approximation").

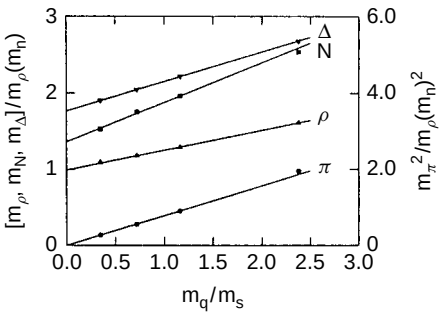


Fig. 3 For a $30 \times 32^2 \times 40$ lattice at $(k_B T)^{-1} = 6.17$, m_π^2, m_ρ, m_N, and m_Δ in units of the physical rho meson mass $m_\rho(m_n)$, as functions of the quark mass m_q in units of the strange quark mass m_s. Particles studied are pion, rho meson, nucleon, and delta baryon, respectively. From Butler et al. (1993)

2. One cannot work at the (physically relevant) very small quark mass m_q, but rather has to take data on the various hadron masses for a range of quark masses and extrapolate these data to zero (Fig. 3).

After a double extrapolation ($m_q \to 0$, lattice spacing at fixed volume $\to 0$), one obtains mass ratios that are in very satisfactory agreement with experiment. For example, for the nucleon the mass ratio for the finite volume is $m_N/m_\rho = 1.285 \pm 0.070$, extrapolated to infinite volume 1.219 ± 0.105, the experimental value being 1.222 (all masses in units of the mass m_ρ of a rho meson).

8 Selected Comments on Applications in Classical Statistical Mechanics of Condensed-Matter Systems

In this section, we mention a few applications very briefly, just in order to give the flavor of the type of work that is done and the kind of questions that are asked and answered by Monte-Carlo simulations. More extensive reviews can be found in the literature (Binder, 1976, 1979, 1984, 1992b).

8.1 Metallurgy and Materials Science

A widespread application of Monte-Carlo simulation in this area is the study of order-disorder phenomena in alloys: One tests analytical approximations to calculate phase diagrams, such as the cluster variation (CV) method, and one tests to what extent a simple model can describe the properties of complicated materials.

An example (Fig. 4) shows the order parameter for long-range order (LRO) and short-range order (SRO, for nearest neighbors) as function of temperature

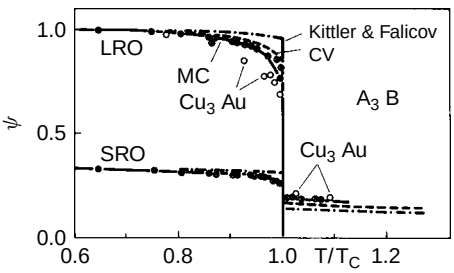


Fig. 4 Long-range order parameter (LRO) and absolute value of nearest-neighbor short-range order parameter (SRO) plotted versus temperature T (in units of the temperature T_c where the first-order transition occurs) for a nearest-neighbor model of binary alloys on the face-centered cubic lattice with A_3B structure. Open circles: experimental data for Cu_3Au; broken and dash-dotted curves: results of analytical theories. Full dots: Monte-Carlo results obtained from a simulation in the semi-grand canonical ensemble (chemical potential difference between A and B atoms treated as independent variable); circles with crosses: values obtained from a canonical ensemble simulation (concentration of B atoms fixed at 25%). From Binder et al. (1981)

for a model of Cu_3Au alloys on the fcc lattice. Here an Ising model with antiferromagnetic interactions between nearest neighbors only is studied, and the Monte-Carlo data (filled symbols and symbols with crosses) are compared to CV calculations (broken curve) and other analytical calculations (dash-dotted curve) and to experiment (open circles). The simulation shows that the analytical approximations describe the ordering of the model only qualitatively. Of course, there is no perfect agreement with the experimental data either; this is to be expected, of course, since in real alloys the interaction range is considerably larger than just extending to nearest neighbors only.

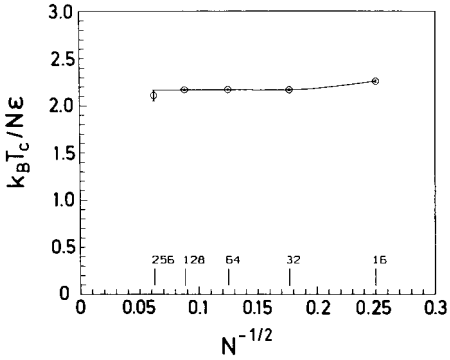


Fig. 5 Normalized critical temperature $k_BT/N\varepsilon$ of a symmetric polymer mixture (N = chain length, ε = energy parameter describing the repulsive interaction between *A*-*B* pairs of monomers) plotted versus $N^{-1/2}$. Data are results of simulations for the bond-fluctuation model, using N in the range from $N = 16$ to $N = 256$, as indicated. The data are consistent with an asymptotic extrapolation $k_BT_c/\varepsilon \approx 2.15N$, while Flory–Huggins theory (in the present units) would yield $k_BT_c/\varepsilon \approx 7N$, and the integral-equation theory $k_BT_c/\varepsilon \propto \sqrt{N}$. From Deutsch and Binder (1992)

8.2 Polymer Science

One can study phase transitions not only for models of magnets or alloys, of course, but also for complex systems such as mixtures of flexible polymers. A question heavily debated in the recent literature is the dependence of the critical temperature of unmixing of a symmetric polymer mixture (both constituents have the same degree of polymerization $N_A = N_B = N$) on chain length N. The classical Flory–Huggins theory predicted $T_c \propto N$, while a recent integral equation theory predicted $T_c \propto \sqrt{N}$ (Schweizer and Curro, 1990). This law would lead in the plot of Fig. 5 to a straight line through the origin. Obviously, the data seem to rule out this behavior, and are rather qualitatively consistent with Flory–Huggins theory (though the latter significantly overestimates the prefactor in the relation $T_c \propto N$).

Polymer physics provides examples for many scientific questions where simulations could contribute significantly to provide a better understanding. Figure 6 provides one more example (Paul et al., 1991a). The problem is to provide truly microscopic evidence for the reptation concept (Doi and Edwards, 1986). This concept implies that, as a result of "entanglements" between chains in a dense melt, each chain moves snakelike along its own contour. This behavior leads to a special behavior of mean square displacements: After a characteristic time τ_e, one should see a crossover from a law $g_1(t) \equiv \langle[\mathbf{r}_i(t) - \mathbf{r}_i(0)]^2\rangle \propto t^{1/2}$ (Rouse model) to a law $g_1(t) \propto t^{1/4}$, and, at a still later time (τ_R), one should see another crossover to $g_1(t) \propto t^{1/2}$ again. At the same time, the center of gravity displacement should also show an intermediate regime of anomalously slow diffusion, $g_3(t) \propto t^{1/2}$. Figure 6 provides qualitative evidence for these predictions – although

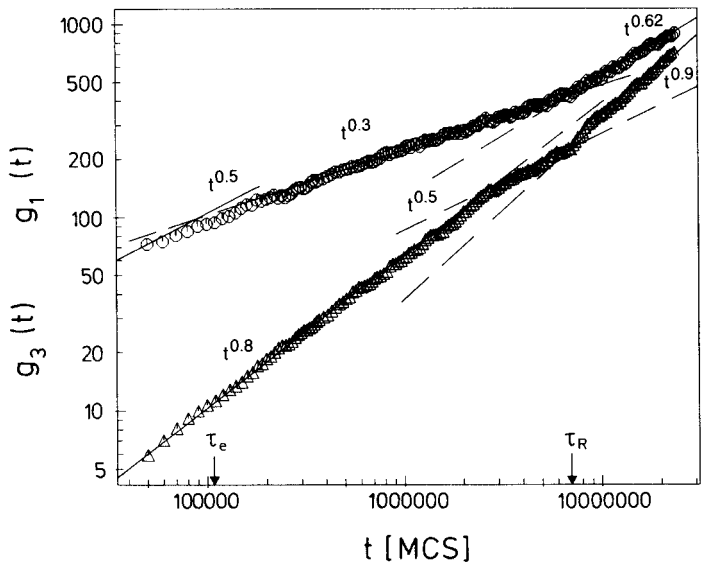


Fig. 6 Log-log plot of the mean square displacements of inner monomers [$g_1(t)$] and of the center of gravity of the chain [$g_3(t)$] versus time t (measured in units of Monte-Carlo steps, while lengths are measured in units of the lattice spacing). Straight lines show various power laws as indicated; various characteristic times are indicated by arrows (see text). Data refer to the bond-fluctuation model on the simple cubic lattice, for an athermal model of a polymer melt with chain length $N = 200$ and a volume fraction $\phi = 0.5$ of occupied lattice sites. From Paul et al. (1991a)

the effective exponents indicated do not quite have the expected values. A challenge for further theoretical explanation is the anomalous center-of-gravity diffusion in the initial Rouse regime, $g_3(t) \propto t^{0.8}$.

While this is an example where dynamic Monte-Carlo simulations are used to check theories – and pose further theoretical questions – one can also compare to experiment if one uses data in suitably normalized form. In Fig. 7, the diffusion constant D of the chains is normalized by its value in the Rouse regime (limit for small N) and plotted versus N/N_e where the characteristic "entanglement chain length" N_e is extracted from τ_e shown in Fig. 6 (see Paul et al., 1991a, b, for details).

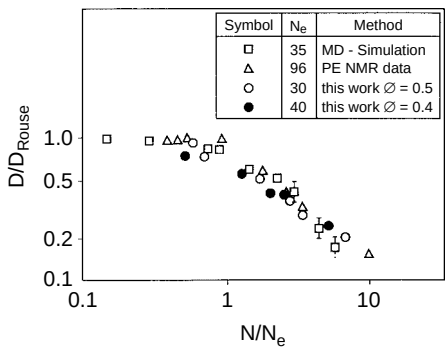


Fig. 7 Log-log plot of the self-diffusion constant D of polymer chains, normalized by the Rouse diffusivity, versus N/N_e (N_e = entanglement chain length, estimated independently and indicated in the inset). Circles: from Monte-Carlo simulations of Paul et al. (1991b); squares: from molecular dynamics simulations (Kremer and Grest, 1990); triangles: experimental data (Pearson et al., 1987). From Paul et al. (1991b)

The Monte-Carlo data presented in this scaled form agree with results from both a molecular dynamics (MD) simulation (Kremer and Grest, 1990) and experiment on polyethylene (PE) (Pearson et al., 1987).

This example also shows that, for slow diffusive motions, Monte-Carlo simulation is competitive to molecular dynamics, although it does not describe the fast atomic motions realistically.

8.3 Surface Physics

Our last example considers phenomena far from thermal equilibrium. Studying the ordering behavior of ordered superstructures, we treat the problem where initially the adatoms are adsorbed at random, and one gradually follows the formation of ordered domains out of initially disordered configurations. In a scattering experiment, one expects to see this by the gradual growth of a peak at the Bragg position $\mathbf{q}_B$. Figure 8 shows a simulation

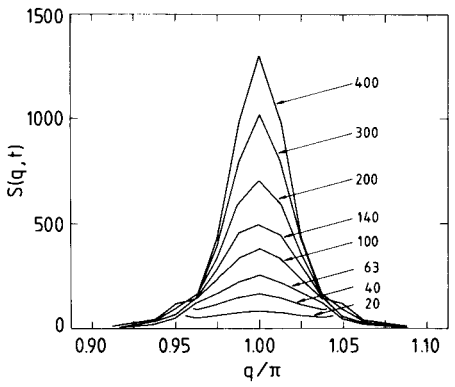


Fig. 8 Structure factor $S(q,t)$ versus wavevector q at various times t (in units MCS per lattice site), after the system was started in a completely random configuration. Temperature (measured in units of the nearest-neighbor repulsion W_{nn}) is 1.33 ($T_c \approx 2.07$ in these units), and coverage $\theta = \frac{1}{2}$. From Sadiq and Binder (1984)

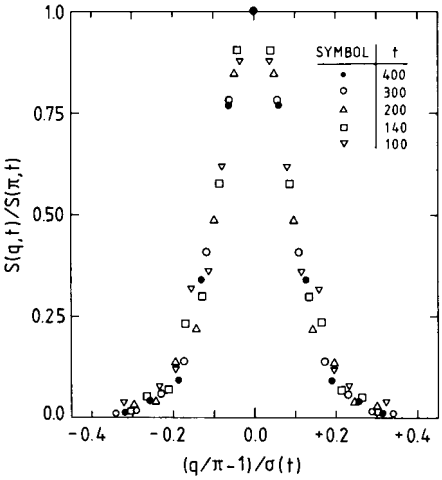


Fig. 9 Structure factor of Fig. 8 plotted in scaled form, normalizing $S(q,t)$ by its peak value $S(\pi, t)$ and normalizing $q/\pi - 1$ by the halfwidth $\sigma(t) = L^{-1}(t)$, where $L(t)$ thus defined is the characteristic domain size. From Sadiq and Binder (1984)

for the case of the (2×1) structure on the square lattice, where the Bragg position is at the Brillouin-zone boundary ($q_B = \pi$ if lengths are measured in units of the lattice spacing). Here a lattice gas with repulsive interactions between nearest and next-nearest neighbors (of equal strength) was used (Sadiq and Binder, 1984), using a single–spin-flip kinetics (if the lattice gas is translated to an Ising spin model), as is appropriate for a description of a monolayer in equilibrium with surrounding gas (the random "spin flips" then correspond to random evaporation-condensation events). Figure 9 presents evidence that these data on the kinetics of ordering satisfy a scaling hypothesis, namely

$$S(q, t) = [L(t)]^2 \tilde{S}(|\mathbf{q} - \mathbf{q}_B| L(t)), \qquad (44)$$

where $\tilde{S}$ is a scaling function. This hypothesis, Eq. (44), was first proposed on the basis of simulations, and later it was

established to describe experimental data as well.

Of course, surface physics provides many more examples where simulations have been useful; see Binder and Landau (1989) for a review.

9 Concluding Remarks

In this article, the basic features of the most widely used numerical techniques that fall into the category of Monte-Carlo calculations were described. There is a vast literature on the subject – the author estimates the number of papers using Monte-Carlo methods in condensed-matter physics of the order of 10^4, in lattice gauge theory of the order of 10^3! Thus many important variants of algorithms could not be treated here, and interesting applications (e.g., the study of neural-network models) were completely omitted.

There also exist other techniques for the numerical simulation of complex systems, which sometimes are an alternative approach to Monte-Carlo simulation. The molecular dynamics (MD) method (numerical integration of Newton's equations) has already been mentioned in the text, and there exist combinations of both methods (''hybrid Monte Carlo,'' ''Brownian dynamics,'' etc.). A combination of MD with the local-density approximation of quantum mechanics is the basis of the Car–Parrinello method.

Problems like that shown in Figs. 8 and 9 can also be formulated in terms of numerically solving appropriate differential equations, which may in turn even be discretized to cellular automata. When planning a Monte Carlo simulation, hence, some thought to the question ''When which method?'' should be given.

Glossary

Critical Slowing Down: Divergence of the relaxation time of dynamic models of statistical mechanics at a second-order phase transition (critical point).

Detailed Balance Principle: Relation linking the transition probability for a move and the transition probability for the inverse move to the ratio of the probability for the occurrence of these states in thermal equilibrium. This condition is sufficient for a Markov process to tend toward thermal equilibrium.

Ergodicity: Property that ensures equality of statistical ensemble averages (such as the ''canonic ensemble'' of statistical mechanics) and time averages along the trajectory of the system through phase space.

Finite-Size Scaling: Theory that describes the finite-size-induced rounding of singularities that would occur at phase transitions in the thermodynamic limit.

Heat-Bath Method: Choice of transition probability where the probability to ''draw'' a trial value for a degree of freedom does not depend on its previous value.

Importance Sampling: Monte-Carlo method that chooses the states that are generated according to the desired probability distribution. For example, for statistical mechanics applications, states are chosen with weights proportional to the Boltzmann factor.

Lattice Gauge Theory: Field theory of quarks and gluons in which space and time are discretized into a four-dimensional lattice, gauge field variables being associated to the links of the lattice.

Master Equation: Rate equation describing the "time" evolution of the probability that a state occurs as a function of a "time" coordinate labeling the sequence of states (in the context of importance-sampling Monte-Carlo methods).

Molecular-Dynamics Method: Simulation method for interacting many-body systems based on numerical integration of the Newtonian equations of motion.

Monte-Carlo Step: Unit of (pseudo) time in (dynamically interpreted) importance sampling where, on the average, each degree of freedom in the system gets one chance to be changed (or "updated").

Random Number Generator (RNG): Computer subroutine to produce pseudorandom numbers that are approximately not correlated with each other and approximately uniformly distributed in the interval from zero to one. RNGs typically are strictly periodic, but the period is large enough that, for practical applications, this periodicity does not matter.

Simple Sampling: Monte-Carlo method that chooses states uniformly and at random from the available phase space.

Transition Probability: Probability that controls the move from one state to the next one in a Monte-Carlo process.

List of Works Cited

Ahrens, J. H., Dieter, U. (1979), *Pseudo Random Numbers*, New York: Wiley.

Barber, M. N. (1983), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 8 New York: Academic, Chap. 2.

Barkema, G. T., Marko, J. F. (1993), *Phys. Rev. Lett.* **71**, 2070–2073.

Berne, B. J., Thirumalai, D. (1986), *Annu. Rev. Phys. Chem.* **37**, 401.

Binder, K. (1976), in: C. Domb, M.S. Green (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 5b, New York: Academic, p. 1.

Binder, K. (1979) (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer.

Binder, K. (1984) (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer.

Binder, K. (1987), *Ferroelectrics* **73**, 43–67.

Binder, K. (1992a), *Annu. Rev. Phys. Chem.* **43**, 33–59.

Binder, K. (1992b) (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer.

Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics: An Introduction*, Berlin: Springer.

Binder, K., Landau, D. P. (1989), in: K. P. Lawley (Ed.), *Molecule-Surface Interaction*, New York: Wiley, pp. 91–152.

Binder, K., Lebowitz, J. L., Phani, M. K., Kalos, M. H. (1981), *Acta Metall.* **29**, 1655–1665.

Bruns, W., Motoc, I., O'Driscoll, K. F. (1981), *Monte Carlo Applications in Polymer Science*, Berlin: Springer.

Butler, F., Chen, H., Sexton, J., Vaccarino, A., Weingarten, D. (1993), *Phys. Rev. Lett.* **70**, 2849–2852.

Carlson, J. (1988), *Phys. Rev. C* **38**, 1879–1885.

Ceperley, D. M., Alder, B. J. (1980), *Phys. Rev. Lett.* **45**, 566–569.

Ceperley, D. M., Kalos, M. H. (1979), in: K. Binder (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer, pp. 145–194.

Ceperley, D. M., Pollock, E. L. (1987), *Can. J. Phys.* **65**, 1416–1420.

Ciccotti, G., Hoover, W. G. (1986) (Eds.), *Molecular Dynamics Simulation of Statistical Mechanical Systems*, Amsterdam: North-Holland.

Compagner, A. (1991), *Am. J. Phys.* **59**, 700–705.

Compagner, A., Hoogland, A. (1987), *J. Comput. Phys.* **71**, 391–428.

De Grand, T. (1992), in: H. Gausterer, C. B. Lang (Eds.), *Computational Methods in Field Theory*, Berlin: Springer, pp. 159–203.

De Raedt, H., von der Linden, W. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, pp. 249–284.

Deutsch, H. P., Binder, K. (1992), *Macromolecules* **25**, 6214–6230.

Doi, M., Edwards, S. F. (1986), *Theory of Polymer Dynamics*, Oxford: Clarendon Press.

Ferrenberg, A. M., Landau, D. P., Wong, Y. J. (1992), *Phys. Rev. Lett.* **69**, 3382–3384.
Frick, M., Pattnaik, P. C., Morgenstern, I., Newns, D. M., von der Linden, W. (1990), *Phys. Rev. B* **42**, 2665–2668.
Gillan, M. J., Christodoulos, F. (1993), *Int. J. Mod. Phys.* **C4**, 287–297.
Gould, H., Tobochnik, J. (1988), *An Introduction to Computer Simulation Methods/Applications to Physical Systems, Parts 1 and 2*, Reading, MA: Addison-Wesley.
Hammersley, J. M., Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Chapman and Hall.
Heermann, D. W. (1986), *Computer Simulation Methods in Theoretical Physics*, Berlin: Springer.
Herrmann, H. J. (1986), *Phys. Rep.* **136**, 153–227.
Herrmann, H. J. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 5.
Hockney, R. W., Eastwood, J. W. (1988), *Computer Simulation using Particles*, Bristol: Adam Hilger.
James, F. (1990), *Comput. Phys. Commun.* **60**, 329–344.
Kalos, M. H. (1984) (Ed.), *Monte Carlo Methods in Quantum Problems*, Dordrecht: Reidel.
Kalos, M. H., Whitlock, P. A. (1986), *Monte Carlo Methods*, Vol. 1, New York: Wiley.
Kirkpatrick, S., Stoll, E. (1981), *J. Comput. Phys.* **40**, 517–526.
Knuth, D. (1969), *The Art of Computer Programming*, Vol. 2, Reading, MA: Addison-Wesley.
Koonin, S. E. (1981), *Computational Physics*, Reading, MA: Benjamin.
Kremer, K., Binder, K. (1988), *Comput. Phys. Rep.* **7**, 259–310.
Kremer, K., Grest, G. S. (1990), *J. Chem. Phys.* **92**, 5057–5086.
Landau, D. P. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 2.
Lehmer, D. H. (1951), in: *Proceedings of the 2nd Symposium on Large-Scale Digital Computing Machinery*, Harvard University, Cambridge, 1951, pp. 142–145.
Levesque, D., Weis, J. J. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, Chap. 6.
Marsaglia, G. A. (1985), in: L. Billard (Ed.), *Computer Science and Statistics: The Interface*, Amsterdam: Elsevier, Chap. 1.
Marsaglia, G. A. (1986), *Proc. Natl. Acad. Sci. U.S.A.* **61**, 25–28.
Marsaglia, G. A., Narasumhan, B., Zaman, A. (1990), *Comput. Phys. Comm.* **60**, 345–349.
Marx, D., Opitz, O., Nielaba, P., Binder, K. (1993), *Phys. Rev. Lett.* **70**, 2908–2911.
Meakin, P. (1988), in: C. Domb, J. L. Lebowitz (Eds.), *Phase Transitions and Critical Phenomena*, Vol. 12, New York: Academic, pp. 336–489.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. M., Teller, E. (1953), *J. Chem. Phys.* **21**, 1087–1092.
Panagiotopoulos, A. Z. (1992), *Molec. Simul.* **9**, 1–23.
Paul, W., Binder, K., Heermann, D. W., Kremer, K. (1991a), *J. Chem. Phys.* **95**, 7726–7740.
Paul, W., Binder, K., Heermann, D. W., Kremer, K. (1991b), *J. Phys. (Paris)* **111**, 37–60.
Pearson, D. S., Verstrate, G., von Meerwall, E., Schilling, F. C. (1987), *Macromolecules* **20**, 113–1139.
Privman, V. (1990) (Ed.), *Finite Size Scaling and Numerical Simulation of Statistical Systems*, Singapore: World Scientific.
Rebbi, C. (1984), in: K. Binder (Ed.), *Application of the Monte Carlo Method in Statistical Physics*, Berlin: Springer, p. 277.
Reger, J. D., Young, A. P. (1988), *Phys. Rev. B* **37**, 5978–5981.
Rosenbluth, M. N., Rosenbluth, A. W. (1955), *J. Chem. Phys.* **23**, 356–362.
Sadiq, A., Binder, K. (1984), *J. Stat. Phys.* **35**, 517–585.
Schmidt, K. E., Ceperley, D. M. (1992), in: K. Binder (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer, pp. 203–248.
Schmidt, K. E., Kalos, M. H. (1984), in: K. Binder (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer, pp. 125–149.
Schweizer, K. S., Curro, J. G. (1990), *Chem. Phys.* **149**, 105–127.
Sokol, P. E., Sosnick, T. R., Snow, W. M. (1989), in: R. E. Silver, P. E. Sokol (Eds.), *Momentum Distributions*, New York: Plenum Press.
Stanley, H. E. (1971), *An Introduction to Phase Transitions and Critical Phenomena*, Oxford: Oxford University Press.
Stauffer, D. (1985), *An Introduction to Percolation Theory*, London: Taylor & Francis.
Suzuki, M. (1986) (Ed.), *Quantum. Monte Carlo Methods*, Berlin: Springer.
Swendsen, R. H., Wang, J. S., Ferrenberg, A. M. (1992), in: K. Binder (Ed.), *The Monte Carlo*

Method in Condensed Matter Physics, Berlin: Springer, Chap. 4.
Tausworthe, R. C. (1965), *Math. Comput.* **19**, 201–208.
Tolman, S., Meakin, P. (1989), *Phys. Rev. A* **40**, 428–437.
Vicsek, T. (1989), *Fractal Growth Phenomena*, Singapore: World Scientific.
West, G. B. (1975), *Phys. Rep.* **18C**, 263–323.
Wilson, K. (1974), *Phys. Rev. D* **10**, 2445–2453.

Further Reading

A textbook describing for the beginner how to learn to write Monte-Carlo programs and to analyze the output generated by them has been written by Binder, K., Heermann, D. W. (1988), *Monte Carlo Simulation in Statistical Physics: An Introduction*, Berlin: Springer. This book emphasizes applications of statistical mechanics such as random walks, percolation, and the Ising model.

A useful book that gives much weight to applications outside of statistical mechanics is Kalos, M. H., Whitlock, P. A. (1986), *Monte Carlo Methods, Vol. 1*, New York: Wiley.

A more general but pedagogic introduction to computer simulation is presented in Gould, H., Tobochnik, J. (1988), *An Introduction to Computer Simulation Methods/Applications to Physical Systems, Parts 1 and 2*, Reading, MA: Addison-Wesley.

A rather systematic collection of applications of Monte-Carlo studies in statistical mechanics and condensed matter physics has been compiled in a series of books edited by the author of this article: Binder, K. (1979) (Ed.), *Monte Carlo Methods in Statistical Physics*, Berlin: Springer; Binder, K. (1984) (Ed.), *Applications of the Monte Carlo Method in Statistical Physics*, Berlin: Springer; and Binder, K. (1992) (Ed.), *The Monte Carlo Method in Condensed Matter Physics*, Berlin: Springer.

Finally we draw attention to two important areas which are only briefly covered in the present article: For quantum problems, see Suzuki, M. H. (1987) (Ed.), *Quantum Monte Carlo Methods*, Berlin: Springer; Doll, J. D., Gubernaitis, J. E. (1990) (Eds.), *Quantum Simulations*, Singapore: World Scientific. For lattice gauge theory, see Bunk, B., Mütter, K. H., Schilling, K. (1986) (Eds.), *Lattice Gauge Theory, A Challenge in Large-Scale Computing*, New York: Plenum; De Grand, T. (1992), in: H. Gausterer, C. B. Lang (Eds.), *Computational Methods in Field Theory*, Berlin: Springer, pp. 159–203; Smit, J., van Baal, P. (1993) (Eds.), *Lattice 92*, Amsterdam: North Holland.

Numerical Methods

Christina C. Christara and Kenneth R. Jackson
Computer Science Department, University of Toronto, Toronto, Ontario, Canada

Introduction ***284***
1 Floating-Point Arithmetic ***284***
1.1 The IEEE Standard *284*
1.2 Rounding Errors *286*
1.3 The Effects of Inexact Arithmetic: Some Illustrative Examples *287*
2 The Direct Solution of Linear Algebraic Systems ***290***
2.1 Gaussian Elimination *290*
2.2 Back Substitution *292*
2.3 The *LU* Factorization *292*
2.4 Forward Elimination *293*
2.5 Scaling and Pivoting *293*
2.6 The Cholesky Factorization *297*
2.7 Banded and Sparse Matrices *298*
2.8 Rounding Errors, Condition Numbers, and Error Bounds *301*
2.9 Iterative Improvement *303*
3 The Iterative Solution of Linear Algebraic Systems ***304***
3.1 Basic Iterative Methods *304*
3.2 The Conjugate-Gradient Method *309*
4 Overdetermined and Underdetermined Linear Systems ***312***
4.1 The Normal Equations for Overdetermined Linear Systems *312*
4.2 The Normal Equations for Underdetermined Linear Systems *312*
4.3 Householder Transformations and the *QR* Factorization *313*
4.4 Using the *QR* Factorization to Solve Overdetermined Linear Systems *314*
4.5 Using the *QR* Factorization to Solve Underdetermined Linear Systems *314*
4.6 The Gram–Schmidt Orthogonalization Algorithm *314*
4.7 Using Gram–Schmidt to Solve Overdetermined Linear Systems *315*
4.8 Using Gram–Schmidt to Solve Underdetermined Linear Systems *315*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

5 Eigenvalues and Eigenvectors of Matrices *315*
5.1 The Power Method *316*
5.2 The *QR* Method *317*
5.3 Transforming a Symmetric Matrix to Tridiagonal Form *318*
5.4 Inverse Iteration *319*
5.5 Other Methods *319*
6 Nonlinear Algebraic Equations and Systems *320*
6.1 Fixed-Point Iteration *320*
6.2 Newton's Method for Nonlinear Equations *320*
6.3 The Secant Method *321*
6.4 The Bisection and Regula Falsi Methods *321*
6.5 Convergence *321*
6.6 Rate of Convergence *322*
6.7 Newton's Method for Systems of Nonlinear Equations *322*
6.8 Modifications and Alternatives to Newton's Method *323*
6.9 Polynomial Equations *323*
6.10 Horner's Rule *324*
7 Unconstrained Optimization *324*
7.1 Some Definitions and Properties *325*
7.2 The Fibonacci and Golden-Section Search Methods *325*
7.3 The Steepest-Descent Method *327*
7.4 Conjugate-Direction Methods *327*
7.5 The Conjugate-Gradient Method *328*
7.6 Newton's Method *328*
7.7 Quasi-Newton Methods *328*
8 Approximation *328*
8.1 Polynomial Approximation *329*
8.2 Polynomial Interpolation *329*
8.2.1 Monomial Basis *330*
8.2.2 Lagrange Basis *330*
8.2.3 Newton Basis and Divided Differences *330*
8.3 Polynomial Interpolation with Derivative Data *331*
8.4 The Error in Polynomial Interpolation *331*
8.5 Piecewise Polynomials and Splines *332*
8.5.1 Constant Splines *332*
8.5.2 Linear Splines *333*
8.5.3 Quadratic Splines *333*
8.5.4 Quadratic Piecewise Polynomials *333*
8.5.5 Cubic Splines *334*
8.5.6 Cubic Hermite Piecewise Polynomials *334*
8.6 Piecewise Polynomial Interpolation *334*
8.6.1 Linear Spline Interpolation *335*
8.6.2 Quadratic Spline Interpolation *335*
8.6.3 Cubic Spline Interpolation *335*

8.6.4 Cubic Hermite Piecewise Polynomial Interpolation *336*
8.7 Least-Squares Approximation *336*
8.7.1 Orthogonal Polynomials *337*
8.7.2 The Gram–Schmidt Orthogonalization Algorithm *337*
8.7.3 Constructing the Least-Squares Polynomial Approximation *337*
9 Numerical Integration – Quadrature *338*
9.1 Simple Quadrature Rules *338*
9.1.1 Some Definitions *338*
9.1.2 Gaussian Quadrature Rules *340*
9.1.3 Translating the Interval of Integration *340*
9.1.4 Comparison of Gaussian and Newton–Cotes Quadrature Rules *341*
9.2 Composite (Compound) Quadrature Rules *341*
9.3 Adaptive Quadrature *343*
9.4 Romberg Integration and Error Estimation *343*
9.5 Infinite Integrals and Singularities *344*
9.6 Monte-Carlo Methods *345*
10 Ordinary Differential Equations *346*
10.1 Initial-Value Problems (IVPs) *346*
10.1.1 Two Simple Formulas *346*
10.1.2 Stiff IVPs *347*
10.1.3 Solving Implicit Equations *349*
10.1.4 Higher-order Formulas *350*
10.1.5 Runge–Kutta Formulas *350*
10.1.6 Linear Multistep Formulas *352*
10.1.7 Adams Formulas *353*
10.1.8 Backward Differentiation Formulas *353*
10.1.9 Other Methods *353*
10.1.10 Adaptive Methods *354*
10.2 Boundary-Value Problems (BVPs) *354*
10.2.1 Shooting Methods *354*
10.2.2 One-Step Methods *356*
10.2.3 Other Methods *356*
11 Partial Differential Equations (PDEs) *357*
11.1 Classes of Problems and PDEs *357*
11.1.1 Some Definitions *358*
11.1.2 Boundary Conditions *359*
11.2 Classes of Numerical Methods for PDEs *359*
11.2.1 Analysis of Numerical Methods for PDEs *360*
11.2.1.1 Convergence Analysis (for BVPs and IVPs) *360*
11.2.1.2 Stability Analysis (for IVPs) *360*
11.2.1.3 Time (Computational) Complexity Analysis *360*
11.2.1.4 Memory Complexity Analysis *361*
11.2.1.5 Overall Efficiency Analysis *361*

11.3 Finite-Difference Methods for BVPs *361*
11.3.1 An Example of a Finite-Difference Method in One Dimension *362*
11.3.2 An Example of a Finite-Difference Method in Two Dimensions *363*
11.4 Finite-Element Methods for BVPs *365*
11.4.1 The Galerkin Method *366*
11.4.2 The Collocation Method *367*
11.5 Finite-Difference Methods for IVPs *369*
11.5.1 An Example of an Explicit One-Step Method for a Parabolic IVP *370*
11.5.2 An Example of an Implicit One-Step Method for a Parabolic IVP *370*
11.5.3 An Example of an Explicit Two-Step Method for a Hyperbolic IVP *371*
11.6 The Method of Lines *372*
11.7 Boundary-Element Methods *373*
11.8 The Multigrid Method *373*
12 Parallel Computation *375*
12.1 Cyclic Reduction *375*
13 Sources of numerical software *377*
Glossary *379*
Mathematical Symbols Used *380*
Abbreviations Used *381*
List of Works Cited *381*
Further Reading *383*

Introduction

Numerical methods are an indispensable tool in solving many problems that arise in science and engineering. In this article, we briefly survey a few of the most common mathematical problems and review some numerical methods to solve them.

As can be seen from the table of contents, the topics covered in this survey are those that appear in most introductory numerical methods books. However, our discussion of each topic is more brief than is normally the case in such texts and we frequently provide references to more advanced topics that are not usually considered in introductory books.

Definitions of the more specialized mathematical terms used in this survey can be found in the Glossary at the end of the article. We also list some mathematical symbols and abbreviations used throughout the survey in the two sections following the Glossary.

1 Floating-Point Arithmetic

In this section, we consider the representation of floating-point numbers, floating-point arithmetic, rounding errors, and the effects of inexact arithmetic in some simple examples. For a more detailed discussion of these topics, see Wilkinson (1965) and Goldberg (1991) or an introductory numerical methods text.

1.1 The IEEE Standard

The approval of the IEEE (Institute of Electrical and Electronics Engineers)

Standard for Binary Floating-Point Arithmetic (IEEE, 1985) was a significant advance for scientific computation. Not only has this led to cleaner floating-point arithmetic than was commonly available previously, thus greatly facilitating the development of reliable, robust numerical software, but, because many computer manufacturers have since adopted the standard, it has significantly increased the portability of programs.

The IEEE standard specifies both single- and double-precision floating-point numbers, each of the form

$$(-1)^s \times b_0.b_1b_2\cdots b_{p-1} \times 2^E, \qquad (1)$$

where $s = 0$ or 1, $(-1)^s$ is the *sign* of the number, $b_i = 0$ or 1 for $i = 0, \ldots, p-1$; $b_0.b_1b_2\cdots b_{p-1}$ is the *significand* (sometimes called the *mantissa*) of the number, and the *exponent* E is an integer satisfying $E_{\min} \le E \le E_{\max}$. In single-precision, $p = 24$, $E_{\min} = -126$ and $E_{\max} = +127$, whereas, in double-precision, $p = 53$, $E_{\min} = -1022$, and $E_{\max} = +1023$. We emphasize that a number written in the form (1) is binary. So, for example, $1.100\cdots0 \times 2^0$ written in the format (1) is equal to the decimal number 1.5

A *normalized* number is either 0 or a floating-point number of the form (1) with $b_0 = 1$ (and so it is not necessary to store the leading bit). In single-precision, this provides the equivalent of 7 to 8 significant decimal digits with positive and negative numbers having magnitudes roughly in the range $[1.2 \times 10^{-38}, 3.4 \times 10^{+38}]$. In double-precision, this is increased to about 16 significant decimal digits and a range of roughly $[2.2 \times 10^{-308}, 1.8 \times 10^{+308}]$.

An *underflow* occurs when an operation produces a nonzero result in the range $(-2^{E_{\min}}, +2^{E_{\min}})$. In the IEEE standard, the default is to raise an underflow exception flag and to continue the computation with the result correctly rounded to the nearest *denormalized* number or zero. A denormalized floating-point number has the form (1) with $E = E_{\min}$ and $b_0 = 0$. Because denormalized numbers use some of the leading digits from the significand to represent the magnitude of the number, there are fewer digits available to represent its significant digits. Using denormalized numbers in this way for underflows is sometimes referred to as *gradual underflow*. Many older non-IEEE machines do not have denormalized numbers and, when an underflow occurs, they either abort the computation or replace the result by zero.

An *overflow* occurs when an operation produces a nonzero result outside the range of floating-point numbers of the form (1). In the IEEE standard, the default is to raise an overflow exception flag and to continue the computation with the result replaced by either $+\infty$ or $-\infty$, depending on the sign of the overflow value. Many older non-IEEE machines do not have $+\infty$ or $-\infty$ and, when an overflow occurs, they usually abort the computation.

The IEEE standard also includes at least two NaNs (Not-a-Number) in both precisions, representing indeterminate values that may arise from invalid or inexact operations such as $(+\infty) + (-\infty)$, $0 \times \infty$, $0/0$, ∞/∞, or $\sqrt{x}$ for $x < 0$. When a NaN arises in this way, the default is to raise an exception flag and to continue the computation. This novel feature is not available on most older non-IEEE machines.

It follows immediately from the format (1) that floating-point numbers are discrete and finite, whereas real numbers are dense and infinite. As a result, an arithmetic operation performed on two floating-point numbers may return a result that cannot be represented exactly in the form (1) in the same precision as the operands.

A key feature of the IEEE standard is that it requires that the basic arithmetic operations $+, -, \times, /$, and $\sqrt{}$ return *properly rounded* results. That is, we may think of the operation as first being done exactly and then properly rounded to the precision of the result. An operation with $\pm\infty$ is interpreted as the limiting case of the operation with an arbitrary large value in place of the ∞, when such an interpretation makes sense; otherwise the result is a NaN. An operation involving one or more NaNs returns a NaN.

The default rounding mode is *round-to-nearest;* that is, the exact result of the arithmetic operation is rounded to the nearest floating-point number, where, in the case of a tie, the floating-point number with the least significant bit equal to 0 is selected. The standard also provides for directed roundings (round-towards- $+\infty$, round-towards- $-\infty$, and round-towards-0), but these are not easily accessed from most programming languages.

Another important feature of the IEEE standard is that comparisons are exact and never overflow or underflow. The comparisons $<, >$, and $=$ work as expected with finite floating-point numbers of the form (1) and $-\infty < x < +\infty$ for any finite floating-point number x. NaNs are unordered and the comparison of a NaN with any other value – including itself – returns false.

The IEEE standard also provides for *extended* single- and double-precision floating-point numbers, but these are not easily accessed from most programming languages, and so we do not discuss them here.

As noted above, the IEEE standard has been widely adopted in the computer industry, but there are several important classes of machines that do not conform to it, including Crays, DEC Vaxes, and IBM mainframes. Although their floating-point numbers are similar to those described above, there are important differences. Space limitations, though, do not permit us to explore these systems here.

1.2 Rounding Errors

Since IEEE standard floating-point arithmetic returns the correctly rounded result for the basic operations $+, -, \times, /$, and $\sqrt{}$, one might expect that rounding errors would never pose a problem, particularly in double-precision computations. Although rounding errors can be ignored in many cases, the examples in the next subsection show that they may be significant even in simple calculations. Before considering these examples, though, we need to define an important machine constant and describe its significance in floating-point computation.

Machine epsilon, often abbreviated *mach-eps*, is the distance from 1 to the next larger floating-point number. For numbers of the form (1), mach-eps $= 2^{1-p}$. So for IEEE single- and double-precision numbers, mach-eps is $2^{-23} \approx 1.19209 \times 10^{-7}$ and $2^{-52} \approx 2.22045 \times 10^{-16}$, respectively.

A common alternative definition of machine epsilon is that it is the smallest positive floating-point number $\in$ such that $1 + \in > 1$ in floating-point arithmetic. We prefer the definition given in the previous paragraph, and use it throughout this section, because it is independent of the rounding mode and so is characteristic of the floating-point number system itself, while the alternative definition given in this paragraph depends on the rounding mode as well. We also assume throughout this section that round-to-nearest is in effect. The discussion below, though, can be modified easily for the alternative

definition of mach-eps and other rounding modes.

It follows immediately from the floating-point format (1) that the absolute distance between floating-point numbers is not uniform. Rather, from (1) and the definition of mach-eps above, we see that the spacing between floating-point numbers in the intervals $[2^k, 2^{k+1})$ and $(-2^{k+1}, -2^k]$ is mach-eps $\times 2^k$ for $E_{\min} \leq k \leq E_{\max}$. Thus, the absolute distance between neighboring nonzero normalized floating-point numbers with the same exponent is uniform, but the floating-point numbers near $2^{E_{\min}}$ are much closer together in an absolute sense than those near $2^{E_{\max}}$. However, the *relative* spacing between all nonzero normalized floating-point numbers does not vary significantly. It is easy to see that, if x_1 and x_2 are any two neighboring nonzero normalized floating-point numbers, then

$$\frac{\text{mach-eps}}{2} \leq \left| \frac{x_1 - x_2}{x_1} \right| \leq \text{mach-eps}. \tag{2}$$

As a result, it is more natural to consider relative, rather than absolute, errors in arithmetic operations on floating-point numbers, as is explained in more detail below.

For $x \in \mathbb{R}$, let $fl(x)$ be the floating-point number nearest to x, where, in the case of a tie, the floating-point number with the least significant bit equal to 0 is selected. The importance of mach-eps stems largely from the observation that, if $fl(x)$ does not overflow or underflow, then

$$fl(x) = x(1 + \delta) \text{ for some } |\delta| \leq u, \tag{3}$$

where u, the *relative roundoff error bound*, satisfies $u = \text{mach-eps}/2$ for round-to-nearest and $u = \text{mach-eps}$ for the other IEEE rounding modes. Rewriting (3) as

$$\delta = \frac{[fl(x) - x]}{x},$$

we see that δ is the relative error incurred in approximating x by $fl(x)$, and so (3) is closely related to (2).

If op is one of $+, -, \times$, or/and x and y are two floating-point numbers, let $fl(x \text{ op } y)$ stand for the result of performing the arithmetic operation x op y in floating-point arithmetic. If no arithmetic exception arises in the floating-point operation, then

$$fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta) \text{ for some } |\delta| \leq u, \tag{4}$$

where the $(x \text{ op } y)$ on the right side of (4) is the exact result of the arithmetic operation. Similarly, if x is a nonnegative normalized floating-point number, then

$$fl(\sqrt{x}) = \sqrt{x}(1 + \delta) \text{ for some } |\delta| \leq u. \tag{5}$$

Again, the u in either (4) or (5) is the relative roundoff error bound and the δ is the relative error incurred in approximating x op y by $fl(x \text{ op } y)$ or $\sqrt{x}$ by $fl(\sqrt{x})$, respectively.

Although the relations (4)–(5) are not quite as tight as the requirement that the basic operations $+, -, \times, /$, and $\sqrt{}$ return the correctly rounded result, they are very useful in deriving error bounds and explaining the effects of rounding errors in computations.

1.3 The Effects of Inexact Arithmetic: Some Illustrative Examples

As noted at the start of the last subsection, since IEEE standard floating-point arithmetic returns the correctly rounded result for the basic operations $+, -, \times, /$, and $\sqrt{}$, one might expect that rounding errors would never pose a problem, particularly in double-precision computations. This,

though, is not the case. In this subsection we consider a simple example that illustrates some of the *pitfalls* of numerical computation.

Suppose that we compute the expression

$$1 + 10^{10} - 10^{10} \tag{6}$$

in single precision from left to right. We first compute $fl(1 + 10^{10}) = 10^{10}$, the correctly rounded single-precision result. Then we use this value to compute $fl(10^{10} - 10^{10}) = 0$, without committing an additional rounding error. Thus, $fl((1 + 10^{10}) - 10^{10}) = 0$, whereas the true result is 1.

The key point to note here is that $fl(1 + 10^{10}) = 10^{10} = (1 + 10^{10})(1 + \delta)$, where $|\delta| = 1/(1 + 10^{10}) < 10^{-10} < u = 2^{-24}$. So the rounding error that we commit in computing $1 + 10^{10}$ is small relative to $1 + 10^{10}$, the true result of the first addition, but the absolute error of 1 associated with this addition is not small compared with the true final answer, 1, thus illustrating the **Rule for Sums**:

> Although a rounding error is always small relative to the result that gives rise to it, it might be large relative to the true final answer if intermediate terms in a sum are large relative to the true final answer.

The Rule for Sums is important to remember when computing more complex expressions such as the truncated Taylor series $T_k(x) = 1 + x + x^2/2 + \cdots + x^k/k! \approx e^x$, where k is chosen large enough so that the *truncation error*

$$e^x - T_k(x) = \sum_{i=k+1}^{\infty} \frac{x^i}{i!}$$

is insignificant relative to e^x. It is easy to prove that, if $x \geq 0$ and k is large enough, then $T_k(x)$ is a good approximation to e^x. However, if $x < 0$ and of moderate magnitude, then the rounding error associated with some of the intermediate terms $x^i/i!$ in $T_k(x)$ might be much larger in magnitude than either the true value of $T_k(x)$ or e^x. As a result, $fl(T_k(x))$, the computed value of $T_k(x)$, might be completely erroneous, no matter how large we choose k. For example, $e^{-15} \approx 3.05902 \times 10^{-7}$, while we computed $fl(T_k(x)) \approx 2.12335 \times 10^{-2}$ in IEEE single-precision arithmetic on a Sun Sparcstation.

A similar problem is less likely to occur with multiplications, provided no overflow or underflow occurs, since from (4)

$$fl(x_1 \cdot x_2 \cdots x_n) = x_1 \cdot x_2 \cdots x_n(1 + \delta_1) \cdots (1 + \delta_{n-1}), \tag{7}$$

where $|\delta_i| \leq u$ for $i = 1, \ldots, n - 1$. Moreover, if $nu \leq 0.1$, then $(1 + \delta_1) \cdots (1 + \delta_{n-1}) = 1 + 1.1n\delta$ for some $\delta \in [-u, u]$. Therefore, unless n is very large, $fl(x_1 \cdot x_2 \cdots x_n)$ is guaranteed to be a good approximation to $x_1 \cdot x_2 \cdots x_n$. However, it is not hard to find examples for which $nu \gg 1$ and $fl(x_1 \cdot x_2 \cdots x_n)$ is a poor approximation to $x_1 \cdot x_2 \cdots x_n$.

The example (6) also illustrates another important phenomenon commonly called *catastrophic cancellation*: all the digits in the second sum, $fl(10^{10} - 10^{10})$, cancel, signaling a catastrophic loss of precision. Catastrophic cancellation refers also to the case that many, but not all, of the digits cancel. This is often a sign that a disastrous loss of accuracy has occurred, but, as in this example when we compute $fl(1 + 10^{10}) = 10^{10}$ and lose the 1, it is often the case that the accuracy is lost before the catastrophic cancellation occurs.

For an example of catastrophic cancellation in a more realistic computation, consider calculating the roots of the quadratic

$ax^2 + bx + c$ by the standard formula

$$r_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{8}$$

We used this formula to compute the roots of the quadratic $x^2 - 10^4 x + 1$ in IEEE single-precision arithmetic on a Sun Sparcstation. The computed roots were 10^4 and 0, the larger of which is accurate, having a relative error of about 10^{-8}, but the smaller one is completely wrong, the true root being about 10^{-4}. A similar result usually occurs whenever $|ac|/b^2 \ll 1$. The root of larger magnitude is usually computed precisely, but the smaller one is frequently very inaccurate as a result of catastrophic cancellation, since $b^2 - 4ac \approx b^2$ and so $|b| - \sqrt{b^2 - 4ac} \approx 0$. Although the second of these relations signals catastrophic cancellation, the loss of precision occurs in the first.

There is an easy remedy for the loss of precision due to catastrophic cancellation in this case. Use (8) to compute r_1, the root of larger magnitude, and then use the alternative formula $r_2 = c/ar_1$ to compute the smaller one. The relative error in r_2 is at most $(1 + u)^2$ times larger than the relative error in r_1, provided that no overflows or underflows occur in computing c/ar_1.

Another point to note is that, if we compute (6) from right to left, instead of left to right, then

$$fl(1 + (10^{10} - 10^{10})) = fl(1 + 0) = 1.$$

It is not particularly significant that this computation gives the correct answer, but what is important is that it illustrates that floating-point addition is not associative, although it is commutative, since $fl(a + b)$ and $fl(b + a)$ are both required to be the correctly rounded value for $a + b = b + a$. Similar results hold for multiplication and division.

Many other fundamental mathematical relations that we take for granted do not hold for floating-point computations. For example, the result that $\sin(x)$ is strictly increasing for $x \in (0, \pi/2)$ cannot hold in any floating-point system in which x and $\sin(x)$ are in the same precision, since there are more floating-point numbers in the domain $(0, \pi/2)$ than there are in (0,1), the range of $\sin(x)$.

Finally, we end this section by noting that overflows and underflows often cause problems in computations. After an overflow, $\pm\infty$ or NaN frequently propagates through the computation. Although this can sometimes yield a useful result, it is more often a signal of an error in the program or its input. On the other hand, continuing the computation with denormalized numbers or zero in place of an underflow can often yield a useful numerical result. However, there are cases when this can be disastrous. For example, if x^2 underflows, but y^2 is not too close to the underflow limit, then $fl(\sqrt{x^2 + y^2})$, the computed value of $\sqrt{x^2 + y^2}$, is still accurate. However, if both x^2 and y^2 underflow to 0, then $fl(\sqrt{x^2 + y^2}) = 0$, although $\sqrt{x^2 + y^2} \geq \max(|x|, |y|)$ may be far from the underflow limit.

It is often possible to ensure that overflows do not occur and that underflows are harmless. For the example considered above, note that $\sqrt{x^2 + y^2} = s\sqrt{(x/s)^2 + (y/s)^2}$ for any scaling factor $s > 0$. If we choose $s = 2^k$ for an integer $k \approx \log_2[\max(|x|, |y|)]$, then neither $(x/s)^2$ nor $(y/s)^2$ can overflow, and any underflow that occurs is harmless, since one of $(x/s)^2$ or $(y/s)^2$ is close to 1. Moreover, in IEEE floating-point arithmetic, multiplying and dividing by $s = 2^k$ does not introduce any additional rounding error into the computation.

A similar problem with overflows and underflows occurs in formula (8) and in many other numerical computations. Overflows can be avoided and underflows can be rendered harmless in computing $\sqrt{b^2 - 4ac}$ in (8) by scaling in much the same way as described above for $\sqrt{x^2 + y^2}$.

2 The Direct Solution of Linear Algebraic Systems

In this section, we consider the *direct* solution of linear algebraic systems of the form $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$) is a nonsingular matrix and x and $b \in \mathbb{R}^n$ (or $\mathbb{C}^n$). A numerical method for solving $Ax = b$ is direct if it computes the exact solution of the system when implemented in exact arithmetic. Iterative methods for nonsingular linear systems and methods for overdetermined and underdetermined systems are considered in Secs. 4 and 5, respectively.

The standard direct methods for solving $Ax = b$ are based on, or closely related to, *Gaussian elimination* (GE), the familiar variable elimination technique that reduces the original system $Ax = b$ to an *upper-triangular system*, $Ux = \tilde{b}$, which has the same solution x. We present a simple form of GE in Sec. 2.1 and show how $Ux = \tilde{b}$ can be solved easily by *back substitution* in Sec. 2.2. We then explain how the simple form of GE presented in Sec. 2.1 for $Ax = b$ relates to the LU factorization of the coefficient matrix A in Sec. 2.3 and to *forward elimination* in Sec. 2.4. Enhancements to this simple form of GE to make it an efficient, robust, reliable numerical method for the solution of linear systems are outlined in Sec. 2.5. The closely related Cholesky factorization for symmetric positive-definite matrices is presented in Sec. 2.6. We consider how to adapt these methods to banded and sparse linear systems in Sec. 2.7. We end with a discussion of the effects of rounding errors on the direct methods in Sec. 2.8 and of *iterative improvement*, a technique to ameliorate these effects, in Sec. 2.9. See Sec. 13 for a discussion of sources of high-quality numerical software for solving systems of linear algebraic equations.

GE can also be applied to singular systems of linear equations or overdetermined or underdetermined linear systems of m equations in n unknowns. However, it is not as robust as the methods discussed in Sec. 5 for these problems, and so we do not present these generalizations of GE here. In addition, we note that there are several mathematically equivalent, but computationally distinct, implementations of GE and the factorizations discussed here. The reader interested in a more comprehensive treatment of these topics should consult an advanced text, such as Golub and Van Loan (1989).

Finally we note that it is generally inadvisable to solve a system $Ax = b$ by first computing A^{-1} and then calculating $x = A^{-1}b$. The techniques discussed in this section are usually both more reliable and more cost effective than methods using A^{-1}. We also note that, although Cramer's rule is a useful theoretical tool, it is an extremely ineffective computational scheme.

2.1 Gaussian Elimination

First, to unify the notation used below, let $A_0 = A$ and $b_0 = b$.

GE for $Ax = b$ proceeds in $n - 1$ stages. For $k = 1, \ldots, n - 1$, we begin stage k

of GE with the reduced system $A_{k-1}x = b_{k-1}$, where columns $1, \ldots, k-1$ of A_{k-1} contain 0's below the main diagonal. That is, for $A_{k-1} = [a_{ij}^{(k-1)}]$, $a_{ij}^{(k-1)} = 0$ for $j = 1, \ldots, k-1$ and $i = j+1, \ldots, n$. This corresponds to the variables $x_1, \ldots, x_{i-1}$ having been eliminated from equation i of $A_{k-1}x = b_{k-1}$ for $i = 2, \ldots, k-1$ and the variables $x_1, \ldots, x_{k-1}$ having been eliminated from the remaining equations $k, \ldots, n$. Moreover, x is the unique solution of both $A_{k-1}x = b_{k-1}$ and $Ax = b$. Note that all the assumptions above hold vacuously for $k = 1$, since, in this case, the "reduced" system $A_0x = b_0$ is just the original system $Ax = b$ from which no variables have yet been eliminated.

During stage k of GE, we further the reduction process by eliminating the variable x_k from row i of $A_{k-1}x = b_{k-1}$ for $i = k+1, \ldots, n$ by multiplying row k of this system by $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ and subtracting it from row i.

Note that the multipliers m_{ik} are not properly defined by $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ if the *pivot element* $a_{kk}^{(k-1)} = 0$. In exact arithmetic, this can happen only if the $k \times k$ leading principal minor of A is singular. We consider how to deal with zero (or nearly zero) pivots in Sec. 2.5. For now, though, we say that this simple form of GE "breaks down" at stage k and we terminate the process.

After the last stage of GE, the original system $Ax = b$ has been reduced to $Ux = \tilde{b}$, where $\tilde{b} = b_{n-1}$ and $U = A_{n-1}$. Note that $U = [u_{ij}]$ is an upper-triangular matrix (i.e., $u_{ij} = 0$ for $1 \le j < i \le n$) with $u_{kk} = a_{kk}^{(k-1)}$ for $k = 1, \ldots, n$. Therefore, if A and all its leading principal minors are nonsingular, then $u_{kk} = a_{kk}^{(k-1)} \neq 0$ for $k = 1, \ldots, n$, so U is nonsingular too. Moreover, in this case, x is the unique solution of both $Ax = b$ and $Ux = \tilde{b}$. The latter system can be solved easily by back substitution, as described in Sec. 2.2.

The complete GE process can be written in pseudocode as shown in Table 1. Note that at stage k of GE, we know the elements $a_{ik}^{(k)} = a_{ik}^{(k-1)} - m_{ik} \cdot a_{kk}^{(k-1)} = 0$ for $i = k+1, \ldots, n$, and so we do not need to perform this calculation explicitly. Consequently, instead of j running for $k, \ldots, n$ in Table 1, as might be expected, j runs from $k+1, \ldots, n$ instead.

To reduce the storage needed for GE, the original matrix A and vector b are often overwritten by the intermediate matrices A_k and vectors b_k, so that at the end of the GE process, the upper-triangular part of A contains U and b contains $\tilde{b}$. The only change required to the algorithm in Table 1 to implement this reduced storage scheme is to remove the superscripts from the coefficients a_{ij} and b_i. As explained below, it is also important to store the multipliers m_{ik}. Fortunately, the $n-k$ multipliers $\{m_{ik} : i = k+1, \ldots, n\}$ created at stage k of GE can be stored in the $n-k$ positions in column k of A below the main diagonal that are eliminated in stage k. Thus, in many implementations of GE, the upper-triangular part of A is overwritten by U and the strictly lower part of A is

Tab. 1 Gaussian elimination (GE) for the system $Ax = b$

```
for k = 1, ..., n − 1 do
    for i = k + 1, ..., n do
        m_ik = a_ik^(k−1) / a_kk^(k−1)
        for j = k + 1, ..., n do
            a_ij^(k) = a_ij^(k−1) − m_ik · a_kj^(k−1)
        end
        b_i^(k) = b_i^(k−1) − m_ik · b_k^(k−1)
    end
end
```

overwritten by the multipliers m_{ik} for $1 \le k < i \le n$.

A straightforward count of the operations in Table 1 shows that GE requires $n(n-1)/2 \approx n^2/2$ divisions to compute the multipliers m_{ik}, $n(2n-1)(n-1)/6 \approx n^3/3$ multiplications and subtractions to compute the coefficients of U, and $n(n-1)/2 \approx n^2/2$ multiplications and subtractions to compute the coefficients of $\tilde{b}$. Since multiplications and subtractions (or additions) occur in pairs so frequently in matrix calculations, we refer to this pair of operations as a *flop*, which is short for *floating-point operation*. Thus, the computational work required to reduce a system $Ax = b$ of n equations in n unknowns to $Ux = \tilde{b}$ is about $n^3/3$ flops. We show in Sec. 2.2 that $Ux = \tilde{b}$ can be solved by *back substitution* using n divisions and about $n^2/2$ flops.

2.2 Back Substitution

Let $U = [u_{ij}]$ be an $n \times n$ nonsingular upper-triangular matrix. That is, $u_{ij} = 0$ for $1 \le j < i \le n$ and $u_{ii} \ne 0$ for $1 \le i \le n$. Then the linear algebraic system $Ux = \tilde{b}$ can be solved easily by back substitution, as shown in Table 2.

It is easy to see from Table 2 that back substitution requires n divisions and $n(n-1)/2 \approx n^2/2$ multiplications and subtractions. So the computational work is about $n^2/2$ flops.

Tab. 2 Back substitution to solve $Ux = \tilde{b}$

for $i = n, \ldots, 1$ do

$$x_i = \left(\tilde{b}_i - \sum_{j=i+1}^{n} u_{ij}x_j\right) \Big/ u_{ii}$$

end

2.3 The *LU* Factorization

Applying Gaussian elimination (GE) as described in Sec. 2.1 to solve the linear algebraic system $Ax = b$ of n equations in n unknowns is closely related to computing the LU factorization of the matrix A, where $L_1 = [l_{ij}^{(1)}]$ is a unit lower-triangular matrix (i.e., $l_{ii}^{(1)} = 1$ for $i = 1, \ldots, n$ and $l_{ij}^{(1)} = 0$ for $1 \le i < j \le n$) and $U_1 = [u_{ij}^{(1)}]$ is an upper-triangular matrix (i.e., $u_{ij}^{(1)} = 0$ for $1 \le j < i \le n$) satisfying

$$A = L_1 U_1. \tag{9}$$

The factorization (9) exists and is unique if and only if all the leading principal minors of A are nonsingular. In this case, it can be shown that the matrix U_1 in (9) is the same as the upper-triangular matrix U produced by GE, and the elements in the strictly lower-triangular part of $L_1 = [l_{ij}^{(1)}]$ satisfy $l_{ij}^{(1)} = m_{ij}$ for $1 \le j < i \le n$, where the m_{ij} are the multipliers used in GE. Moreover, the U_1 in (9) is nonsingular if A is; L_1 is always nonsingular if the factorization exists.

From the discussion in Sec. 2.1, it follows that computing the LU factorization of an $n \times n$ matrix A in this way requires $n(n-1)/2 \approx n^2/2$ divisions and $n(2n-1)(n-1)/6 \approx n^3/3$ multiplications and subtractions. Thus, the computational work required to calculate it is about $n^3/3$ flops.

If we need to solve $m > 1$ systems $Ax_i = b_i$, $i = 1, \ldots, m$, (or $AX = B$, where X and $B \in \mathbb{R}^{n\times m}$ or $\mathbb{C}^{n\times m}$), we may obtain significant computational savings by computing the LU factorization of A once only and using the factors L_1 and U_1 to solve each system $Ax_i = b_i$ for $i = 1, \ldots, m$ by first solving $L_1\tilde{b}_i = b_i$ for

$\tilde{b}_i$ by *forward elimination*, as described in Sec. 2.4, and then solving $U_1 x_i = \tilde{b}_i$ for x_i by back substitution, as outlined in Sec. 2.2. This procedure is essentially the same as performing GE to reduce A to U once only, saving the multipliers $\{m_{ik}\}$ used in the process, and then, for each system $Ax_i = b_i$, using the multipliers to perform the same transformation on b_i to produce $\tilde{b}_i$ and solving $Ux_i = \tilde{b}_i$ for x_i by back substitution. With either of these procedures, the computational work required to solve all m systems $Ax_i = b_i$ is about $n^3/3 + mn^2$ flops, whereas, if we apply GE as outlined in Sec. 2.1 to each system $Ax_i = b_i$, recomputing U each time, the computational work required to solve all m systems $Ax_i = b_i$ is about $m(n^3/3 + n^2)$ flops, which is much greater if m and/or n is large.

Finally, note that we intentionally used the same symbol $\tilde{b}_i$ for the solution of $L\tilde{b}_i = b_i$ and the transformed right-side vector produced by GE, since these vectors are identical.

2.4 Forward Elimination

Let $L = [l_{ij}]$ be an $n \times n$ lower-triangular matrix (i.e., $l_{ij} = 0$ for $1 \le i < j \le n$). If L is nonsingular too, then $l_{ii} \neq 0$ for $1 \le i \le n$ and so the linear algebraic system $L\tilde{b} = b$ can be solved easily by forward elimination, as shown in Table 3.

It is easy to see from Table 3 that forward elimination requires n divisions and $n(n-1)/2 \approx n^2/2$ multiplications and subtractions. So the computational work is about $n^2/2$ flops.

Tab. 3 Forward elimination too solve $L\tilde{b} = b$

for $i = 1, \ldots, n$ do

$$\tilde{b}_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij}\tilde{b}_j\right) \Big/ l_{ii}$$

end

If $L = [l_{ij}]$ is unit lower-triangular (i.e., $l_{ii} = 1$ for $i = 1, \ldots, n$ as well as L being lower-triangular), as is the case for the L produced by the LU factorization described in Sec. 2.3, then the division by l_{ii} in Table 3 is not required, reducing the operation count slightly to about $n^2/2$ flops. However, we have presented the forward-elimination procedure in Table 3 with general $l_{ii} \neq 0$, since other schemes, such as the Cholesky factorization described in Sec. 2.6, produce a lower-triangular matrix that is typically not unit lower-triangular.

The name "forward elimination" for this procedure comes from the observation that it is mathematically equivalent to the forward-elimination procedure used in GE to eliminate the variables $x_1, \ldots, x_{i-1}$ from equation i of the original system $Ax = b$ to produce the reduced system $Ux = \tilde{b}$.

2.5 Scaling and Pivoting

As noted in Sec. 2.1, the simple form of Gaussian elimination (GE) presented there may "break down" at stage k if the pivot $a_{kk}^{(k-1)} = 0$. Moreover, even if $a_{kk}^{(k-1)} \neq 0$, but $|a_{kk}^{(k-1)}| \ll |a_{ik}^{(k-1)}|$ for some $i \in \{k+1, \ldots, n\}$, then $|m_{ik}| = |a_{ik}^{(k-1)}/a_{kk}^{(k-1)}| \gg 1$. So multiplying row k of A_{k-1} by m_{ik} and subtracting it from row i may produce large elements in the resulting row i of A_k, which in turn may produce still larger elements during later stages of the GE process. Since, as noted in Sec. 2.8, the bound on the rounding errors in the GE process is proportional to the largest element that occurs in A_k for $k = 0, \ldots, n-1$, creating large elements during the GE reduction process may introduce excessive rounding

error into the computation, resulting in an unstable numerical process and destroying the accuracy of the LU factorization and the computed solution x of the linear system $Ax = b$. We present in this section *scaling* and *pivoting* strategies that enhance GE to make it an efficient, robust, reliable numerical method for the solution of linear systems.

Scaling, often called *balancing* or *equilibration*, is the process by which the equations and unknowns of the system $Ax = b$ are scaled in an attempt to reduce the rounding errors incurred in solving the problem and improve its conditioning, as described in Sec. 2.8. The effects can be quite dramatic.

Typically, scaling is done by choosing two $n \times n$ diagonal matrices $D_1 = [d_{ij}^{(1)}]$ and $D_2 = [d_{ij}^{(2)}]$ (i.e., $d_{ij}^{(1)} = d_{ij}^{(2)} = 0$ for $i \neq j$) and forming the new system $\hat{A}\hat{x} = \hat{b}$, where $\hat{A} = D_1 A D_2^{-1}$, $\hat{x} = D_2 x$, and $\hat{b} = D_1 b$. Thus, D_1 scales the rows and D_2 scales the unknowns of $Ax = b$, or, equivalently, D_1 scales the rows and D_2^{-1} scales the columns of A. Of course, the solution of $Ax = b$ can be recovered easily from the solution of $\hat{A}\hat{x} = \hat{b}$, since $x = D_2^{-1}\hat{x}$. Moreover, if the diagonal entries $d_{11}^{(1)}, \ldots, d_{nn}^{(1)}$ of D_1 and $d_{11}^{(2)}, \ldots, d_{nn}^{(2)}$ of D_2 are chosen to be powers of the base of the floating-point number system (i.e., powers of 2 for IEEE floating-point arithmetic), then scaling introduces no rounding errors into the computation.

One common technique is to scale the rows only by taking $D_2 = I$ and choosing D_1 so that largest element in each row of the scaled matrix $\hat{A} = D_1 A$ is about the same size. A slightly more complicated procedure is to scale the rows and columns of A so that the largest element in each row and column of $\hat{A} = D_1 A D_2^{-1}$ is about the same size.

These strategies, although usually helpful, are not foolproof: it is easy to find examples for which row scaling or row and column scaling as described above makes the numerical solution worse. The best strategy is to scale on a problem-by-problem basis depending on what the source problem says about the significance of each coefficient a_{ij} in $A = [a_{ij}]$. See an advanced text such as Golub and Van Loan (1989) for a more detailed discussion of scaling.

For the remainder of this section, we assume that scaling, if done at all, has already been performed.

The most commonly used pivoting strategy is *partial pivoting*. The only modification required to stage k of GE described in Sec. 2.1 to implement GE with partial pivoting is to first search column k of A_{k-1} for the largest element $a_{ik}^{(k-1)}$ on or below the main diagonal. That is, find $i \in \{k, \ldots, n\}$ such that $|a_{ik}^{(k-1)}| \geq |a_{\mu k}^{(k-1)}|$ for $\mu = k, \ldots, n$. Then interchange equations i and k in the reduced system $A_{k-1}x = b_{k-1}$ and proceed with stage k of GE as described in Sec. 2.1.

After the equation interchange described above, the pivot element $a_{kk}^{(k-1)}$ satisfies $|a_{kk}^{(k-1)}| \geq |a_{ik}^{(k-1)}|$ for $i = k, \ldots, n$. So, if $a_{kk}^{(k-1)} \neq 0$, then the multiplier $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ must satisfy $|m_{ik}| \leq 1$ for $i = k+1, \ldots, n$. Thus no large multipliers can occur in GE with partial pivoting.

On the other hand, if the pivot element $a_{kk}^{(k-1)} = 0$, then $a_{ik}^{(k-1)} = 0$ for $i = k, \ldots, n$, whence A_{k-1} is singular and so A must be too. Thus, GE with partial pivoting never "breaks down" (in exact arithmetic) if A is nonsingular.

Partial pivoting adds a little overhead only to the GE process. At stage k, we must perform $n - k$ comparisons to

determine the row i with $|a_{ik}^{(k-1)}| \geq |a_{jk}^{(k-1)}|$ for $j = k, \ldots, n$. Thus, GE with partial pivoting requires a total of $n(n-1)/2 \approx n^2/2$ comparisons. In addition, we must interchange rows i and k if $i > k$, or use some form of indirect addressing if the interchange is not performed explicitly. On the other hand, exactly the same number of arithmetic operations must be executed whether or not pivoting is performed. Therefore, if n is large, the added cost of pivoting is small compared with performing approximately $n^3/3$ flops to reduce A to upper-triangular form.

Complete pivoting is similar to partial pivoting except that the search for the pivot at stage k of GE is not restricted to column k of A_{k-1}. Instead, in GE with complete pivoting, we search the $(n-k) \times (n-k)$ lower right block of A_{k-1} for the largest element. That is, find i and $j \in \{k, \ldots, n\}$ such that $|a_{ij}^{(k-1)}| \geq |a_{\mu\nu}^{(k-1)}|$ for $\mu = k, \ldots, n$ and $\nu = k, \ldots, n$. Then interchange equations i and k and variables j and k in the reduced system $A_{k-1}x_{k-1} = b_{k-1}$ and proceed with stage k of GE as described in Sec. 2.1. Note that the vector x_{k-1} in the reduced system above is a reordered version of the vector of unknowns x in the original system $Ax = b$, incorporating the variable interchanges that have occurred in stages $1, \ldots, k-1$ of GE with complete pivoting.

After the equation and variable interchanges described above, the pivot element $a_{kk}^{(k-1)}$ satisfies $|a_{kk}^{(k-1)}| \geq |a_{ik}^{(k-1)}|$ for $i = k, \ldots, n$. So, if $a_{kk}^{(k-1)} \neq 0$, the multiplier $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ must satisfy $|m_{ik}| \leq 1$ for $i = k+1, \ldots, n$, as is the case with partial pivoting. However, with complete pivoting, the multipliers tend to be even smaller than they are with partial pivoting, since the pivots tend to be larger, and so the numerical solution might suffer less loss of accuracy due to rounding errors, as discussed further in Sec. 2.8.

After the row and column interchanges in stage k, the pivot element $a_{kk}^{(k-1)} = 0$ only if $a_{ij}^{(k-1)} = 0$ for $i = k, \ldots, n$ and $j = k, \ldots, n$ in which case A_{k-1} is singular and so A must be too. Thus, like GE with partial pivoting, GE with complete pivoting never "breaks down" (in exact arithmetic) if A is nonsingular.

Moreover, if A is singular and $a_{ij}^{(k-1)} = 0$ for $i = k, \ldots, n$ and $j = k, \ldots, n$, then the GE process can be terminated at this stage and the factorization computed so far used to advantage in determining a solution (or approximate solution) to the singular system $Ax = b$. However, this is not as robust a technique as the methods discussed in Sec. 4 for overdetermined problems, and so we do not discuss this further here. The reader interested in this application of GE should consult an advanced text such as Golub and Van Loan (1989).

Complete pivoting, unlike partial pivoting, adds significantly to the cost of the GE process. At stage k, we must perform $(n-k+1)^2 - 1$ comparisons to determine the row i and column j with $|a_{ij}^{(k-1)}| \geq |a_{\mu\nu}^{(k-1)}|$ for $\mu = k, \ldots, n$ and $\nu = k, \ldots, n$. Thus, GE with partial pivoting requires a total of $n(n-1)(2n+5)/6 \approx n^3/3$ comparisons. On the other hand, exactly the same number of arithmetic operations must be executed whether or not pivoting is performed. So the cost of determining the pivots is comparable to the cost of performing approximately $n^3/3$ flops required to reduce A to upper-triangular form. In addition, we must interchange rows i and k if $i > k$ and columns j and k if $j > k$ or use some form of indirect addressing if the interchange is not performed explicitly. Thus even though GE with complete

pivoting has better roundoff-error properties than GE with partial pivoting, GE with partial pivoting is used more often in practice.

As is the case for the simple version of GE presented in Sec. 2.1, GE with partial or complete pivoting is closely related to computing the LU factorization of the matrix A. However, in this case, we must account for the row or row and column interchanges by extending (9) to

$$P_2 A = L_2 U_2 \tag{10}$$

for partial pivoting and

$$P_3 A Q_3^{\mathrm{T}} = L_3 U_3 \tag{11}$$

for complete pivoting, where P_2 and P_3 are permutation matrices that record the row interchanges performed in GE with partial and complete pivoting, respectively; Q_3^{T} is a permutation matrix that records the column interchanges performed in GE with complete pivoting; L_2 and L_3 are unit lower-triangular matrices with the $n-k$ multipliers from stage k of GE with partial and complete pivoting, respectively, in column k below the main diagonal, but permuted according to the row interchanges that occur in stages $k+1, \ldots, n-1$ of GE with partial and complete pivoting, respectively; and U_2 and U_3 are the upper-triangular matrices produced by GE with partial and complete pivoting, respectively.

A permutation matrix P has exactly one 1 in each row and column and all other elements equal to 0. It is easy to check that $PP^{\mathrm{T}} = I$ so that P is nonsingular and $P^{\mathrm{T}} = P^{-1}$. That is, P is an orthogonal matrix. Also note that we do not need a full $n \times n$ array to store P: The information required to form or multiply by $P = [P_{ij}]$ can be stored in an n-vector $p = [p_i]$, where $p_i = j$ if and only if $P_{ij} = 1$ and $P_{ik} = 0$ for $k \neq j$.

The factorizations (9), (10), and (11) are all called *LU factorizations*; (10) is also called a *PLU factorization*. Unlike (9), the LU factorizations (10) and (11) always exist. P_2, P_3, Q_3, L_2, and L_3 are always nonsingular; U_2 and U_3 are nonsingular if and only if A is nonsingular. Moreover, the factorizations (10) and (11) are unique if there is a well-defined choice for the pivot if more than one element of maximal size occurs in the search for the pivot and if there is a well-defined choice for the multipliers in L if the pivot is zero.

The LU factorization (10) can be used to solve the linear system $Ax = b$ by first computing $\hat{b}_2 = P_2 b$, then solving $L_2 \tilde{b}_2 = \tilde{b}_2$ by forward elimination, and finally solving $U_2 x = \tilde{b}_2$ by back substitution. The steps are similar if we use the LU factorization (11) instead of (10), except that the back substitution $U_3 \hat{x}_3 = \tilde{b}_3$ yields the permuted vector of unknowns $\hat{x}_3$. The original vector of unknowns x can be recovered by $x = Q_3^{\mathrm{T}} \hat{x}_3$. We have used $\tilde{b}_2$ and $\tilde{b}_3$ for the intermediate results above to emphasize that this is the same as the vector $\tilde{b}$ that is obtained if we perform GE with partial and complete pivoting, respectively, on the original system $Ax = b$.

As noted earlier for (9), if we need to solve $m > 1$ systems $Ax_i = b_i$, $i = 1, \ldots, m$ (or $AX = B$, where $X, B \in \mathbb{R}^{n \times m}$ or $\mathbb{C}^{n \times m}$), we may obtain significant computational savings by computing the LU factorization of A once only. The same observation applies to the LU factorizations (10) and (11).

We end by noting that not having to pivot to ensure numerical stability can be a great advantage in some cases – for example, when factoring a banded or sparse matrix,

as described in Sec. 2.7. Moreover, there are classes of matrices for which pivoting is not required to ensure numerical stability. Three such classes are complex Hermitian positive-definite matrices, real symmetric positive-definite matrices, and diagonally dominant matrices.

2.6 The Cholesky Factorization

In this subsection, we present the *Cholesky factorization* of a real symmetric positive-definite $n \times n$ matrix A. It is straightforward to modify the scheme for complex Hermitian positive-definite matrices.

Recall that $A \in \mathbb{R}^{n\times n}$ is symmetric if $A = A^{\mathrm{T}}$, where A^{T} is the transpose of A, and it is positive definite if $x^{\mathrm{T}}Ax > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$. The Cholesky factorization exploits these properties of A to compute a lower-triangular matrix L satisfying

$$A = LL^{\mathrm{T}}. \tag{12}$$

The similar *LDL factorization* computes a unit lower-triangular matrix $\tilde{L}$ and a diagonal matrix D satisfying

$$A = \tilde{L}D\tilde{L}^{\mathrm{T}}. \tag{13}$$

We present the *dot product* form of the Cholesky factorization in Table 4. It is derived by equating the terms of $A = [a_{ij}]$ to those of LL^{T} in the order $(1, 1), (2, 1), \ldots, (n, 1), (2, 2), (3, 2), \ldots, (n, 2), \ldots, (n, n)$ and using the lower-triangular structure of $L = [l_{ij}]$ (i.e., $l_{ij} = 0$ for $1 \le i < j \le n$). Other forms of the Cholesky factorization are discussed in advanced texts such as Golub and Van Loan (1989).

It can be shown that, if A is symmetric positive-definite, then

$$a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 > 0$$

Tab. 4 The Cholesky factorization of a symmetric positive-definite matrix A

for $j = 1, \ldots, n$ do
  $l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$
  for $i = j + 1, \ldots, n$ do
    $l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}\right) \Big/ l_{jj}$
  end
end

(in exact arithmetic) each time this expression is computed in the Cholesky factorization. Therefore, we may take the associated square root to be positive, whence $l_{jj} > 0$ for $j = 1, \ldots, n$. With this convention, the Cholesky factorization is unique.

Moreover, it follows from

$$l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\right)^{1/2}$$

that

$$a_{jj} = \sum_{k=1}^{j} l_{jk}^2.$$

So the elements in row j of the Cholesky factor L are bounded by $\sqrt{a_{jj}}$ even if we do not pivot. Consequently, as noted in Sec. 2.5, it is customary to compute the Cholesky factorization without pivoting.

Note that the method in Table 4 accesses the lower-triangular part of A only, and so only those elements need to be stored. Moreover, if we replace l_{jj} and l_{ij} by a_{jj} and a_{ij}, respectively, in Table 4, then the modified algorithm overwrites the lower-triangular part of A with the Cholesky factor L.

A straightforward count of the operations in Table 4 shows that the Cholesky factorization requires n square roots, $n(n-1)/2 \approx n^2/2$ divisions, and $n(n-$

$1)(n+1)/6 \approx n^3/6$ multiplications and subtractions. This is approximately half the arithmetic operations required to compute the LU factorization of A. Of course, the storage required for the Cholesky factorization is also about half that required for the LU factorization.

The Cholesky, LDL, and LU factorizations are closely related. Let $D_1 = [d_{ij}^{(1)}]$ be the diagonal matrix with the same diagonal elements as $L = [l_{ij}]$ (i.e., $d_{jj}^{(1)} = l_{jj}$ for $j = 1, \ldots, n$ and $d_{ij}^{(1)} = 0$ for $i \neq j$). D_1 is nonsingular, since, as noted above, $l_{jj} > 0$ for $j = 1, \ldots, n$. Moreover, $\tilde{L} = LD_1^{-1}$ is unit lower-triangular. If we also let $D = D_1 D_1 = D_1 D_1^{\mathrm{T}}$, then

$$\begin{aligned} A &= LL^{\mathrm{T}} = (\tilde{L}D_1)(\tilde{L}D_1)^{\mathrm{T}} \\ &= \tilde{L}D_1 D_1^{\mathrm{T}} \tilde{L}^{\mathrm{T}} = \tilde{L}D\tilde{L}^{\mathrm{T}}, \end{aligned}$$

where $\tilde{L}D\tilde{L}^{\mathrm{T}}$ is the LDL factorization of A. Furthermore, if we let $U = D\tilde{L}^{\mathrm{T}}$, then $\tilde{L}U$ is the LU factorization of A.

The LDL factorization can be computed directly by equating the terms of A to those of $\tilde{L}D\tilde{L}^{\mathrm{T}}$, just as we did above for the Cholesky factorization. This leads to a scheme similar to that shown in Table 4, but without any square roots, although it has $n(n-1)/2 \approx n^2/2$ more multiplications. Thus the cost of computing the factorization remains about $n^3/6$ flops and the storage requirement remains about $n^2/2$.

An advantage of the LDL factorization is that it can be applied to a symmetric indefinite matrix. The Cholesky factorization is not applicable in this case, since LL^{T} is always symmetric positive-semidefinite. However, since LDL^{T} is always symmetric, pivoting must be restricted to ensure that the reordered matrix PAQ^{T} is symmetric. The simplest way to maintain symmetry is to use *symmetric pivoting* in which $Q = P$. This, though, restricts the choice of the pivot at stage k of the LDL factorization to a_{jj} for $j = k, \ldots, n$. As a result, in some cases, the LDL factorization may incur much more rounding error than GE with partial or complete pivoting.

2.7 Banded and Sparse Matrices

Significant savings in both computational work and storage can often be obtained in solving $Ax = b$ by taking advantage of zeros in the coefficient matrix A. We outline in this subsection how these savings may be realized for banded and more general sparse matrices.

To begin, note that an $n \times n$ matrix A is said to be *sparse* if the number of nonzero elements in A is much less than n^2, the total number of elements in A. *Banded matrices* are an important subclass of sparse matrices in which the nonzero elements of the matrix are restricted to a band around the main diagonal of the matrix. The *lower bandwidth* of a matrix $A = [a_{ij}]$ is the smallest integer p such that $a_{ij} = 0$ for $i - j > p$, the *upper bandwidth* of A is the smallest integer q such that $a_{ij} = 0$ for $j - i > q$, and the *bandwidth* of A is $1 + p + q$. Clearly, if p and $q \ll n$, then A is sparse, since A has at most $(1 + p + q)n \ll n^2$ nonzero elements.

Banded and more general sparse matrices arise in many important applications. For example, quadratic spline interpolation, as described in Sec. 8.6.2, requires the solution of a linear system $T_c = g$, where T is a tridiagonal matrix (i.e., banded with $p = q = 1$) and c is the vector of coefficients for the quadratic spline interpolant. Banded matrices also arise in the solution of boundary-value problems for ordinary differential equations. See Sec. 11.3.1 for an example of a system $T_u = g$, where T

is a symmetric positive-definite tridiagonal matrix. More general sparse matrices arise in the numerical solution of partial differential equations. See Sec. 11.3.2 for an example of the matrix associated with the standard five-point difference scheme for Poisson's equation.

If A is large and sparse, it is common to store only the nonzero elements of A, since this greatly reduces the storage requirements. This is easy to do if A is banded, since we can map the elements from the band of A to a $(1+p+q)\times n$ array representing A in a *packed format*, several of which are commonly used in practice. If $A=[a_{ij}]$ is a general sparse matrix, then we require a more general *sparse-matrix data structure* that stores each nonzero element a_{ij} of A along with some information used to recover the indices i and j.

We consider Gaussian elimination (GE) for a banded matrix first. To begin, note that the reduction process described in Sec. 2.1 maintains the band structure of A. In particular, $a_{ik}^{(k-1)}=0$ for $i-k>p$, whence the multipliers $m_{ik}=a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ in Table 1 need to be calculated for $i=k+1,\ldots\min(k+p,n)$ only and the i loop can be changed accordingly. Similarly $a_{kj}^{(k-1)}=0$ for $j-k>q$, and so the reduction $a_{ij}^{(k)}=a_{ij}^{(k-1)}-m_{ij}a_{kj}^{(k-1)}$ in Table 1 needs to be calculated for $j=k+1,\ldots,\min(k+q,n)$ only and the j loop can be charged accordingly. It therefore follows from a straightforward operation count that GE modified as described above for banded matrices requires $np-p(p+1)/2 \approx np$ divisions to compute the multipliers m_{ik}, either $npq-p(3q^2+3q+p^2-1)/6 \approx npq$ (if $p\le q$) or $npq-q(3p^2+3p+q^2-1)/6 \approx npq$ (if $p\ge q$) multiplications and subtractions to compute the coefficients of U, and $np-p(p+1)2 \approx np$ multiplications and subtractions to compute the coefficients of $\tilde{b}$. Furthermore, note that, if we use this modified GE procedure to compute the LU factorization of A, then the lower-triangular matrix L has lower bandwidth p and the upper-triangular matrix U has upper bandwidth q. As noted in Sec. 2.1, it is common to overwrite A with L and U. This can be done even if A is stored in packed format, thereby achieving significant reduction in storage requirements.

The back-substitution method shown in Table 2 and the forward-elimination method shown in Table 3 can be modified similarly so that each requires n divisions, the back-substitution method requires $nq-q(q+1)/2 \approx nq$ multiplications and subtractions, while the forward-elimination method requires $np-p(p+1)/2\approx np$ multiplications and subtractions. In addition, recall that the n divisions are not needed in forward elimination if L is unit lower-triangular, as is the case for the modified LU factorization described here.

A similar modification of the Cholesky method shown in Table 4 results in a procedure that requires n square roots, $np-p(p+1)/2\approx np$ divisions, and $(n-p)p(p+1)/2+(p-1)p(p+1)/6\approx np^2/2$ multiplications and subtractions. In deriving these operation counts, we used $p=q$, since the matrix A must be symmetric for the Cholesky factorization to be applicable. Moreover, the Cholesky factor L has lower bandwidth p. Thus, as for the general case, the Cholesky factorization of a band matrix requires about half as many arithmetic operations and about half as much storage as the LU factorization, since packed storage can also be used for the Cholesky factor L.

If partial pivoting is used in the LU factorization of A, then the upper bandwidth of U may increase to $p+q$. The associated matrix L is a permuted version of a lower-triangular matrix with lower bandwidth p. Both factors L and U can be stored in packed format in a $(1+2p+q)\times n$ array. The operation count for the LU factorization, forward elimination, and back solution is the same as though A were a banded matrix with upper bandwidth $p+q$ and lower bandwidth p. Thus, if $p>q$, both computational work and storage can be saved by factoring A^{T} instead of A and using the LU factors of A^{T} to solve $Ax=b$. If complete pivoting is used in the LU factorization of A, then L and U may fill in so much that there is little advantage to using a band solver. Consequently, it is advantageous not to pivot when factoring a band matrix, provided this does not lead to an unacceptable growth in rounding errors. As noted in Sec. 2.5, it is not necessary to pivot for numerical stability if A is complex Hermitian positive definite, real symmetric positive definite, or column-diagonally dominant. If pivoting is required, it is advantageous to use partial, rather than complete, pivoting, again provided this does not lead to an unacceptable growth in rounding errors.

Extending GE to take advantage of the zeros in a general sparse matrix is considerably more complicated than for banded matrices. The difficulty is that, when row k of A_{k-1} is multiplied by m_{ik} and added to row i of A_{k-1} to eliminate $a_{ik}^{(k-1)}$ in stage k of GE, as described in Sec. 2.1, some zero elements in row i of A_{k-1} may become nonzero in the resulting row i of A_k. These elements are said to *fill in* and are collectively referred to as *fill-in* or *fill*.

However, pivoting can often greatly reduce the amount of fill-in. To see how this comes about, the interested reader may wish to work through an example with the *arrow-head* matrix

$$A=\begin{pmatrix} 5 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since A is a real symmetric positive-definite matrix, there is no need to pivot for numerical stability. It is easy to see that the Cholesky and LU factors of A completely fill in. Interchanging the first and last rows and the first and last columns of A corresponds to forming the permuted matrix $B=PAP^{\mathrm{T}}$, where $P=I-[1,0,0,0,-1]^{\mathrm{T}}[1,0,0,0,-1]$ is a permutation matrix. Since B is also a real symmetric positive-definite matrix, there is no need to pivot for numerical stability when factoring B. However, in this case, the Cholesky and LU factors of B suffer no fill-in at all. This small example can be generalized easily to arbitrarily large arrowhead matrices having the similar properties that the LU factors of A completely fill in while those of the permuted matrix $B=PAP^{\mathrm{T}}$ suffer no fill-in at all.

If we use an appropriate sparse-matrix data structure to store only the nonzeros elements of a sparse matrix A and its LU factors, then reducing the fill-in reduces both the storage needed for the LU factors and the computational work required to calculate them. It also reduces the computational work required for forward elimination and back substitution, since the methods shown in Tables 2 and 3 can be modified easily so that they use the nonzero elements in L and U only, thereby avoiding multiplications by zero and the associated subtractions.

Therefore, the goal in a sparse LU or Cholesky factorization is to reorder the

rows and columns of A to reduce the amount of fill-in. If we need to pivot to ensure numerical stability, then this might conflict with pivoting to reduce fill-in. Unfortunately, even without this complication, finding the optimal reordering to minimize the fill-in is computationally too expensive to be feasible in general. There are, though, many good *heuristic* methods to reorder the rows and columns of A that greatly reduce the fill-in and computational work in many important cases. The reader seeking a more complete description of sparse-matrix factorizations should consult a text on this topic, such as Duff et al. (1986) or George and Liu (1981).

We end this subsection with an example illustrating the importance of exploiting the zeros in a sparse matrix A. Consider the linear system derived in Sec. 11.3.2 by discretizing Poisson's equation on an $m \times m$ grid. A is an $n \times n$ symmetric positive-definite matrix with $n = m^2$ and $5m^2 - 4m \approx 5m^2 = 5n$ nonzero elements, out of a total of n^2 elements in A. So, if m is large, A is very sparse. Moreover, if we use the *natural ordering* for the equations and variables in the system, then A, as shown in Sec. 11.3.2, is a banded matrix with lower and upper bandwidth $m = \sqrt{n}$.

The computational work and storage required to solve this linear system are shown in Table 5. We consider three cases:

Tab. 5 Computational work and storage required to solve the linear system $Ax = b$ derived from discretizing Poisson's equation on an $m \times m$ grid

	Factor	*Solve*	*Store A*	*Store L*
Dense	$m^6/6$	m^4	$m^4/2$	$m^4/2$
Banded	$m^4/2$	$2m^3$	m^3	m^3
Sparse	$O(m^3)$	$O(m^2 \log m)$	$5m^2$	$O(m^2 \log m)$

1. dense, the zeros in A are not exploited at all;
2. banded, we use a band solver that exploits the band structure of A;
3. sparse, we use a general sparse solver together with the *nested dissection ordering* (see George and Liu, 1981) for the equations and variables in the system.

The columns labeled factor and solve, respectively, give the approximate number of flops needed to compute the Cholesky factorization of the matrix A and to solve the linear system given the factors. The columns labeled store A and store L, respectively, give the approximate number of storage locations needed to store A and its Cholesky factor L.

2.8 Rounding Errors, Condition Numbers, and Error Bounds

In this subsection, we consider the effects of rounding errors in solving $Ax = b$. In doing so, we use vector and matrix norms extensively. Therefore, we recommend that, if you are not acquainted with norms, you familiarize yourself with this topic before reading this subsection. Most introductory numerical methods texts or advanced books on numerical linear algebra contain a section on vector and matrix norms.

The analysis of the effects of rounding errors in solving $Ax = b$ usually proceeds in two stages. First we establish that the computed solution $\tilde{x}$ is the exact solution of a perturbed system

$$(A + E)\tilde{x} = b + r \tag{14}$$

with bounds on the size of E and r. Then we use (14) together with bounds on the size of A and A^{-1} to bound the error $x - \tilde{x}$.

The first step is called a *backward error analysis*, since it casts the error in the solution back onto the problem and allows us to relate the effects of rounding errors in the computed solution $\tilde{x}$ to other errors in the problem, such as measurement errors in determining the coefficients of A and b. Another advantage of proceeding in this two-stage fashion is that, if $\tilde{x}$ is not sufficiently accurate, it allows us to determine whether this is because the numerical method is faulty or whether the problem itself is unstable.

Throughout this section we assume that A is an $n \times n$ nonsingular matrix and $nu < 0.1$, where u is the *relative roundoff error bound* for the machine arithmetic used in the computation (see Sec. 1.2).

If we use Gaussian elimination (GE) with either partial or complete pivoting together with forward elimination and back substitution to solve $Ax = b$, then it can be shown that the computed solution $\tilde{x}$ satisfies (14) with $r = 0$ and

$$\|E\|_\infty \leq 8n^3\gamma\|A\|_\infty u + O(u^2), \qquad (15)$$

where

$$\gamma = \max_{i,j,k} \frac{|a_{ij}^{(k)}|}{\|A\|_\infty}$$

is the *growth factor* and $A_k = [a_{ij}^{(k)}]$ for $k = 0, \ldots, n-1$ are the intermediate reduced matrices produced during the GE process (see Secs. 2.1 and 2.5). It can be shown that $\gamma \leq 2^{n-1}$ for GE with partial pivoting and $\gamma \leq [n(2 \times 3^{1/2} \times 4^{1/3} \times \cdots n^{1/(n-1)})]^{1/2} \ll 2^{n-1}$ for GE with complete pivoting. Moreover, the former upper bound can be achieved. However, for GE with both partial and complete pivoting, the actual error incurred is usually much smaller than the bound (15) suggests: it is typically the case that $\|E\|_\infty \propto \|A\|_\infty u$. Thus, GE with partial pivoting usually produces a computed solution with a small backward error, but, unlike GE with complete pivoting, there is no guarantee that this will be the case.

If A is column-diagonally dominant, then applying GE without pivoting to solve $Ax = b$ is effectively the same as applying GE with partial pivoting to solve this system. Therefore, all the remarks above for GE with partial pivoting apply in this special case.

If we use the Cholesky factorization without pivoting together with forward elimination and back substitution to solve $Ax = b$, where A is a symmetric positive-definite matrix, then it can be shown that the computed solution $\tilde{x}$ satisfies (14) with $r = 0$ and

$$\|E\|_2 \leq c_n\|A\|_2 u, \qquad (16)$$

where c_n is a constant of moderate size that depends on n only. Thus the Cholesky factorization without pivoting produces a solution with a small backward error in all cases.

Similar bounds on the backward error for other factorizations and matrices with special properties, such as symmetric or band matrices, can be found in advanced texts, such as Golub and Van Loan (1989).

It is also worth noting that we can easily compute an *a posteriori* backward error estimate of the form (14) with $E = 0$ by calculating the *residual* $r = A\tilde{x} - b$ after computing $\tilde{x}$. Typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$.

Now we use (14) together with bounds on the size of A and A^{-1} to bound the error $x - \tilde{x}$. To do so, we first introduce the *condition number* $\kappa(A) = \|A\| \ \|A^{-1}\|$ associated with the problem $Ax = b$. Although $\kappa(A)$ clearly depends on the

matrix norm used, it is roughly of the same magnitude for all the commonly used norms, and it is the magnitude only of $\kappa(A)$ that is important here. Moreover, for the result below to hold, we require only that the matrix norm associated with $\kappa(A)$ is sub-multiplicative (i.e., $\|AB\| \leq \|A\| \, \|B\|$ for all A and $B \in \mathbb{C}^{n\times n}$) and that it is consistent with the vector norm used (i.e., $\|Av\| \leq \|A\| \, \|v\|$ for all $A \in \mathbb{C}^{n\times n}$ and $v \in \mathbb{C}^n$). It follows immediately from these two properties that $\kappa(A) \geq 1$ for all $A \in \mathbb{C}^{n\times n}$. More importantly, it can be shown that, if $\|E\|/\|A\| \leq \delta$, $\|r\|/\|b\| \leq \delta$, and $\delta\kappa(A) = r < 1$, then $A + E$ is nonsingular and

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{2\delta}{1 - r}\kappa(A). \qquad (17)$$

Moreover, for any given A, there are some b, E, and r for which $\|x - \tilde{x}\|/\|x\|$ is as large as the right side of (17) suggests it might be, although this is not the case for all b, E, and r. Thus we see that, if $\kappa(A)$ is not too large, then small relative errors $\|E\|/\|A\|$ and $\|r\|/\|b\|$ ensure a small relative error $\|x - \tilde{x}\|/\|x\|$, and so the problem is *well-conditioned*. On the other hand, if $\kappa(A)$ is large, then $\|x - \tilde{x}\|/\|x\|$ might be large even though $\|E\|/\|A\|$ and $\|r\|/\|b\|$ are small, and so the problem is *ill-conditioned*. Thus, as the name suggests, the condition number $\kappa(A)$ gives a good measure of the conditioning – or stability – of the problem $Ax = b$.

Combining the discussion above with the earlier observation that typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$, we get the general rule of thumb that, if $u \approx 10^{-d}$ and $\kappa_\infty = \|A\|_\infty \|A^{-1}\|_\infty \approx 10^q$, then $\tilde{x}$ contains about $d - q$ correct digits.

Many routines for solving linear systems provide an estimate of $\kappa(A)$, although most do not compute $\|A\| \, \|A^{-1}\|$ directly, since they do not compute A^{-1}.

There are many other useful inequalities of the form (17). The interested reader should consult an advanced text, such as Golub and Van Loan (1989).

2.9 Iterative Improvement

The basis of *iterative improvement* is the observation that, if x_1 is an approximate solution to $Ax = b$, we can form the residual $r_1 = b - Ax_1$, which satisfies $r_1 = A(x - x_1)$, and then solve $Ad_1 = r_1$ for the difference $d_1 = x - x_1$ and finally compute the improved solution $x_2 = x_1 + d_1$. In exact arithmetic, $x_2 = x$, but, in floating-point arithmetic, $x_2 \neq x$ normally. So we can repeat the process using x_2 in place of x_1 to form another improved solution x_3, and so on. Moreover, if we have factored A to compute x_1, then, as noted in Secs. 2.3 and 2.5, there is relatively little extra computational work required to compute a few iterations of iterative improvement.

The catch here is that, as noted in Sec. 2.8, typically $\|r\|_\infty \propto \|b\|_\infty u$ if GE with partial or complete pivoting or the Cholesky factorization is used to compute $\tilde{x}$. So, if we compute $r_1 = b - Ax_1$ in the same precision, then r_1 will contain few if any correct digits. Consequently, using it in iterative improvement usually does not lead to a reduction of the error in x_2, although it may lead to a smaller E in (14) in some cases. However, if we compute $r_k = b - Ax_k$ in double precision for $k = 1, 2, \ldots$, then iterative improvement may be quite effective. Roughly speaking, if the relative roundoff error bound $u \approx 10^{-d}$ and the condition number $\kappa(A) = \|A\| \, \|A^{-1}\| \approx 10^q$, then after k iterations of iterative improvement the computed solution x_k typically has

about $\min(d, k(d-q))$ correct digits. Thus, if $\kappa(A)$ is large, but not too large, and, as a result, the initial computed solution x_1 is inaccurate, but not completely wrong, then iterative improvement can be used to obtain almost full single-precision accuracy in the solution.

The discussion above can be made more rigorous by noting that iterative improvement is a basic iterative method of the form (19)–(20) for solving $Ax = b$ and applying the analysis in Sec. 3.1.

3 The Iterative Solution of Linear Algebraic Systems

In this section, we consider *iterative* methods for the numerical solution of linear algebraic systems of the form $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$) is a nonsingular matrix, x, and $b \in \mathbb{R}^n$ (or $\mathbb{C}^n$). Such schemes compute a sequence of approximations $x_1, x_2, \ldots$ to x in the hope that $x_k \to x$ as $k \to \infty$. Direct methods for solving $Ax = b$ are considered in Sec. 2.

Iterative methods are most frequently used when A is large and sparse, but not banded with a small bandwidth. Such matrices arise frequently in the numerical solution of partial differential equations (PDEs). See, for example, the matrix shown in Sec. 11.3.2 that is associated with the standard five-point difference scheme for Poisson's equation. In many such cases, iterative methods are more efficient than direct methods for solving $Ax = b$: They usually use far less storage and often require significantly less computational work as well.

We discuss *basic iterative methods* in Sec. 3.1 and the *conjugate-gradient acceleration* of these schemes in Sec. 3.2. A more complete description and analysis of these methods and other iterative schemes are provided in Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981), Young (1971), and Varga (1962). See Sec. 13 for a discussion of sources of high-quality numerical software for solving systems of linear algebraic equations.

We discuss multigrid methods in Sec. 11.8, because these iterative schemes are so closely tied to the PDE that gives rise to the linear system $Ax = b$ to which they are applied.

3.1 Basic Iterative Methods

Many iterative methods for solving $Ax = b$ are based on splitting the matrix A into two parts, M and N, such that $A = M - N$ with M nonsingular. M is frequently called the *splitting matrix*. Starting from an initial guess x_0 for x, we compute $x_1, x_2, \ldots$ recursively from

$$Mx_{k+1} = Nx_k + b. \tag{18}$$

We call such a scheme a *basic iterative method*, but it is often also referred to as a *linear stationary method of the first degree*.

Since $N = M - A$, (18) can be rewritten as

$$Mx_{k+1} = (M - A)x_k + b = Mx_k + (b - Ax_k),$$

which is equivalent to

$$Md_k = r_k, \tag{19}$$

$$x_{k+1} = x_k + d_k, \tag{20}$$

where $r_k = b - Ax_k$ is the residual at iteration k. Although (18) and (19)–(20) are mathematically equivalent, it might be computationally more effective to implement a method in one form than the other.

Clearly, for either (18) or (19)–(20) to be effective,

1. it must be much easier to solve systems with M than with A, and
2. the iterates $x_1, x_2, \ldots$ generated by (18) or (19)–(20) must converge quickly to x, the solution of $Ax = b$.

To address point **2**, first note that, since $Ax = b$ and $A = M - N$, $Mx = Nx + b$. So, if the sequence $x_1, x_2, \ldots$ converges, it must converge to x. To determine whether the sequence $x_1, x_2, \ldots$ converges and, if so, how fast, subtract (18) from $Mx = Nx + b$ and note that the error $e_k = x - x_k$ satisfies the recurrence $Me_{k+1} = Ne_k$, or equivalently $e_{k+1} = Ge_k$, where

$$G = M^{-1}N = I - M^{-1}A$$

is the associated *iteration matrix*. So

$$e^k = G^k e_0. \tag{21}$$

Using (21), we can show that, starting from any initial guess x_0, the sequence $x_1, x_2, \ldots$ generated by (18) converges to x if and only if $\rho(G) < 1$, where

$$\rho(G) = \max\{|\lambda| : \lambda \text{ an eigenvalue of } G\}$$

is the *spectral radius* of G. Moreover, $\rho(G)$ is the "asymptotically average" amount by which the error e_k decreases at each iteration. Consequently, $(\log \in)/\log \rho(G)$ is a rough estimate of the number of iterations of (18) required to reduce the initial error e_0 by a factor $\in$. Thus, it is common to define

$$R(G) = -\log \rho(G) \tag{22}$$

to be the *rate of convergence* (sometimes called the *asymptotic rate of convergence* or the *asymptotic average rate of convergence*) of the iteration (18).

One useful general result is that, if $A = M - N$ is Hermitian positive-definite and if the Hermitian matrix $M^{\mathrm{H}} + N$ is positive-definite too, then $\rho(G) < 1$ and the associated iteration (18) converges.

Possibly the simplest iterative scheme is the *RF method* (a variant of Richardson's method) for which $M = I$ and $N = I - A$, whence (18) reduces to

$$x_{k+1} = x_k + r_k,$$

where $r_k = b - Ax_k$ is the residual at iteration k. From the general discussion above, it follows that this scheme converges if and only if $\rho(I - A) < 1$. Because of this severe constraint on convergence, this scheme is not often effective in its own right, but it can be used productively as the basis for polynomial acceleration, as discussed in Sec. 3.2.

We describe the Jacobi, Gauss–Seidel, SOR, and SSOR methods next, and then consider their convergence. In describing them, we use the notation $A = D - L - U$, where D is assumed to be nonsingular and consists of the diagonal elements of A for the point variant of each method or the diagonal submatrices of A for the block variant. L and U are the negatives of the strictly lower- and upper-triangular parts of A, respectively, either point or block, as the case may be. Typically, the block variant of each method converges faster than the point version, but requires more computational work per iteration. Thus, it is usually not clear without additional analysis which variant will be more effective.

The *Jacobi iteration* takes $M_{\mathrm{J}} = D$ and $N_{\mathrm{J}} = L + U$, resulting in the recurrence

$$Dx_{k+1} = (L + U)x_k + b. \tag{23}$$

The associated iteration matrix is $G_{\mathrm{J}} = D^{-1}(L + U)$. The *Gauss–Seidel iteration*

takes $M_{GS} = D - L$ and $N_{GS} = U$, resulting in the recurrence

$$(D - L)x_{k+1} = Ux_k + b. \tag{24}$$

The associated iteration matrix is $G_{GS} = (D - L)^{-1}U$. Although (24) may at first appear a little more complicated than (23), it is in fact easier to implement in practice on a sequential machine, since one can overwrite x_k when computing x_{k+1} in (24), whereas this is generally not possible for (23). However, for the Gauss–Seidel iteration, the jth component of x_{k+1} might depend on the ith component of x_{k+1} for $i < j$, because of the factor L on the left side of (24). This often inhibits vectorization and parallelization of the Gauss–Seidel iteration. Note that the Jacobi iteration has no such dependence, and so might be more effective on a vector or parallel machine.

Relaxation methods for $Ax = b$ can be written in the form

$$x_{k+1} = x_k + \omega(\hat{x}_{k+1} - x_k), \tag{25}$$

where $\omega \neq 0$ is the relaxation parameter and $\hat{x}_{k+1}$ is computed from x_k by some other iterative method. The best known of these schemes is *successive overrelaxation* (SOR) for which

$$D\hat{x}_{k+1} = Lx_{k+1} + Ux_k + b. \tag{26}$$

Equations(25) and (26) can be combined to give

$$\left(\frac{1}{\omega}D - L\right)x_{k+1} = \left(\frac{1-\omega}{\omega}D + U\right)x_k + b, \tag{27}$$

which is an iteration of the form (18) with $M_{SOR}(\omega) = (1/\omega)D - L$ and $N_{SOR}(\omega) = [(1-\omega)/\omega]D + U$. It follows immediately from (24) and (27) that the SOR iteration reduces to the Gauss–Seidel method if $\omega = 1$. Moreover, because of the similarity between (24) and (27), the SOR iteration shares with the Gauss–Seidel method the implementation advantages and disadvantages noted above.

Overrelaxation corresponds to choosing $\omega > 1$ in (25) or (27), while underrelaxation corresponds to choosing $\omega \in (0, 1)$. Historically, $\omega > 1$ was used in SOR for the solution of elliptic PDEs – hence the name successive *over*relaxation – but underrelaxation is more effective for some problems. See for example Young (1971) for a more complete discussion.

The *symmetric SOR* (SSOR) method takes one half step of SOR with the equations solved in the standard order followed by one half step of SOR with the equations solved in the reverse order:

$$\left(\frac{1}{\omega}D - L\right)x_{k+1/2} = \left(\frac{1-\omega}{\omega}D + U\right)x_k + b, \tag{28}$$

$$\left(\frac{1}{\omega}D - U\right)x_{k+1} = \left(\frac{1-\omega}{\omega}D + L\right)x_{k+1/2} + b. \tag{29}$$

These two half steps can be combined into one step of the form (18) with

$$M_{SSOR}(\omega) = \frac{\omega}{2-\omega}\left(\frac{1}{\omega}D - L\right)D^{-1}\left(\frac{1}{\omega}D - U\right)$$

and

$$N_{SSOR}(\omega) = \frac{(\omega-1)^2}{\omega(2-\omega)}D + \frac{1-\omega}{2-\omega}(L+U) + \frac{\omega}{2-\omega}LD^{-1}U.$$

Note that, if $A = D - L - U$ is a real, symmetric, positive-definite matrix and $\omega \in (0, 2)$, then $M_{\mathrm{SSOR}}(\omega)$ is a real, symmetric, positive-definite matrix too, since, in this case, $\omega/(2-\omega) > 0$, both D and D^{-1} are symmetric positive-definite, and $(1/\omega)D - L = [(1/\omega)D - U]^{\mathrm{T}}$ is nonsingular. This property of $M_{\mathrm{SSOR}}(\omega)$ plays an important role in the effective acceleration of the SSOR iteration, as discussed in Sec. 3.2.

We now consider the convergence of the Jacobi, Gauss–Seidel, SOR, and SSOR methods. It is easy to show that the Jacobi iteration (23) converges if A is either row- or column-diagonally dominant. It can also be shown that the Gauss–Seidel iteration (24) converges if A is Hermitian positive-definite. Furthermore, if A is *consistently ordered*, a property enjoyed by a large class of matrices, including many that arise from the discretization of PDEs [see, for example, Young (1971) for details], it can be shown that the Gauss–Seidel iteration converges twice as fast as the Jacobi iteration if either one converges.

The iteration matrix associated with the SOR iteration (27) is

$$G_{\mathrm{SOR}}(\omega) = [M_{\mathrm{SOR}}(\omega)]^{-1} N_{\mathrm{SOR}}(\omega) = \left(\frac{1}{\omega}D - L\right)^{-1}\left(\frac{1-\omega}{\omega}D + U\right).$$

For any nonsingular A with nonsingular D, it can be shown that $\rho(G_{\mathrm{SOR}}(\omega)) \geq |\omega - 1|$ with equality possible if and only if all eigenvalues of $G_{\mathrm{SOR}}(\omega)$ have magnitude $|\omega - 1|$. So, if $\omega \in \mathbb{R}$, as is normally the case, a necessary condition for the convergence of SOR is $\omega \in (0, 2)$. It can also be shown that, if A is Hermitian, D is positive-definite, and $\omega \in \mathbb{R}$, then the SOR iteration converges if and only if A is positive-definite and $\omega \in (0, 2)$. Furthermore, if A is consistently ordered and all the eigenvalues of $G_{\mathrm{J}} = D^{-1}(L + U)$ are real and lie in $(-1, 1)$, then the optimal choice of the SOR parameter ω is

$$\omega_0 = \frac{2}{1 + \sqrt{1 - [\rho(G_{\mathrm{J}})]^2}} \in (1, 2)$$

and

$$\begin{aligned}\rho(G_{\mathrm{SOR}}(\omega_0)) &= \omega_0 - 1 \\ &= \left(\frac{\rho(G_{\mathrm{J}})}{1 + \sqrt{1 - [\rho(G_{\mathrm{J}})]^2}}\right)^2 \\ &= \min_{\omega} \rho(G_{\mathrm{SOR}}(\omega)) < \rho(G_1) \\ &= \rho(G_{\mathrm{GS}}) = [\rho(G_{\mathrm{J}})]^2 \\ &< \rho(G_{\mathrm{J}}). \end{aligned} \tag{30}$$

In many cases, though, it is not convenient to calculate the optimal ω. Hageman and Young (1981) discuss heuristics for choosing a "good" ω.

If A is a real, symmetric, positive-definite matrix, then the SSOR iteration (28)–(29) converges for any $\omega \in (0, 2)$. Moreover, determining the precise value of the optimal ω is not nearly as critical for SSOR as it is for SOR, since, unlike SOR, the rate of convergence of SSOR is relatively insensitive to the choice of ω. However, for SSOR to be effective,

$$\rho(D^{-1}LD^{-1}U) \leq \frac{1}{4}$$

should be satisfied – or nearly so. If this is the case, then a good value for ω is

$$\omega_1 = \frac{2}{1 + \sqrt{2[1 - \rho(G_{\mathrm{J}})]}}$$

and

$$\rho(G_{\mathrm{SSOR}}(\omega_1)) \leq \frac{1 - \{[1 - \rho(G_{\mathrm{J}})]/2\}^{1/2}}{1 + \{[1 - \rho(G_{\mathrm{J}})]/2\}^{1/2}}, \tag{31}$$

where $G_{\mathrm{SSOR}}(\omega) = [M_{\mathrm{SSOR}}(\omega)]^{-1} N_{\mathrm{SSOR}}(\omega)$ is the SSOR iteration matrix.

For a large class of problems, including many that arise from the discretization of elliptic PDEs, $\rho(G_{\mathrm{J}}) = 1 - \varepsilon$ for some ε satisfying $0 < \varepsilon \ll 1$. For such problems, if (30) is valid and (31) holds with $=$ in place of $\leq$, then

$$\rho(G_{\mathrm{J}}) = 1 - \varepsilon,$$

$$\rho(G_{\mathrm{GS}}) = (\rho(G_{\mathrm{J}}))^2 \approx 1 - 2\varepsilon,$$

$$\rho(G_{\mathrm{SSOR}}(\omega_1)) \approx 1 - 2\sqrt{\varepsilon/2},$$

$$\rho(G_{\mathrm{SOR}}(\omega_0)) \approx 1 - 2\sqrt{2\varepsilon},$$

whence the rates of convergence for these schemes are

$$R(G_{\mathrm{J}}) \approx \varepsilon,$$

$$R(G_{\mathrm{GS}}) = 2R(G_{\mathrm{J}}) \approx 2\varepsilon,$$

$$R(G_{\mathrm{SSOR}}(\omega_1)) \approx 2\sqrt{\varepsilon/2},$$

$$R(G_{\mathrm{SOR}}(\omega_0)) \approx 2\sqrt{2\varepsilon} \approx 2R(G_{\mathrm{SSOR}}(\omega_1)),$$

showing that the Gauss–Seidel iteration converges twice as fast as the Jacobi iteration, the SOR iteration converges about twice as fast as the SSOR iteration, and the SOR and SSOR iterations converge much faster than either the Gauss–Seidel or Jacobi iteration. However, the SSOR iteration often has the advantage, not normally shared by the SOR method, that it can be accelerated effectively, as discussed in Sec. 3.2.

Another class of basic iterative methods is based on *incomplete factorizations*. For brevity, we describe only the subclass of *incomplete Cholesky factorizations* (ICFs) here; the other schemes are similar. See an advanced text, such as Axelsson (1994), for details.

For an ICF to be effective, A should be symmetric positive-definite (or nearly so), large, and sparse. If A is banded, then the band containing the nonzeros should also be sparse, as is the case for the discretization of Poisson's equation shown in Sec. 11.3.2. The general idea behind the ICFs is to compute a lower-triangular matrix L_{ICF} such that $M_{\mathrm{ICF}} = L_{\mathrm{ICF}} L_{\mathrm{ICF}}^{\mathrm{T}}$ is in some sense close to A and L_{ICF} is much sparser than L, the true Cholesky factor of A. Then employ the iteration (19)–(20) to compute a sequence of approximations $x_1, x_2, \ldots$ to x, the solution of $Ax = b$. Note that (19) can be solved efficiently by forward elimination and back substitution as described in Secs. 2.4 and 2.2, respectively, since the factorization $M_{\mathrm{ICF}} = L_{\mathrm{ICF}} L_{\mathrm{ICF}}^{\mathrm{T}}$ is known. This scheme, or an accelerated variant of it, is often very effective if it converges rapidly and L_{ICF} is much sparser than L.

A simple, but often effective, way of computing $L_{\mathrm{ICF}} = [l_{ij}]$ is to apply the Cholesky factorization described in Table 4, but to set $l_{ij} = 0$ whenever $a_{ij} = 0$, where $A = [a_{ij}]$. Thus, L_{ICF} has the same sparsity pattern as the lower-triangular part of A, whereas the true Cholesky factor L of A might suffer significant fill-in, as described in Sec. 2.7. Unfortunately, this simple ICF is not always stable.

As noted in Sec. 2.9, iterative improvement is a basic iterative method of this form, although, for iterative improvement, the error $N = M - A$ is due entirely to rounding errors, whereas, for the incomplete factorizations considered here, the error $N = M - A$ is typically also due to dropping elements from the factors of M to reduce fill-in.

For a more complete discussion of ICFs and other incomplete factorizations, their convergence properties, and their potential for acceleration, see an advanced text such as Axelsson (1994).

We end this section with a brief discussion of *alternating-direction implicit* (ADI)

methods. A typical example of a scheme of this class is the Peaceman–Rachford method

$$(H + \alpha_n I)x_{n+1/2} = b - (V - \alpha_n I)x_n, \tag{32}$$

$$(V + \alpha'_n I)x_{n+1} = b - (H - \alpha'_n I)x_{n+1/2}, \tag{33}$$

where $A = H + V$, $\alpha_n > 0$, $\alpha'_n > 0$, and A, H, and V are real, symmetric, positive-definite matrices. For many problems, it is possible to choose H, V, and $\{\alpha_n, \alpha'_n\}$, so that the iteration (32)–(33) converges rapidly, and it is much cheaper to solve (32) and (33) than it is to solve $Ax = b$. For example, for the standard five-point discretization of a separable two-dimensional elliptic PDE, H and V can be chosen to be essentially tridiagonal and the rate of convergence of (32)–(33) is proportional to $1/\log h^{-1}$, where h is the mesh size used in the discretization. In contrast, the rate of convergence for SOR with the optimal ω is proportional to h. Note that $h \ll 1/\log h^{-1}$ for $0 < h \ll 1$, supporting the empirical evidence that ADI schemes converge much more rapidly than SOR for many problems. However, ADI schemes are not applicable to as wide a class of problems as SOR is.

For a more complete discussion of ADI schemes, see an advanced text such as Varga (1962) or Young (1971).

3.2 The Conjugate-Gradient Method

The *conjugate-gradient* (CG) *method* for the solution of the linear system $Ax = b$ is a member of a broader class of methods often called polynomial acceleration techniques or Krylov-subspace methods. (The basis of these names is explained below.) Although many schemes in this broader class are very useful in practice, we discuss only CG here, but note that several of these schemes, including Chebyshev acceleration and GMRES, apply to more general problems than CG. The interested reader should consult an advanced text such as Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981) for a more complete discussion of polynomial acceleration techniques and Krylov-subspace methods. The close relationship between CG and the Lanczos method is discussed in Golub and Van Loan (1989).

The *preconditioned conjugate-gradient* (PCG) *method* can be viewed either as an acceleration technique for the basic iterative method (18) or as CG applied to the preconditioned system $M^{-1}Ax = M^{-1}b$, where the splitting matrix M of (18) is typically called a preconditioning matrix in this context. We adopt the second point of view in this subsection.

An instructive way of deriving PCG is to exploit its relationship to the minimization technique of the same name described in Sec. 7.5. To this end, assume that the basic iterative method (18) is *symmetrizable*. That is, there exists a real, nonsingular matrix W such that $S = WM^{-1}AW^{-1}$ is a real, symmetric, positive-definite (SPD) matrix. Consider the quadratic functional $F(y) = \frac{1}{2}y^{\mathrm{T}}Sy - y^{\mathrm{T}}\hat{b}$, where $\hat{b} = WM^{-1}b$ and $y \in \mathbb{R}^n$. It is easy to show that the unique minimum of $F(y)$ is the solution $\hat{x}$ of $S\hat{x} = \hat{b}$. It follows immediately from the relations for S and $\hat{b}$ given above that $x = W^{-1}\hat{x}$ is the solution of both the preconditioned system $M^{-1}Ax = M^{-1}b$ and the original system $Ax = b$. If we take $M = W = I$, then $S = W(M^{-1}A)W^{-1} = A$ and, if we assume that A is a real SPD matrix, PCG reduces to the standard (unpreconditioned) CG method.

It is easy to show that if both A and M are real SPD matrices, then $M^{-1}A$ is symmetrizable. Moreover, many important practical problems, such as the numerical solution of self-adjoint elliptic PDEs, give rise to matrices A that are real SPD. Furthermore, if A is a real SPD matrix, then the splitting matrix M associated with the RF, Jacobi, and SSOR [with $\omega \in (0, 2)$] iterations is a real SPD matrix too, as is the M given by an incomplete Cholesky factorization, provided it exists. Hence, these basic iterative methods are symmetrizable in this case and so PCG can be used to accelerate their convergence. In contrast, the SOR iteration with the optimal ω is generally not symmetrizable. So PCG is not even applicable and more general Krylov-subspace methods normally do not accelerate its convergence.

We assume throughout the rest of this subsection that A and M are real SPD matrices, because this case arises most frequently in practice and also because it simplifies the discussion below. Using this assumption and applying several mathematical identities, we get the computationally effective variant of PCG shown in Table 6. Note that W does not appear explicitly in this algorithm. Also note that, if we choose $M = I$, then $\tilde{r}_k = r_k$ and the PCG method reduces to the unpreconditioned CG method for $Ax = b$.

Many other mathematically equivalent forms of PCG exist. Moreover, as noted above, a more general form of PCG can be used if $M^{-1}A$ is symmetrizable, without either A or M being a real SPD matrix. For a discussion of the points, see an advanced text such as Hageman and Young (1981).

Before considering the convergence of PCG, we introduce some notation. The *energy norm* of a vector $y \in \mathbb{R}^n$ with respect to a real SPD matrix B is $\|y\|_{B^{1/2}} = (y^{\mathrm{T}}By)^{1/2}$.

Tab. 6 The preconditioned conjugate-gradient (PCG) method for solving $Ax = b$

```
choose an initial guess x_0
compute r_0 = b − Ax_0
solve M r̃_0 = r_0
set p_0 = r̃_0
for k = 0, 1, ... until convergence do
   α_k = r_k^T r̃_k / p_k^T A p_k
   x_{k+1} = x_k + α_k p_k
   r_{k+1} = r_k − α_k A p_k
   solve M r̃_{k+1} = r_{k+1}
   β_k = r_{k+1}^T r̃_{k+1} / r_k^T r̃_k
   p_{k+1} = r̃_{k+1} + β_k p_k
end
```

The *Krylov subspace* of degree $k-1$ generated by a vector v and a matrix W is $\mathscr{K}_k(v, W) = \mathrm{span}\{v, Wv, \ldots, W^{k-1}v\}$.

It is easy to show that the r_k that occurs in Table 6 is the residual $r_k = b - Ax_k$ associated with x_k for the system $Ax = b$ and that $\tilde{r}_k = M^{-1}r_k = M^{-1}b - M^{-1}Ax_k$ is the residual associated with x_k for the preconditioned system $M^{-1}Ax = M^{-1}b$. Let $e_k = x - x_k$ be the error associated with x_k. It can be shown that the x_k generated by PCG is a member of the shifted Krylov subspace $x_0 + \mathscr{K}_k(\tilde{r}_0, M^{-1}A) \equiv \{x_0 + v : v \in \mathscr{K}_k(\tilde{r}_0, M^{-1}A)\}$. Hence, PCG is in the broader class of Krylov-subspace methods characterized by this property. Moreover, it can be shown that the x_k generated by PCG is the unique member of $x_0 + \mathscr{K}_k(\tilde{r}_0, M^{-1}A)$ that minimizes the energy norm of the error $\|e_k\|_{A^{1/2}} = (e_k^{\mathrm{T}}Ae_k)^{1/2}$ over all vectors of the form $e'_k = x - x'_k$, where x'_k is any other member of $x_0 + \mathscr{K}_k(\tilde{r}_0, M^{-1}A)$. Equivalently, $e_k = P_k^*(M^{-1}A)e_0$, where $P_k^*(z)$ is the polynomial that minimizes $\|P_k(M^{-1}A)e_0\|_{A^{1/2}}$ over all polynomials $P_k(z)$ of degree k that satisfy $P_k(0) = 1$. This result is the basis of the characterization that PCG is the optimal polynomial acceleration scheme for the basic iterative method (18).

In passing, note that the iterate x_k generated by the basic iteration (18) is in the shifted Krylov subspace $x_0 + \mathscr{K}_k(\tilde{r}_0, M^{-1}A)$ also and that, by (21), the associated error satisfies $e_k = (I - M^{-1}A)^k e_0$. So (18) is a Krylov-subspace method too and its error satisfies the polynomial relation described above with $P_k(z) = (1-z)^k$. Thus, the associated PCG method is guaranteed to accelerate the convergence of the basic iteration (18) – at least when the errors are measured in the energy norm.

The characterization of its error discussed above can be very useful in understanding the performance of PCG. For example, it can be used to prove the *finite termination property* of PCG. That is, if $M^{-1}A$ has m distinct eigenvalues, then $x_m = x$, the exact solution of $Ax = b$. Since $M^{-1}A$ has n eigenvalues, $m \le n$ always and $m \ll n$ sometimes. We caution the reader, though, that the argument used to prove this property of PCG assumes exact arithmetic. In floating-point arithmetic, we rarely get $e_m = 0$, although we frequently get $e_{\tilde{m}}$ small for some $\tilde{m}$ possibly a little larger than m.

The proof of the finite termination property can be extended easily to explain the rapid convergence of PCG when the eigenvalues of $M^{-1}A$ fall into a few small clusters. So a preconditioner M is good if the eigenvalues of $M^{-1}A$ are much more closely clustered than those of the unpreconditioned matrix A.

Because of the finite termination property, both CG and PCG can be considered direct methods. However, both are frequently used as iterative schemes, with the iteration terminated long before $e_m = 0$. Therefore, it is important to understand how the error decreases with k, the iteration count. To this end, first note that the characterization of the error ensures that $\|e_{k+1}\|_{A^{1/2}} \le \|e_k\|_{A^{1/2}}$ with equality only if $e_k = 0$. That is, PCG is a *descent* method in the sense that some norm of the error decreases on every iteration. Not all iterative methods enjoy this useful property.

The characterization of the error can also be used to show that

$$\|e_k\|_{A^{1/2}} \le 2\left(\frac{\sqrt{\lambda_n/\lambda_1} - 1}{\sqrt{\lambda_n/\lambda_1} + 1}\right)^k \|e_0\|_{A^{1/2}}, \tag{34}$$

where λ_n and λ_1 are the largest and smallest eigenvalues, respectively, of $M^{-1}A$. (Note that $\lambda_n, \lambda_1 \in \mathbb{R}$ and $\lambda_n \ge \lambda_1 > 0$ since $M^{-1}A$ is symmetrizable.) It follows easily from the definition of the energy norm and (34) that

$$\|e_k\|_2 \le 2\sqrt{\kappa_2(A)}\left(\frac{\sqrt{\lambda_n/\lambda_1} - 1}{\sqrt{\lambda_n/\lambda_1} + 1}\right)^k \|e_0\|_2, \tag{35}$$

where $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A in the 2-norm (see Sec. 2.8). Although (35) is generally not as tight as (34), it may be more relevant to the practitioner. It follows from either (34) or (35) that, in this context, a preconditioner M is good if λ_n/λ_1 is (much) closer to 1 than is the ratio of the largest to smallest eigenvalues of A.

Unlike many other iterative methods, such as SOR, PCG does not require an estimate of any parameters, although some stopping procedures for PCG require an estimate of the extreme eigenvalues of A or $M^{-1}A$. See an advanced text such as Axelsson (1994), Golub and Van Loan (1989), Hageman and Young (1981) for details.

4 Overdetermined and Underdetermined Linear Systems

A linear system $Ax = b$, with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ given and $x \in \mathbb{R}^n$ unknown, is called *overdetermined* if $m > n$. Such a system typically has no solution. However, there is always an x that minimizes $\|b - Ax\|_2$. Such an x is called a *least-squares* solution to the overdetermined linear system $Ax = b$. Moreover, if $\text{rank}(A) = n$, then the least-squares solution is unique.

A linear system $Ax = b$, as above, is called *underdetermined* if $m < n$. Such a system typically has infinitely many solutions. If $Ax = b$ has a solution, then there is an x of minimum Euclidean norm, $\|x\|_2$, that satisfies $Ax = b$. Such an x is called a *least-squares* solution to the underdetermined linear system $Ax = b$. It can be shown that, if $\text{rank}(A) = m$, then the least-squares solution is unique.

In the following subsections, we describe methods to compute the least-squares solution of overdetermined and underdetermined linear systems, assuming that the matrix A has full rank, i.e., $\text{rank}(A) = \min(m, n)$. For the more general case of $\text{rank}(A) < \min(m, n)$, see an advanced text such as Golub and Van Loan (1989).

4.1 The Normal Equations for Overdetermined Linear Systems

Let $Ax = b$ be an overdetermined linear system with m equations in n unknowns and with $\text{rank}(A) = n$. The x that minimizes $\|b - Ax\|_2$ satisfies

$$\frac{\partial[(b - Ax)^{\mathrm{T}}(b - Ax)]}{\partial x_j} = 0 \text{ for } j = 1, \ldots, n.$$

These relations can be rewritten in matrix form as $A^{\mathrm{T}}(b - Ax) = 0$ or equivalently

$$A^{\mathrm{T}}Ax = A^{\mathrm{T}}b, \tag{36}$$

which is called the system of *normal equations* for the overdetermined linear system $Ax = b$. It can be shown that the matrix $A^{\mathrm{T}}A$ is symmetric and positive-definite if $\text{rank}(A) = n$. So the system (36) has a unique solution, which is the least-squares solution to $Ax = b$.

Note that the matrix $A^{\mathrm{T}}A$ is $n \times n$. Computing the matrix $A^{\mathrm{T}}A$ requires about $mn^2/2$ flops (floating-point operations), while computing $A^{\mathrm{T}}b$ requires about mn flops, and solving (36) requires about $n^3/6$ flops, if we assume that the Cholesky factorization (see Sec. 2.6) is used.

The condition number of the matrix $A^{\mathrm{T}}A$ is often large. More specifically, it can be shown that if A is square and nonsingular, then the condition number of $A^{\mathrm{T}}A$ is the square of the condition number of A (see Sec. 2.8). As a result, the approach described above of forming and solving the normal equations (36) often leads to a serious loss of accuracy. Therefore, in Secs. 4.4 and 4.7, we discuss more stable alternatives for solving least-squares problems.

4.2 The Normal Equations for Underdetermined Linear Systems

Let $Ax = b$ be an underdetermined linear system with m equations in n unknowns and with $\text{rank}(A) = m$. It can be shown that the least-squares solution x to $Ax = b$ can be written in the form $A^{\mathrm{T}}y$ for some $y \in \mathbb{R}^m$ that satisfies

$$AA^{\mathrm{T}}y = b. \tag{37}$$

This is a linear system of size $m \times m$, called the system of *normal equations* for the underdetermined linear system $Ax = b$. It can be shown that the matrix AA^{T} is symmetric and positive-definite if $\mathrm{rank}(A) = m$. So the system (37) has a unique solution. The unique least-squares solution to $Ax = b$ is $x = A^{\mathrm{T}}y$.

Computing the matrix AA^{T} requires about $nm^2/2$ flops, while computing $A^{\mathrm{T}}y$ requires about mn flops, and solving (37) requires about $m^3/6$ flops, if the Cholesky factorization (see Sec. 2.6) is used.

As is the case for overdetermined systems, the method of forming and solving the normal equations (37) to compute the least-squares solution to $Ax = b$ is numerically unstable in some cases. In Secs. 4.5 and 4.8, we discuss more stable alternatives.

4.3 Householder Transformations and the *QR* Factorization

An *orthogonal transformation* is a linear change of variables that preserves the length of vectors in the Euclidean norm. Examples are a rotation about an axis or a reflection across a plane. The following is an example of an orthogonal transformation $y = (y_1, y_2)$ to $x = (x_1, x_2)$:

$$x_1 = 0.6y_1 + 0.8y_2,$$
$$x_2 = 0.8y_1 - 0.6y_2.$$

It is easy to see that $\|x\|_2 = \|y\|_2$.

An *orthogonal matrix* is an $m \times n$ matrix Q with the property $Q^{\mathrm{T}}Q = I$. Note that, if (and only if) Q is square, i.e., $m = n$, the relation $Q^{\mathrm{T}}Q = I$ is equivalent to $QQ^{\mathrm{T}} = I$ and so $Q^{\mathrm{T}} = Q^{-1}$. An orthogonal transformation of y to x can be written as $x = Qy$, where Q is an orthogonal matrix. In the above example, the corresponding orthogonal matrix is

$$Q = \begin{pmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{pmatrix}.$$

If Q is an orthogonal matrix, then

$$\|x\|_2^2 = \|Qy\|_2^2 = (Qy)^{\mathrm{T}}(Qy) = y^{\mathrm{T}}(Q^{\mathrm{T}}Q)y$$
$$= y^{\mathrm{T}}y = \|y\|_2^2,$$

whence $\|x\|_2 = \|y\|_2$. This property is exploited in the methods described below to solve least-squares problems. Moreover, the numerical stability of these schemes is due at least in part to the related observation that orthogonal transformations do not magnify rounding error.

A *Householder transformation* or *Householder reflection* is an orthogonal matrix of the form $H = I - 2ww^{\mathrm{T}}$, where $\|w\|_2 = 1$. Note that, when H is 2×2, the effect of H on a vector x is equivalent to reflecting the vector x across the plane perpendicular to w and passing through the origin of x. A Householder reflection can be used to transform a nonzero vector into one containing mainly zeros.

An $m \times n$ matrix $R = [r_{ij}]$ is *right triangular* if $r_{ij} = 0$ for $i > j$. Note that if (and only if) R is square, i.e., $m = n$, the terms right triangular and upper triangular are equivalent.

Let A be an $m \times n$ matrix with $m \geq n$. The QR factorization of A expresses A as the product of an $m \times m$ orthogonal matrix Q and an $m \times n$ right-triangular matrix R. It can be computed by a sequence $H_1, H_2, \ldots, H_n$ of Householder transformations to reduce A to right-triangular form R. More specifically, this variant of the QR factorization proceeds in n steps (or $n - 1$ if $n = m$). Starting with $A_0 = A$, at step k for $k = 1, \ldots, n$, H_k is applied to the partially processed matrix A_{k-1} to zero the components $k + 1$ to m of column k of A_{k-1}. $Q =$

$H_1 H_2, \ldots, H_n$ and $R = A_n$, the last matrix to be computed. For the details of the QR factorization algorithm using Householder transformations or other elementary orthogonal transformations, see Hager (1988) or Golub and Van Loan (1989).

4.4 Using the *QR* Factorization to Solve Overdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m \geq n$ and $\text{rank}(A) = n$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense, if $m > n$). Let $A = QR$ be the QR factorization of A. Note that

$$\begin{aligned} \|Ax - b\|_2 &= \|QRx - b\|_2 \\ &= \|Q(Rx - Q^{\mathrm{T}}b)\|_2 \\ &= \|Rx - Q^{\mathrm{T}}b\|_2. \end{aligned}$$

Therefore, we can use the QR factorization of A to reduce the problem of solving $Ax = b$ to that of solving $Rx = Q^{\mathrm{T}}b$, a much simpler task. We solve the latter by first computing $y = Q^{\mathrm{T}}b$. Let $\hat{y}$ be the vector consisting of the first n components of y and $\hat{R}$ the upper-triangular matrix consisting of the first n rows of R. Now solve $\hat{R}x = \hat{y}$ by back substitution (see Sec. 2.2). Then x is the least-squares solution to both $Rx = Q^{\mathrm{T}}b$ and $Ax = b$.

It can be shown that the QR factorization algorithm applied to an $n \times n$ linear system requires about $2n^3/3$ flops, which is about twice as many as the LU factorization needs. However, QR is a more stable method than LU and it requires no pivoting. The QR factorization algorithm applied to an $m \times n$ linear system requires about twice as many flops as forming and solving the normal equations. However, QR is a more stable method than solving the normal equations.

4.5 Using the *QR* Factorization to Solve Underdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m < n$ and $\text{rank}(A) = m$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense). Obtain the QR factorization of A^{T} by the QR factorization algorithm: $A^{\mathrm{T}} = QR$, where Q is an $n \times n$ orthogonal matrix and R is an $n \times m$ right-triangular matrix. Let $\hat{R}$ be the upper-triangular matrix consisting of the first m rows of R. Solve $\hat{R}^{\mathrm{T}}\hat{y} = b$ by back substitution (see Sec. 2.2) and let $y = (\hat{y}^{\mathrm{T}}, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^n$. Note that y is the vector of minimal Euclidean norm that satisfies $R^{\mathrm{T}}y = b$. Finally compute $x = Qy$ and note that x is the vector of minimal Euclidean norm that satisfies $R^{\mathrm{T}}Q^{\mathrm{T}}x = b$ or equivalently $Ax = b$. That is, x is the least-squares solution to $Ax = b$.

4.6 The Gram–Schmidt Orthogonalization Algorithm

The *Gram–Schmidt orthogonalization* algorithm is an alternative to QR. The modified version of the algorithm presented below is approximately twice as fast as QR and more stable than solving the normal equations.

Assume that $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and that A has n linearly independent columns $\{a_j\}_{j=1}^n$. The Gram–Schmidt algorithm applied to A generates an $m \times n$ orthogonal matrix Q and an $n \times n$ upper-triangular matrix R satisfying $A = QR$. In Table 7, we present a stable version of the algorithm, often called the *modified Gram–Schmidt algorithm*.

Tab. 7 The modified Gram–Schmidt algorithm

```
for j = 1, ..., n do                    (q_j and a_j are
  q_j = a_j                               columns j of Q
  for i = 1, ..., j − 1 do                and A,
    r_ij = q_i^T q_j                      respectively)
    q_j = q_j − r_ij q_i
  end
  r_jj = ||q_j||_2
  q_j = q_j / r_jj
end
```

4.7 Using Gram–Schmidt to Solve Overdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m \geq n$ and $\operatorname{rank}(A) = n$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense, if nonsquare). The method for solving $Ax = b$ using the Gram–Schmidt algorithm is similar to that described in Sec. 4.4 for the QR factorization.

First compute the QR factorization of A by the Gram–Schmidt algorithm. Next compute $y = Q^{\mathrm{T}}b$ and solve $Rx = y$ by back substitution (see Sec. 2.2). Then x is the least-squares solution to $Ax = b$.

It can be shown that the Gram–Schmidt algorithm applied to an $n \times n$ linear system requires about $n^3/3$ flops, which is about the same number of arithmetic operations as the LU factorization algorithm needs and about half of what the QR factorization algorithm requires. Moreover, the modified Gram–Schmidt algorithm, as presented in Table 7, is relatively stable. The Gram–Schmidt algorithm applied to an $m \times n$ linear system requires about the same number of flops as that needed to form and to solve the normal equations, but the Gram–Schmidt algorithm is more stable.

4.8 Using Gram–Schmidt to Solve Underdetermined Linear Systems

Assume that A is an $m \times n$ matrix, with $m < n$ and $\operatorname{rank}(A) = m$. Let $Ax = b$ be the linear system to be solved (in the least-squares sense). The method for solving $Ax = b$ using the Gram–Schmidt algorithm is similar to that described in Sec. 4.5 for the QR factorization.

First compute the QR factorization of A^{T} by the Gram–Schmidt algorithm. Next solve $R^{\mathrm{T}}y = b$ by back substitution (see Sec. 2.2) and set $x = Qy$. Then x is the least-squares solution to $Ax = b$.

5 Eigenvalues and Eigenvectors of Matrices

Given an $n \times n$ matrix A, $\lambda \in \mathbb{C}$, and $x \in \mathbb{C}^n$ satisfying $Ax = \lambda x$, λ is called an *eigenvalue* of A and x is called an *eigenvector* of A. The relation $Ax = \lambda x$ can be rewritten as $(A - \lambda I)x = 0$, emphasizing that λ is an eigenvalue of A if and only if $A - \lambda I$ is singular and that an eigenvector x is a nontrivial solution to $(A - \lambda I)x = 0$. It also follows that the eigenvalues of A are the roots of $\det(A - \lambda I) = 0$, the *characteristic equation* of A. The polynomial $p(\lambda) = \det(A - \lambda I)$ of degree n is called the *characteristic polynomial* of A and plays an important role in the theory of eigenvalues.

An $n \times n$ matrix A has precisely n eigenvalues, not necessarily distinct. It also has at least one eigenvector for each distinct eigenvalue. Note also that, if x is an eigenvector of A, then so is any (scalar) multiple of x and the corresponding eigenvalue is the same. We often choose an eigenvector of norm one in some vector norm, often the Euclidean norm, as the representative.

Two matrices A and B are similar if $A = W^{-1}BW$ for some nonsingular matrix W. The matrix W is often referred to as a *similarity transformation*. It is easy to see that, if $Ax = \lambda x$, then $B(Wx) = \lambda(Wx)$. Thus, similar matrices have the same eigenvalues and their eigenvectors are related by the similarity transformation W. Similarity transformations are often used in numerical methods for eigenvalues and eigenvectors to transform a matrix A into another one B that has the same eigenvalues and related eigenvectors, but whose eigenvalues and eigenvectors are in some sense easier to compute than those of A.

Eigenvalues play a major role in the study of convergence of iterative methods (see Sec. 3). Eigenvalues and eigenvectors are also of great importance in understanding the stability and other fundamental properties of many physical systems.

The matrix A is often large, sparse, and symmetric. These properties can be exploited to great advantage in numerical schemes for calculating the eigenvalues and eigenvectors of A.

A common approach to calculate the eigenvalues (and possibly the eigenvectors) of a matrix A consists of two stages. First, the matrix A is transformed to a similar but simpler matrix B, usually tridiagonal, if A is symmetric (or Hermitian), or Hessenberg, if A is nonsymmetric (or non-Hermitian). Then, the eigenvalues (and possibly the eigenvectors) of B are calculated. An exception to this approach is the power method (see Sec. 5.1).

A standard procedure for computing the eigenvectors of a matrix A is to calculate the eigenvalues first then use them to compute the eigenvectors by inverse iteration (see Sec. 5.4). Again, an exception to this approach is the power method (see Sec. 5.1).

Before describing numerical methods for the eigenvalue problem, we comment briefly on the sensitivity of the eigenvalues and eigenvectors to perturbations in the matrix A, since a backward error analysis can often show that the computed eigenvalues and eigenvectors are the exact eigenvalues and eigenvectors of a slightly perturbed matrix $\hat{A} = A + E$, where E is usually small relative to A. In general, if A is symmetric, its eigenvalues are well-conditioned with respect to small perturbations E. That is, the eigenvalues of $\hat{A}$ and A are very close. This, though, is not always true of the eigenvectors of A, particularly if the associated eigenvalue is a multiple eigenvalue or close to another eigenvalue of A. If A is nonsymmetric, then both its eigenvalues and eigenvectors may be poorly conditioned with respect to small perturbations E. Therefore, the user of a computer package for calculating the eigenvalues of a matrix should be cautious about the accuracy of the numerical results. For a further discussion of the conditioning of the eigenvalue problem, see Wilkinson (1965).

5.1 The Power Method

The power method is used to calculate the eigenvalue of largest magnitude of a matrix A and the associated eigenvector. Since matrix–vector products are the dominant computational work required by the power method, this scheme can exploit the sparsity of the matrix to great advantage.

Let λ be the eigenvalue of A of largest magnitude and let x be an associated eigenvector. Also let z_0 be an initial guess for some multiple of x. The power method, shown in Table 8, is an

Tab. 8 The power method

```
Pick z_0
for k = 1, 2, ... do
    w_k = A z_{k-1}
    Choose m ∈ {1, ..., n} such that
        |(w_k)_m)| ≥ |(w_k)_i)| for i = 1, ..., n
    z_k = w_k/(w_k)_m
    μ_k = (w_k)_m/(z_{k-1})_m
    test stopping criterion
end
```

iterative scheme that generates a sequence of approximations $z_1, z_2, \ldots$ to some multiple of x and another sequence of approximations $\mu_1, \mu_2, \ldots$ to λ. In the scheme shown in Table 8, z_k is normalized so that the sequence $z_1, z_2, \ldots$ converges to an eigenvector x of A satisfying $\|x\|_\infty = 1$. Normalizations of this sort are frequently used in eigenvector calculations.

The power method is guaranteed to converge if A has a single eigenvalue λ of largest magnitude. The rate of convergence depends on $|\lambda_2|/|\lambda|$, where λ_2 is the eigenvalue of A of next largest magnitude. With some modifications the power method can be used when A has more than one eigenvalue of largest magnitude.

After the absolutely largest eigenvalue of A has been calculated, the power method can be applied to an appropriately deflated matrix to calculate the next largest eigenvalue and the associated eigenvector of A, and so on. However, this approach is inefficient if all or many eigenvalues are needed. The next sections describe more general-purpose methods for eigenvalue and eigenvector computations. For an introduction to the power method, see Atkinson (1989). For further reading, see Golub and Van Loan (1989) or Wilkinson (1965).

5.2 The *QR* Method

The QR method is widely used to calculate all the eigenvalues of a matrix A. It employs the QR factorization algorithm presented briefly in Sec. 4.4. Here, we recall that, given an $n \times n$ matrix A, there is a factorization $A = QR$, where R is an $n \times n$ upper-triangular matrix and Q an $n \times n$ orthogonal matrix.

The QR method, shown in Table 9, is an iterative scheme to compute the eigenvalues of A. It proceeds by generating a sequence $A_1, A_2, \ldots$ of matrices, all of which are similar to each other and to the starting matrix $A_0 = A$. The sequence converges either to a triangular matrix with the eigenvalues of A on its diagonal or to an almost triangular matrix from which the eigenvalues can be calculated easily.

For a real, nonsingular matrix A with no two (or more) eigenvalues of the same magnitude, the QR method is guaranteed to converge. The iterates A_k converge to an upper-triangular matrix with the eigenvalues of A on its diagonal. If A is symmetric, the iterates converge to a diagonal matrix. The rate of convergence depends on $\max\{|\lambda_{i+1}/\lambda_i| : i = 1, \ldots, n-1\}$, where $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$.

If two or more eigenvalues of A have the same magnitude, the sequence $A_1, A_2, \ldots$ may not converge to an upper-triangular matrix. If A is symmetric, the sequence

Tab. 9 The *QR* method

```
Set A_0 = A
for k = 1, 2, ... do
  Compute the QR factorization of
  A_{k-1} = Q_k R_k
  Set A_k = R_k Q_k
  test stopping criterion
end
```

converges to a block-diagonal matrix, with blocks of order 1 or 2, from which the eigenvalues of A can be calculated easily. If A is nonsymmetric, the problem is more complicated. See Wilkinson (1965) or Parlett (1968) for details.

In many cases, the QR method is combined with a technique known as *shifting* to accelerate convergence. A discussion of this procedure can be found in Atkinson (1989), Golub and Van Loan (1989), Parlett (1980), and Wilkinson (1965).

The QR method for eigenvalues requires the computation of the QR factorization of a matrix and a matrix–matrix multiplication at each iteration. For a large matrix A, this is an expensive computation, making the method inefficient. To improve its efficiency, the matrix A is normally preprocessed, reducing it to a simpler form. If A is symmetric, it is reduced to a similar tridiagonal matrix, as described in Sec. 5.3. If A is nonsymmetric, it is reduced to a similar upper Hessenberg matrix by a comparable algorithm. It can be shown that, if $A_0 = A$ is in Hessenberg or tridiagonal form, then all the iterates A_k generated by the QR method will also be in Hessenberg or tridiagonal form, respectively. With appropriate implementation techniques, the QR factorization and the matrix–matrix products applied to matrices with these special structures require fewer operations than would be needed for arbitrary dense matrices.

With the modifications discussed above, the QR method is an efficient, general-purpose scheme for calculating the eigenvalues of a dense matrix. The eigenvectors can also be calculated if all the similarity transformations employed in the QR process are stored. Alternatively, the eigenvectors can be calculated by inverse iteration, as described in Sec. 5.4.

5.3 Transforming a Symmetric Matrix to Tridiagonal Form

As noted above, the eigenvalues of a symmetric (or Hermitian) matrix A are often calculated by first transforming A into a similar tridiagonal matrix T. Householder transformations (see Sec. 4.3) are usually employed to perform this task.

The algorithm to obtain the matrix T, given A, resembles the QR factorization algorithm described briefly in Sec. 4.3. It proceeds in $n-2$ steps and generates a sequence $H_1, H_2, \ldots, H_{n-2}$ of Householder transformations to reduce A to tridiagonal form T. Starting with $A^{(0)} = A$, at step k for $k = 1, \ldots, n-2$, we form $A_k = H_k A_{k-1} H_k$, where H_k is chosen to zero the components $k+2$ to n of both row k and column k of A_{k-1}. T is A_{n-2}, the last matrix computed by this process. Note that $T = H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2}$ is similar to A because each H_k is symmetric and orthogonal.

It can be shown that the reduction to tridiagonal form by Householder transformations is a stable computation in the sense that the computed T is the exact T for a slightly perturbed matrix $\hat{A} = A + E$, where E is usually small relative to A. As a result, it can be shown that the eigenvalues of A and T differ very little. For a brief introduction to tridiagonal reduction, see Atkinson (1989). For further reading, see Golub and Van Loan (1989) or Wilkinson (1965). For other methods to reduce A to tridiagonal form, such as planar rotation orthogonal matrices, see Golub and Van Loan (1989).

Similar schemes can be used to reduce a nonsymmetric (or non-Hermitian) matrix to Hessenberg form.

5.4 Inverse Iteration

Inverse iteration is the standard method to calculate the eigenvectors of a matrix A, once its eigenvalues have been calculated. This scheme can be viewed as the power method (see Sec. 5.1) applied to the matrix $(A - \tilde{\lambda} I)^{-1}$ instead of A, where $\tilde{\lambda}$ is an (approximate) eigenvalue of A.

To see how inverse iteration works, let $\tilde{\lambda}$ be an approximation to a simple eigenvalue λ of A and let x be the associated eigenvector. Also let z_0 be an initial guess to some multiple of x. Inverse iteration, shown in Table 10, is an iterative method that generates a sequence of approximations $z_1, z_2, \ldots$ to some multiple of x. In the scheme shown in Table 10, z_k is normalized so that the sequence $z_1, z_2, \ldots$ converges to an eigenvector x of A satisfying $\|x\|_\infty = 1$. As noted already in Sec. 5.1, normalizations of this sort are frequently used in eigenvector calculations.

Note that, if $\tilde{\lambda} = \lambda$, then $A - \tilde{\lambda} I$ is singular. Moreover, if $\tilde{\lambda} \approx \lambda$, then $A - \tilde{\lambda} I$ is "nearly" singular. As a result, we can expect the system $(A - \tilde{\lambda} I) w_k = z_{k-1}$ to be very poorly conditioned (see Sec. 2.8). This, though, is not a problem in this context, since any large perturbation in the solution of the ill-conditioned system is in the direction of the desired eigenvector x.

The sequence of approximate eigenvectors $z_1, z_2, \ldots$ is typically calculated by first performing an LU decomposition of $A - \tilde{\lambda} I$, often with pivoting, before the start of the iteration and then using the same LU factorization to perform a forward elimination followed by a back substitution to solve $(A - \tilde{\lambda} I) w_k = z_{k-1}$ for $k = 1, 2, \ldots$ (see Sec. 2). We emphasize that only one LU factorization of $A - \tilde{\lambda} I$ is needed to solve all $(A - \tilde{\lambda} I) w_k = z_{k-1}$, $k = 1, 2, \ldots$.

For a brief discussion of the inverse iteration, including its stability and rate of convergence, see Atkinson (1989). For further reading, see Wilkinson (1965).

Tab. 10 Inverse iteration

```
Pick z_0
for k = 1, 2, ... do
    Solve (A − λ̃I)w_k = z_{k−1}
    z_k = w_k/‖w_k‖_∞
    test stopping criterion
end
```

5.5 Other Methods

Another way to calculate the eigenvalues of A is to compute the roots of the characteristic polynomial $p(\lambda) = \det(A - \lambda I)$. The techniques discussed in Sec. 6.9 can be used for this purpose. However, this approach is often less stable than the techniques described above and it is usually not more efficient. Therefore, it is not normally recommended.

An obvious method for calculating an eigenvector x, once the corresponding eigenvalue λ is known, is to solve the system $(A - \lambda I)x = 0$. Since this system is singular, one approach is to delete one equation from $(A - \lambda I)x = 0$ and replace it by another linear constraint, such as $x_j = 1$ for some component j of x. However, it can be shown [see Wilkinson (1965)] that this method is not always stable and can lead to very poor numerical results in some cases. Thus, it is not recommended either.

Several other methods for calculating the eigenvalues and eigenvectors of a matrix have been omitted from our discussion because of space limitations. We should, though, at least mention one of them, the Jacobi method, a simple, rapidly convergent iterative scheme, applicable

to symmetric matrices, including sparse ones.

The eigenvalue problem for large sparse matrices is a very active area of research. Although the QR method described in Sec. 5.2 is effective for small- to medium-sized dense matrices, it is computationally very expensive for large matrices, in part because it does not exploit sparsity effectively. The Lanczos and Arnoldi methods are much better suited for large sparse problems. For a discussion of these methods, see Cullum and Willoughby (1985), Parlett (1980), Saad (1992), and Scott (1981).

6 Nonlinear Algebraic Equations and Systems

Consider a nonlinear algebraic equation or system $f(x) = 0$, where $f : \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$. A *root* of $f(x)$ is a value $\alpha \in \mathbb{R}^n$ satisfying $f(\alpha) = 0$. A nonlinear function may have one, many, or no roots.

Most numerical methods for computing approximations to roots of a nonlinear equation or system are *iterative* in nature. That is, the scheme starts with some initial guess x_0 and computes new successive approximations x_k, $k = 1, 2, \ldots$, by some formula until a *stopping criterion* such as

$$\|f(x_k)\| \leq \varepsilon,$$
$$\|f(x_k)\| \leq \varepsilon \|f(x_0)\|,$$
$$\|x_k - x_{k-1}\| \leq \varepsilon,$$
$$\|x_k - x_{k-1}\| \leq \varepsilon \|x_k\|, \text{ or}$$
$$k > \text{maxit}$$

is satisfied, where ε is the error tolerance and maxit is the maximum number of iterations allowed.

6.1 Fixed-Point Iteration

A *fixed point* of a function $g(x)$ is a value α satisfying $\alpha = g(\alpha)$. A *fixed-point iteration* is a scheme of the form $x_k = g(x_{k-1})$ that uses the most recent approximation x_{k-1} to the fixed point α to compute a new approximation x_k to α. In this context, the function g is also called the *iteration function*.

One reason for studying fixed-point iterations is that given a function $f(x)$, it is easy to find another function $g(x)$ such that α is a root of $f(x)$ if and only if it is a fixed point of $g(x)$. For example, take $g(x) = x - f(x)$. Many root-finding methods can be viewed as fixed-point iterations.

Given an iteration function g, a fixed-point scheme starts with an initial guess x_0 and proceeds with the iteration as follows:

for $k = 1, 2, \ldots$ do
    $x_k = g(x_{k-1})$
    test stopping criterion
end

6.2 Newton's Method for Nonlinear Equations

We consider scalar equations (i.e., $n = 1$) first, and extend the results to systems of equations (i.e., $n > 1$) in Sec. 6.7.

Newton's method is a fixed-point iteration based on the iteration function $g(x) = x - f(x)/f'(x)$, where $f'(x)$ is the first derivative of f. More specifically, the new approximation x_k to the root α of f is computed by the formula

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})},$$

which uses the previous approximation x_{k-1} to α. Geometrically, Newton's method

approximates the nonlinear function $f(x)$ by its tangent (a straight line) at the current approximation x_{k-1} and takes x_k to be the intersection of the tangent with the x axis. That is, x_k is the root of the local linear approximation to $f(x)$. Newton's method is applicable if and only if f is differentiable and f' is nonzero at the point of approximation.

6.3 The Secant Method

The secant method is applicable to scalar equations only and is not a fixed-point iteration. The new approximation x_k to the root α of f is computed using two previous approximations x_{k-1} and x_{k-2} by the formula

$$x_k = x_{k-1} - f(x_{k-1})\frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}.$$

The secant method can be viewed as a variant of Newton's method in which $f'(x_{k-1})$ is approximated by $[f(x_{k-1}) - f(x_{k-2})]/[x_{k-1} - x_{k-2}]$. Geometrically, the secant method approximates the nonlinear function $f(x)$ by the chord subtending the graph of f at the two most recently computed points of approximation x_{k-1} and x_{k-2} and takes x_k to be the intersection of the chord with the x axis. That is, x_k is the root of the local linear approximation to $f(x)$. The secant method is applicable if and only if $f(x)$ is continuous and takes different values at x_{k-1} and x_{k-2}. To start, the secant method requires initial guesses for x_0 and x_1. These are usually chosen close to each other and must satisfy $f(x_0) \neq f(x_1)$.

6.4 The Bisection and Regula Falsi Methods

The bisection method is not a fixed-point iteration. It is applicable to a scalar equation $f(x) = 0$ if and only if $f(x)$ is continuous and there are two points L and R for which $f(L)f(R) \leq 0$. These conditions guarantee the existence of at least one root of $f(x)$ in the interval $[L, R]$. Without loss of generality, let $L < R$. To start, the bisection method approximates the root by the midpoint $M = (L + R)/2$ of $[L, R]$ and halves the interval at each iteration as follows.

forever do $M = (L + R)/2$ if $f(L)f(M) \leq 0$ then $R = M$ else $L = M$ test stopping criterion end

Note that this iteration maintains the property $f(L)f(R) \leq 0$, as L and R are changed. So, when the algorithm terminates, a root of f is guaranteed to be in $[L, R]$. $M = (L + R)/2$ is often taken as the approximation to the root.

Several root-finding methods are similar to bisection. For example, *regula falsi* chooses

$$M = \frac{f(R)L - f(L)R}{f(R) - f(L)} \tag{38}$$

but is otherwise the same as bisection. Note that the M computed from (38) is the intersection of the chord subtending the graph of $f(x)$ at L and R with the x axis and so is guaranteed to lie in $[L, R]$, since the property $f(L)f(R) \leq 0$ is maintained throughout the iteration even though L and R are changed.

6.5 Convergence

Iterative methods for nonlinear equations can be guaranteed to converge under certain conditions, although they may diverge in some cases.

The bisection method converges whenever it is applicable, but if $f(x)$ has more than one root in the interval of application, there is no guarantee which of the roots the method will converge to.

The convergence of a fixed-point iteration depends critically on the properties of the iteration function $g(x)$. If g is smooth in an interval containing a fixed point α and $|g'(\alpha)| < 1$, then there is an $m \in [0, 1)$ and a neighborhood I around the fixed point α in which $|g'(x)| \leq m < 1$. In this case, the fixed-point iteration $x_k = g(x_{k-1})$ converges to α if $x_0 \in I$. To give an intuitive understanding why this is so, we assume that $x_0, \ldots, x_{k-1} \in I$ and note that

$$\begin{aligned} x_k - \alpha &= g(x_{k-1}) - g(\alpha) \\ &= g'(\xi_k)(x_{k-1} - \alpha), \end{aligned}$$

where we have used $x_k = g(x_{k-1})$, $\alpha = g(\alpha)$ and, by the mean-value theorem, ξ_k is some point in I between x_{k-1} and α. Thus, $|g'(\xi_k)| \leq m < 1$ and so $|x_k - \alpha| \leq m|x_{k-1} - \alpha| \leq m^k|x_0 - \alpha|$, whence $x_k \in I$ too and $x_k \to \alpha$ as $k \to \infty$.

A more formal statement of this theorem and other similar results giving sufficient conditions for the convergence of fixed-point iterations can be found in many introductory numerical methods textbooks. See for example Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980).

Newton's method converges if the conditions for convergence of a fixed-point iteration are met. [For Newton's method, the iteration function is $g(x) = x - f(x)/f'(x)$.] However, it can be shown that Newton's method converges quadratically (see Sec. 6.6) to the root α of f if the initial guess x_0 is chosen sufficiently close to α, f is smooth, and $f'(x) \neq 0$ close to α. It can also be shown that Newton's method converges from any starting guess in some cases. A more formal statement of these and other similar results can be found in many introductory numerical methods textbooks. See for example Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980). For a deeper discussion of this topic, see Dennis and Schnabel (1983).

6.6 Rate of Convergence

The rate of convergence of a sequence $x_1, x_2, \ldots$ to α is the largest number $p \geq 1$ satisfying

$$\|x_{k+1} - \alpha\| \leq C\|x_k - \alpha\|^p \quad \text{as } k \to \infty$$

for some constant $C > 0$. If $p = 1$, we also require that $C < 1$. The larger the value of p the faster the convergence, at least asymptotically. Between two converging sequences with the same rate p, the faster is the one with the smaller C.

A fixed-point iteration with iteration function g converges at a rate p with $C = |g^{(p)}(\alpha)|/p$ if $g \in \mathscr{C}^p$, $g^{(i)}(\alpha) = 0$, for $i = 0, 1, 2, \ldots, p-1$, and $g^{(p)}(\alpha) \neq 0$, where $g^{(i)}(x)$ is the ith derivative of $g(x)$. Thus, Newton's method usually converges quadratically, i.e., $p = 2$ with $C = |g''(\alpha)|/2$, where $g(x) = x - f(x)/f'(x)$. If $f'(\alpha) = 0$, Newton's method typically converges linearly. If $f'(\alpha) \neq 0$ and $f''(\alpha) = 0$, Newton's method converges at least cubically, i.e., $p \geq 3$. The secant method converges at a superlinear rate of $p = (1 + \sqrt{5})/2 \approx 1.618$, i.e., faster than linear but slower than quadratic. Bisection converges linearly, i.e., $p = 1$ and $C = 1/2$.

6.7 Newton's Method for Systems of Nonlinear Equations

Newton's method for a scalar nonlinear equation (see Sec. 6.2) can be extended to a system of nonlinear equations with $f: \mathbb{R}^n \to \mathbb{R}^n$, for $n > 1$. In this case, the

new iterate x_k is computed by

$$x_k = x_{k-1} - [J(x_{k-1})]^{-1} f(x_{k-1}),$$

where $J(x)$ is the *Jacobian* of f, an $n \times n$ matrix with its (i, j) entry equal to $\partial f_i(x)/\partial x_j$. To implement this scheme effectively, the matrix–vector product $z = [J(x_{k-1})]^{-1} f(x_{k-1})$ should *not* be calculated by first computing the inverse of the Jacobian and then performing the matrix–vector multiplication, but rather by first solving the linear system $[J(x_{k-1})]z_k = f(x_{k-1})$ by Gaussian elimination (see Sec. 2.1) or some other effective method for solving linear equations (see Secs. 2 and 3) and then setting $x_k = x_{k-1} - z_k$.

6.8 Modifications and Alternatives to Newton's Method

Solving the linear system $[J(x_{k-1})]z_k = f(x_{k-1})$ requires first evaluating all the partial derivatives of all components of f at the point x_{k-1} and then solving the linear system by performing Gaussian elimination (see Sec. 2.1) or some other effective method for solving linear equations (see Secs.2 and 3). In the unmodified Newton's method, this procedure is repeated at every iteration, requiring $O(n^3)$ flops (floating-point operations) if Gaussian elimination is used. There exist variants of Newton's method that reduce the computational work per iteration significantly. Even though these schemes typically converge more slowly, they often dramatically reduce the cost of solving a nonlinear system, particularly if $n \gg 1$.

The *chord Newton* method, often called the *simplified Newton method*, holds $J(x_{k-1})$ fixed for several steps, thus avoiding many Jacobian evaluations and LU factorizations. However, it still requires one f evaluation and both a forward elimination and a back substitution (see Secs. 2.4 and 2.2, respectively) at each iteration.

Some other variants approximate the Jacobian by matrices that are easier to compute and simpler to solve. For example, the Jacobian may be approximated by its diagonal, giving rise to a *Jacobi-like Newton's* iteration, or by its lower-triangular part, giving rise to a *Gauss–Seidel-like Newton's* scheme. (See Sec. 3 for a discussion of Jacobi and Gauss–Seidel iterations for linear equations.)

Quasi-Newton schemes, which avoid the computation of partial derivatives, are alternatives to Newton's method. Some quasi-Newton schemes approximate partial derivatives by finite differences. For example,

$$[J(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x) \approx \frac{f_i(x + \delta e_j) - f_i(x)}{\delta},$$

where δ is a small nonzero number and $e_j \in \mathbb{R}^n$ is the jth unit vector, with a 1 in component j and 0's in all other entries.

Possibly the best-known quasi-Newton scheme is *Broyden's* method. It does not require the computation of any partial derivatives nor the solution of any linear systems. Rather, it uses one evaluation only of f and a matrix–vector multiply, requiring $O(n^2)$ flops, per iteration. Starting with an initial guess for the inverse of the Jacobian, $J(x_0)$, it updates its approximation to the inverse Jacobian at every iteration. Broyden's method can be viewed as an extension of the secant method to $n > 1$ dimensions. For a brief description of the algorithm, see Hager (1988).

6.9 Polynomial Equations

Polynomial equations are a special case of nonlinear equations. A polynomial of

degree k has exactly k roots, counting multiplicity. One may wish to compute all roots of a polynomial or only a few select ones. The methods for nonlinear equations described above can be used in either case, although more effective schemes exist for this special class of problems. The efficient evaluation of the polynomial and its derivative is discussed below in Sec. 6.10.

Deflation is often used to compute roots of polynomials. It starts by locating one root r_1 of $p(x)$ and then proceeds recursively to compute the roots of $\hat{p}(x)$, where $p(x) = (x - r_1)\hat{p}(x)$. Note that, by the fundamental theorem of algebra, given a root r_1 of $p(x)$, $\hat{p}(x)$ is uniquely defined and is a polynomial of degree $k - 1$. However, deflation may be unstable unless implemented carefully. See an introductory numerical methods textbook for details.

Localization techniques can be used to identify regions of the complex plane that contain zeros. Such techniques include the Lehmer–Schur method, Laguerre's method, and methods based on Sturm sequences [see Householder (1970)]. Localization can be very helpful, for example, when searching for a particular root of a polynomial or when implementing deflation, since in the latter case, the roots should be computed in increasing order of magnitude to ensure numerical stability.

The roots of a polynomial $p(x)$ can also be found by first forming the *companion matrix* A of the polynomial $p(x)$ – the eigenvalues $\{\lambda_i : i = 1, \ldots, n\}$ of A are the roots of $p(x)$ – and then finding all, or a select few, of the eigenvalues of A. See Sec. 5 for a discussion of the computation of eigenvalues and an introductory numerical methods book, such as Hager (1988), for a further discussion of this root-finding technique.

6.10 Horner's Rule

Horner's rule, also called *nested multiplication* or *synthetic division*, is an efficient method to evaluate a polynomial and its derivative. Let $p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ be the polynomial of interest, and α the point of evaluation. The following algorithm computes $p(\alpha)$ in z_0 and $p'(\alpha)$ in y_1 efficiently:

$$
\begin{aligned}
&z_n = a_n \\
&y_n = a_n \\
&\text{for } j = n-1, \ldots, 1 \text{ do} \\
&\qquad z_j = \alpha z_{j+1} + a_j \\
&\qquad y_j = \alpha y_{j+1} + z_j \\
&\text{end} \\
&z_0 = z_1 \alpha + a_0
\end{aligned}
$$

7 Unconstrained Optimization

The *optimization problem* is to find a value $x^* \in \mathbb{R}^n$ that either *minimizes* or *maximizes* a function $f: \mathbb{R}^n \to \mathbb{R}$. We consider only the minimization problem here, since maximizing $f(x)$ is equivalent to minimizing $-f(x)$.

Sometimes the minimizer must satisfy constraints such as $g_i(x) = 0, i = 1, \ldots, m_1$, or $h_j(x) \geq 0, j = 1, \ldots, m_2$, where g_i and $h_j : \mathbb{R}^n \to \mathbb{R}$. Thus, the general minimization problem can be written as

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to

$$
\begin{aligned}
g_i(x) &= 0, \quad i = 1, \ldots, m_1, \\
h_j(x) &\geq 0, \quad j = 1, \ldots, m_2.
\end{aligned}
$$

If any of the functions f, g_i, or h_j are nonlinear then the minimization problem is *nonlinear*; otherwise, it is called a *linear programming* problem. If there are no constraints, the minimization problem is called *unconstrained*; otherwise, it is called *constrained*.

In this section, we present numerical methods for nonlinear unconstrained minimization problems (NLUMPs) only. For a more detailed discussion of these schemes, see an advanced text such as Dennis and Schnabel (1983).

7.1 Some Definitions and Properties

Let $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. The *gradient* ∇f of the function f is the vector of the n first partial derivatives of f:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)^{\mathrm{T}}.$$

Note $\nabla f(x) : \mathbb{R}^n \to \mathbb{R}^n$.

A point x^* is a *critical point* of f if $\nabla f(x^*) = 0$. A point x^* is a *global minimum* of f if $f(x^*) \le f(x)$ for all $x \in \mathbb{R}^n$. A point x^* is a *local minimum* of f if $f(x^*) \le f(x)$ for all x in a neighborhood of x^*. If x^* is a local minimum of f, then it is also a critical point of f, assuming $\nabla f(x^*)$ exists, but the converse is not necessarily true.

The *Hessian* $\nabla^2 f(x) = H(x)$ of f is an $n \times n$ matrix with entries

$$H_{ij}(x) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

If f is smooth, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i},$$

and so the Hessian is symmetric. A critical point x^* is a local minimum of f if $H(x^*)$ is symmetric positive-definite.

Numerical methods for solving NLUMPs are all iterative in character. Some try to compute roots of the gradient, i.e., critical points of f, which are also local minima. These methods are related to schemes used to solve nonlinear equations described in Sec. 6.

Minimization methods can be classified into three main categories:

- *direct-search methods*, which make use of function values only,
- *gradient methods*, which make use of function and derivative values, and
- *Hessian methods*, which make use of function, derivative, and second-derivative values.

Methods from each class are discussed below.

7.2 The Fibonacci and Golden-Section Search Methods

The Fibonacci and golden-section search methods belong to the class of direct-search methods. They are similar to the bisection method for nonlinear equations described in Sec. 6.4, although important differences exist. They are used in one-dimensional minimization only.

The Fibonacci and golden-section search methods are applicable if $f(x)$ is continuous and unimodal in the interval of interest $[a, b]$, where by *unimodal* we mean that $f(x)$ has exactly one local minimum $x^* \in [a, b]$ and $f(x)$ strictly decreases for $x \in [a, x^*]$ and strictly increases for $x \in [x^*, b]$.

The main idea behind these methods is the following. Let x_1 and x_2 be two points satisfying $a < x_1 < x_2 < b$. If $f(x_1) \ge f(x_2)$, then x^* must lie in $[x_1, b]$.

On the other hand, if $f(x_1) \leq f(x_2)$ then x^* must lie in $[a, x_2]$. Thus, by evaluating f at a sequence of test points $x_1, x_2, \ldots, x_k$, we can successively reduce the length of the interval in which we know the local minimum lies, called the *interval of uncertainty*.

The Fibonacci and golden-section search methods differ only in the way the sequence of test points is chosen. Before describing the two methods, we give two more definitions.

The *coordinate* v of a point x *relative to an interval* $[a, b]$ is $v = (x - a)/(b - a)$.

The *Fibonacci numbers* are defined by the initial condition $F_0 = F_1 = 1$ and the recurrence $F_k = F_{k-1} + F_{k-2}$ for $k \geq 2$.

The Fibonacci search method applied to a function f on an interval $[a, b]$ starts with two points x_k and x_{k-1} satisfying $(x_k - a)/(b - a) = F_{k-1}/F_k$ and $(x_{k-1} - a)/(b - a) = F_{k-2}/F_k = 1 - F_{k-1}/F_k$, whence $a < x_{k-1} < x_k < b$. It then computes the sequence $x_{k-2}, x_{k-3}, \ldots, x_1$ as follows. After evaluating $f(x_k)$ and $f(x_{k-1})$, the interval of uncertainty is either $[a, x_k]$ or $[x_{k-1}, b]$.

- If the interval of uncertainty is $[a, x_k]$, then x_{k-1} belongs to it and $(x_{k-1} - a)/(x_k - a) = F_{k-2}/F_{k-1}$. In this case, x_{k-2} is chosen to satisfy $(x_{k-2} - a)/(x_k - a) = F_{k-3}/F_{k-1}$. The method proceeds to the next iteration with the two points x_{k-1} and x_{k-2} in the interval $[a, x_k]$, with relative coordinates F_{k-2}/F_{k-1} and F_{k-3}/F_{k-1}, respectively. Note that $f(x_{k-1})$ is already computed, so that only $f(x_{k-2})$ needs to be evaluated in the next iteration.
- If the interval of uncertainty is $[x_{k-1}, b]$, then x_k belongs to it and $(x_k - x_{k-1})/(b - x_{k-1}) = F_{k-3}/F_{k-1}$. In this case, x_{k-2} is chosen to satisfy $(x_{k-2} - x_{k-1})/(b - x_{k-1}) = F_{k-2}/F_{k-1}$. The method proceeds to the next iteration with the two points x_k and x_{k-2} in the interval $[x_{k-1}, b]$, with relative coordinates F_{k-3}/F_{k-1} and F_{k-2}/F_{k-1}, respectively. Note that $f(x_k)$ is already computed, so that only $f(x_{k-2})$ needs to be evaluated in the next iteration.

Thus, at the start of the second iteration, the situation is similar to that at the start of the first iteration, except that the length of the interval of uncertainty has been reduced. Therefore, the process described above can be repeated.

Before the last iteration, the interval of uncertainty is $[c, d]$ and x_1 is chosen to be $(c + d)/2 + \in$, for some small positive number $\in$.

As noted above, the Fibonacci search method requires one function evaluation per iteration. The main disadvantage of the method is that the number of iterations k must be chosen at the start of the method. However, it can be proved that, given k, the length of the final interval of uncertainty is the shortest possible. Thus, in a sense, it is an optimal method.

The golden-section search method also requires one function evaluation per iteration, but the number of iterations k does not need to be chosen at the start of the method. It produces a sequence of test points $x_1, x_2, \ldots$ and stops when a predetermined accuracy is reached.

Let $r = (\sqrt{5} - 1)/2 \approx 0.618$ be the positive root of the quadratic $x^2 + x - 1$. Note that $1/r = (\sqrt{5} + 1)/2 \approx 1.618$ is the famous *golden ratio*. For the Fibonacci search method, it can be proved that, for large k, the coordinates of the two initial points x_k and x_{k-1} relative to $[a, b]$ are approximately r and $1 - r$, respectively. Thus, if, at each iteration, the two points are chosen with these coordinates relative to the interval of uncertainty, the resulting method, called the golden-section search method, is

an approximation to the Fibonacci search. Moreover, if a point has coordinate $1-r$ relative to $[a, b]$, then it has coordinate r relative to $[a, a+(b-a)r]$. Similarly, if a point has coordinate r relative to $[a, b]$, then it has coordinate $1-r$ relative to $[a+(b-a)(1-r), b]$. This property is exploited in the golden-section search method, enabling it to use one function evaluation only per iteration.

Both methods described above are guaranteed to converge whenever they are applicable and both have a linear rate of convergence (see Sec. 6.6). On the average, the length of the interval of uncertainty is multiplied by r at each iteration. For a further discussion of these methods, see an introductory numerical methods text such as Kahaner et al. (1989).

7.3 The Steepest-Descent Method

The steepest-descent (SD) method belongs to the class of gradient schemes. It is applicable whenever the partial derivatives of f exist and f has at least one local minimum. Whenever it is applicable, it is guaranteed to converge to some local minimum if the partial derivatives of f are continuous. However, it may converge slowly for multidimensional problems. Moreover, if f possesses more than one local minimum, there is no guarantee to which minimum SD will converge.

At iteration k, the SD method performs a search for the minimum of f along the line $x_k - \alpha\nabla f(x_k)$, where α is a scalar variable and $-\nabla f(x_k)$ is the direction of the steepest descent of f at x_k. Note that, since α is scalar, minimizing $f(x_k - \alpha\nabla f(x_k))$ with respect to (w.r.t.) α is a one-dimensional minimization problem. If α^* is the minimizer, x_{k+1} is taken to be $x_k - \alpha^*\nabla f(x_k)$. A brief outline of SD is given in Table 11. See Buchanan and Turner (1992), Johnson and Riess (1982), or Ortega (1988) for further details.

Tab. 11 The steepest-descent method

Pick an initial guess x_0 and a tolerance $\in$
for $k = 1, \ldots,$ maxit do
  $s_{k-1} = -\nabla f(x_{k-1})$
  if $\|s_{k-1}\| \leq \in$ exit loop
  find $\alpha^* \in \mathbb{R}$ that minimizes $f(x_{k-1} + \alpha s_{k-1})$
  $x_k = x_{k-1} + \alpha^* s_{k-1}$
end

7.4 Conjugate-Direction Methods

The definition of conjugate directions is given w.r.t. a symmetric positive-definite (SPD) matrix A: the vectors (directions) u and v are A-conjugate if $u^{\mathrm{T}}Av = 0$. Thus, conjugate directions are orthogonal or perpendicular directions w.r.t. an inner product $(u, v) = u^{\mathrm{T}}Av$ and, as such, are often associated with some shortest-distance property.

The conjugate-direction (CD) methods form a large class of minimization schemes. Their common characteristic is that the search direction at every iteration is conjugate to previous search directions. Proceeding along conjugate search directions guarantees, in some sense, finding the shortest path to the minimum.

The CD methods are guaranteed to converge in at most n iterations for the SPD quadratic function $f(x) = c + b^{\mathrm{T}}x - \frac{1}{2}x^{\mathrm{T}}Ax$, where A is an $n \times n$ SPD matrix.

There are several techniques to construct conjugate directions, each one giving rise to a different CD method. The best-known is *Powell's* method [see Buchanan and Turner (1992) or Ortega (1988)].

Tab. 12 The conjugate-gradient (CG) method

```
Pick an initial guess x_0 and a tolerance ∈
Initialize s_0 = 0 and β = 1
for k = 1, ..., maxit do
    if ‖∇f(x_{k-1})‖ ≤∈ exit loop
    s_k = −∇f(x_{k-1}) + β s_{k-1}
    find α* ∈ ℝ that minimizes f(x_{k-1} + α s_k)
    x_k = x_{k-1} + α* s_k
    β = ‖∇f(x_k)‖²/‖∇f(x_{k-1})‖²
end
```

7.5 The Conjugate-Gradient Method

As the name implies, at each iteration, the conjugate-gradient (CG) method takes information from the gradient of f to construct conjugate directions. Its search direction is a linear combination of the direction of steepest descent and the search direction of the previous iteration. A brief outline of CG is given in Table 12.

The CG method is guaranteed to converge in at most n iterations for the SPD quadratic function $f(x) = c + b^{\mathrm{T}}x - \frac{1}{2}x^{\mathrm{T}}Ax$, where A is a $n \times n$ SPD matrix. See Sec. 3.2, Golub and Van Loan (1989), or Ortega (1988) for a more detailed discussion of this minimization technique applied to solve linear algebraic systems.

There exist several variants of the CG method, most of which are based on slightly different ways of computing the step size β.

7.6 Newton's Method

Newton's method for minimizing f is just Newton's method for nonlinear systems applied to solve $\nabla f(x) = 0$. The new iterate x_k is computed by $x_k = x_{k-1} - [H(x_{x-1})]^{-1} \times \nabla f(f_{k-1})$, where $H(x_{k-1})$ is the Hessian of f at x_{k-1}. See Sec. 6.7 for further details.

If f is convex, then the Hessian is SPD and the search direction generated by Newton's method at each iteration is a descent (downhill) direction. Thus, for any initial guess x_0, Newton's method is guaranteed to converge quadratically, provided f is sufficiently smooth.

For a general function f, there is no guarantee that Newton's method will converge for an arbitrary initial guess x_0. However, if started close enough to the minimum of a sufficiently smooth function f, Newton's method normally converges quadratically, as noted in Sec. 6.6. There exist several variants of Newton's method that improve upon the reliability of the standard scheme.

7.7 Quasi-Newton Methods

At every iteration, Newton's method requires the evaluation of the Hessian and the solution of a linear system $[H(x_{k-1})]s_k = -\nabla f(x_{k-1})$ for the search direction s_k. Quasi-Newton methods update an approximation to the inverse Hessian at every iteration, thus reducing the task of solving a linear system to a simple matrix–vector multiply. The best-known of these schemes are the *Davidon–Fletcher–Powell* (DFP) and the *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) methods. See Buchanan and Turner (1992) or Dennis and Schnabel (1983) for further details.

8 Approximation

It is often desirable to find a function $f(x)$ in some class that approximates the data $\{(x_i, y_i) : i = 1, \ldots, n\}$. That is,

$f(x_i) \approx y_i$, $i = 1, \ldots, n$. If f matches the data exactly, that is, f satisfies the *interpolation relations* or *interpolation conditions* $f(x_i) = y_i$, $i = 1, \ldots, n$, then f is called the *interpolating function* or the *interpolant* of the given data.

Similarly, it is often desirable to find a simple function $f(x)$ in some class that approximates a more complex function $y(x)$. We say that f *interpolates* y, or f is the *interpolant* of y, at the points x_i, $i = 1, \ldots, n$, if $f(x_i) = y(x_i)$, $i = 1, \ldots, n$. The problem of computing the interpolant f of y at the points x_i, $i = 1, \ldots, n$, reduces to the problem of computing the interpolant f of the data $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $y_i = y(x_i)$.

An interpolant does not always exist and, when it does, it is not necessarily unique. However, a unique interpolant does exist in many important cases, as discussed below.

A standard approach for constructing an interpolant is to choose a set of *basis functions* $\{b_1(x), b_2(x), \ldots, b_n(x)\}$ and form a *model*

$$f(x) = \sum_{j=1}^{n} a_j b_j(x),$$

where the numbers a_j are unknown coefficients. For f to be an interpolant, it must satisfy $f(x_i) = y_i$, $i = 1, \ldots, n$, which is equivalent to

$$\sum_{j=1}^{n} a_j b_j(x_i) = y_i, \quad i = 1, \ldots, n.$$

These n conditions form a system $B\mathbf{a} = \mathbf{y}$ of n linear equations in n unknowns, where $\mathbf{a} = (a_1, a_2, \ldots, a_n)^{\mathrm{T}}$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, and $B_{ij} = b_j(x_i)$, $i, j = 1, \ldots, n$. If B is nonsingular, then the interpolant of the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ w.r.t. the basis functions $b_1(x), b_2(x), \ldots, b_n(x)$ exists and is unique. On the other hand, if B is singular, then either the interpolant may fail to exist or there may be infinitely many interpolants.

8.1 Polynomial Approximation

Polynomial approximation is the foundation for many numerical procedures. The basic idea is that, if we want to apply some procedure to a function, such as integration (see Sec. 9), we approximate the function by a polynomial and apply the procedure to the approximating polynomial. Polynomials are often chosen as approximating functions because they are easy to evaluate (see Sec. 6.10), to integrate, and to differentiate. Moreover, polynomials approximate well more complicated functions, provided the latter are sufficiently smooth. The following mathematical result ensures that arbitrarily accurate polynomial approximations exist for a broad class of functions.

WEIERSTRASS THEOREM: If $g(x) \in \mathscr{L}[a, b]$, then, for every $\in > 0$, there exists a polynomial $p_n(x)$ of degree $n = n(\in)$ such that $\max\{|g(x) - p_n(x)| : x \in [a, b]\} \leq \in$.

8.2 Polynomial Interpolation

Techniques to construct a polynomial interpolant for a set of data $\{(x_i, y_i) : i = 1, \ldots, n\}$ are discussed below. Here we state only the following key result.

THEOREM: If the points $\{x_i : i = 1, \ldots, n\}$ are distinct, then there exists a unique polynomial of degree at most $n - 1$ that interpolates the data $\{(x_i, y_i) : i =$

$1, \ldots, n\}$. (There are no restrictions on the y_i's.)

8.2.1 Monomial Basis

One way to construct a polynomial that interpolates the data $\{(x_i, y_i) : i = 1, \ldots, n\}$ is to choose as basis functions the monomials $b_j(x) = x^{j-1}, j = 1, \ldots, n$, giving rise to the model $p_{n-1}(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_n x^{n-1}$. As noted above, the interpolation conditions take the form $B\mathbf{a} = \mathbf{y}$. In this case, B is the *Vandermonde matrix* for which $B_{ij} = x_i^{j-1}$, $i, j = 1, \ldots, n$, where we use the convention that $x^0 = 1$ for all x.

It can be shown that the Vandermonde matrix B is nonsingular if and only if the points $\{x_i : i = 1, \ldots, n\}$ are distinct. If B is nonsingular, then, of course, we can solve the system $B\mathbf{a} = \mathbf{y}$ to obtain the coefficients $a_j, j = 1, \ldots, n$, for the unique interpolant of the data.

It can also be shown that, although the Vandermonde matrix is nonsingular for distinct points, it can be ill-conditioned, particularly for large n. As a result, the methods described below are often computationally much more effective than the scheme described here.

8.2.2 Lagrange Basis

An alternative to the monomial basis functions discussed above is the Lagrange basis polynomials

$$b_j(x) = l_j(x) = \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} \; j = 1, \ldots, n,$$

which are of degree $n - 1$ and satisfy

$$b_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Hence $B = I$ and the system $B\mathbf{a} = \mathbf{y}$ of interpolation conditions has the obvious unique solution $a_j = y_j, j = 1, \ldots, n$.

Note that changing the basis from monomials to Lagrange polynomials does not change the resulting interpolating polynomial $p_{n-1}(x)$, since the interpolant is unique. It only affects the representation of $p_{n-1}(x)$.

8.2.3 Newton Basis and Divided Differences

Another useful basis is the set of Newton polynomials

$$b_j(x) = \prod_{i=1}^{j-1} (x - x_i) \quad j = 1, \ldots, n.$$

The coefficients a_j of the interpolating polynomial written with the Newton basis are relatively easy to compute by a recursive algorithm using *divided differences*. Before describing this form of the interpolating polynomial, though, we must introduce divided differences.

Given a function f with $f(x_i) = y_i$, $i = 1, \ldots, n$, define the divided difference with one point by

$$f[x_i] = y_i \quad i = 1, \ldots, n.$$

If $x_{i+1} \neq x_i$, define the divided difference with two points by

$$f[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \quad i = 1, \ldots, n-1.$$

If $x_i \neq x_{i+k}$, define the divided difference with $k + 1$ points by

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \quad i = 1, \ldots, n-k.$$

We can extend this definition of divided differences to sets $\{x_i : i = 1, \ldots, n\}$ with repeated values by noting that

$$\lim_{x_{i+1} \to x_i} f[x_{i+1}, x_i] = \lim_{x_{i+1} \to x_i} \frac{y_{i+1} - y_i}{x_{i+1} - x_i} = f'(x_i).$$

So, if $x_i = x_{i+1}$, we define $f[x_i, x_{i+1}] = f'(x_i)$. Similarly, it can be shown that

$$\lim_{\substack{x_{i+1} \to x_i \\ \vdots \\ x_{i+k} \to x_i}} f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

So, if $x_i = x_{i+1} = \cdots = x_{i+k}$, we define $f[x_i, x_{i+1}, \ldots, x_{i+k}] = f^{(k)}(x_i)/k!$.

Using divided differences, the coefficients a_j can be computed as $a_j = f[x_1, \ldots, x_j]$. An advantage of the divided-difference form is that it extends easily to the interpolation of derivatives as well as function values, as discussed below. Another advantage is that if the coefficients a_j, $j = 1, \ldots, n$, are computed from n data points, it is easy to add more data points and construct a higher-degree interpolating polynomial without redoing the whole computation, since the new data can be added easily to the existing divided-difference table and the interpolating polynomial extended.

8.3 Polynomial Interpolation with Derivative Data

In this section, we consider *Hermite* or *osculatory interpolation*, which requires that a function and its first derivative be interpolated at the points $\{x_i : i = 1, \ldots, n\}$. The key result is stated below.

THEOREM: Given the data $\{(x_i, y_i, y_i') : i = 1, \ldots, n\}$, where the points $\{x_i : i = 1, \ldots, n\}$ are distinct, there exists a unique polynomial interpolant $p_{2n-1}(x)$ of degree at most $2n - 1$ that satisfies $p_{2n-1}(x_i) = y_i$, $i = 1, \ldots, n$, and $p'_{2n-1}(x_i) = y'_i$, $i = 1, \ldots, n$.

The techniques used to construct such a polynomial are similar to those described in Sec. 8.2. The following choices of basis functions are often used.

1. Monomials: $b_j(x) = x^{j-1}$, $j = 1, \ldots, 2n$;
2. Generalized Lagrange basis polynomials: $b_j(x) = [1 - 2(x - x_j)l'_j(x_j)][l_j(x)]^2$, $j = 1, \ldots, n$, and $b_{n+j}(x) = (x - x_j)[l_j(x)]^2$, $j = 1, \ldots, n$;
3. Newton basis polynomials:

$$b_j(x) = \prod_{i=1}^{j-1} (x - x_i)^2, \quad j = 1, \ldots, n,$$

and

$$b_{n+j}(x) = \left(\prod_{i=1}^{j-1} (x - x_i)^2 \right) (x - x_j), \quad j = 1, \ldots, n.$$

More general forms of polynomial interpolants are discussed in some numerical methods books. See for example Davis (1975).

8.4 The Error in Polynomial Interpolation

Two key results for the error in polynomial interpolation are given in the following two theorems. For their proofs, see an introductory numerical methods text such as Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980). Before stating the results, though, we must define $\mathrm{spr}[x, x_1, x_2, \ldots, x_n]$ to be the smallest interval containing $x, x_1, x_2, \ldots, x_n$.

THEOREM: Let $g(x) \in \mathscr{L}^n$, and let $p_{n-1}(x)$ be the polynomial of degree at most $n-1$ that interpolates $g(x)$ at the n distinct points $x_1, x_2, \ldots, x_n$. Then, for any x,

$$g(x) - p_{n-1}(x) = \frac{g^{(n)}(\xi)}{n!} \prod_{i=1}^{n} (x - x_i),$$

where $g^{(n)}(x)$ is the nth derivative of $g(x)$ and ξ is some point in $\mathrm{spr}[x, x_1, x_2, \ldots, x_n]$.

THEOREM: Let $g(x) \in \mathscr{L}^{2n}$, and let $p_{2n-1}(x)$ be the polynomial of degree at most $2n-1$ that interpolates $g(x)$ and $g'(x)$ at the n distinct points $x_1, x_2, \ldots, x_n$. Then, for any x,

$$g(x) - p_{2n-1}(x) = \frac{g^{(2n)}(\xi)}{(2n)!} \prod_{i=1}^{n} (x - x_i)^2,$$

where $g^{(2n)}(x)$ is the $2n$th derivative of $g(x)$ and ξ is some point in $\mathrm{spr}[x, x_1, x_2, \ldots, x_n]$.

Note that there is a close relationship between the error in polynomial interpolation and the error in a Taylor series. As a result, polynomial interpolation is normally effective if and only if g can be approximated well by a Taylor series.

More specifically, the polynomial interpolation error can be large if the derivative appearing in the error formula is big or if $\mathrm{spr}[x, x_1, x_2, \ldots, x_n]$ is big, particularly if the point x of evaluation is close to an end point of $\mathrm{spr}[x_1, x_2, \ldots, x_n]$ or outside this interval.

8.5 Piecewise Polynomials and Splines

Given a set of *knots* or *grid points* $\{x_i : i = 1, \ldots, n\}$ satisfying $a = x_0 < x_1 < \cdots < x_n = b$, $s(x)$ is a *piecewise polynomial* (PP) of degree N w.r.t. the knots $\{x_i, i = 0, \ldots, n\}$ if $s(x)$ is a polynomial of degree N on each interval (x_{i-1}, x_i), $i = 1, \ldots, n$. A polynomial of degree N is always a PP of degree N, but the converse is not necessarily true.

A *spline* $s(x)$ of degree N is a PP of degree N. The term spline usually implies the continuity of $s(x)$, $s'(x), \ldots, s^{(N-1)}(x)$ at the knots $\{x_0, x_1, \ldots, x_n\}$. In this case, $s \in \mathscr{L}^{N-1}$, the space of continuous functions with $N-1$ continuous derivatives. Sometimes, though, the terms PP and spline are used interchangeably.

Let $s(x)$ be a PP of degree N, and assume that $s(x), s'(x), \ldots, s^{(K)}(x)$ are continuous at the knots $\{x_0, x_1, \ldots, x_n\}$. Since $s(x)$ is a polynomial of degree N on each of the n subintervals (x_i, x_{i-1}), $i = 1, \ldots, n$, it is defined by $D = n(N+1)$ coefficients. To determine these coefficients, we take into account the continuity conditions that s and its K derivatives must satisfy at the $n-1$ interior knots $\{x_1, \ldots, x_{n-1}\}$. There are $K+1$ such conditions at each interior knot, giving rise to $C = (n-1)(K+1)$ conditions. Thus, we need $M = D - C = n(N - K) + K + 1$ properly chosen additional conditions to determine the coefficients of s.

In the following we give examples of PPs and splines and their associated basis functions.

8.5.1 Constant Splines

The constant PP model function

$$\phi(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

is a constant spline w.r.t. the knots $\{0, 1\}$. Note that it is not continuous at the

knots, so that $\phi \in \mathscr{L}^{-1}$. The constant PP functions

$$\phi_i(x) = \phi\left(\frac{(x-a)}{h-i+1}\right), \quad i = 1, \ldots, n,$$

are constant splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$, and form a set of basis functions for the space of constant splines w.r.t. these knots.

8.5.2 Linear Splines

The linear PP model function

$$\phi(x) = \begin{cases} x & \text{for } 0 \le x \le 1, \\ 2 - x & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

is a linear spline w.r.t. the knots $\{0, 1, 2\}$. Note that ϕ is continuous, but ϕ' does not exist at the knots, so that $\phi \in \mathscr{L}^0$. The linear PP functions

$$\phi_i(x) = \phi\left(\frac{(x-a)}{h-i+1}\right), \quad i = 0, \ldots, n, \tag{39}$$

are linear splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$, and form a set of basis functions for the space of linear splines w.r.t. these knots.

8.5.3 Quadratic Splines

The quadratic PP model function

$$\phi(x) = \begin{cases} \dfrac{x^2}{2} & \text{for } 0 \le x \le 1, \\ \dfrac{[x^2 - 3(x-1)^2]}{2} & \text{for } 1 \le x \le 2, \\ \dfrac{[x^2 - 3(x-1)^2 + 3(x-2)^2]}{2} & \text{for } 2 \le x \le 3, \\ 0 & \text{elsewhere,} \end{cases}$$

is a quadratic spline w.r.t. the knots $\{0, 1, 2, 3, \}$. Note that ϕ and ϕ' are continuous, but ϕ'' does not exist at the knots, so that $\phi \in \mathscr{L}^1$. The quadratic PP functions

$$\phi_i(x) = \phi\left[\frac{(x-a)}{h-i+2}\right], \quad i=0, \ldots, n+1 \tag{40}$$

are quadratic splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$, and form a set of basis functions for the space of quadratic splines w.r.t. these knots.

8.5.4 Quadratic Piecewise Polynomials

The quadratic PP model functions

$$\phi(x) = \begin{cases} x(2x-1) & \text{for } 0 \le x \le 1, \\ (x-2)(2x-3) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

$$\psi(x) = \begin{cases} -4x(x-1) & \text{for } 0 \le x \le 1, \\ 0 & \text{elsewhere,} \end{cases}$$

are continuous, but neither ϕ' nor ψ' exists at their knots $\{0, 1, 2\}$ and $\{0, 1\}$, respectively. So ϕ and $\psi \in \mathscr{L}^0$. The functions

$$\phi_i(x) = \begin{cases} \phi((x-a)/h - i/2 + 1) \\ \qquad \text{for } i \text{ even,} \\ \psi((x-a)/h - (i+1)/2 + 1) \\ \qquad \text{for } i \text{ odd,} \\ i = 0, \ldots, 2n \end{cases}$$

are quadratic PPs w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$. Note that ϕ_i is continuous, but ϕ_i' does not exist at the knots, and so $\phi_i \in \mathscr{L}^0$. These functions form a basis for the space of quadratic PPs in $\mathscr{L}^0$ w.r.t. these knots.

8.5.5 Cubic Splines

The cubic PP model function

$$\phi(x) = \begin{cases} x^3/6 & \text{for } 0 \le x \le 1, \\ \dfrac{[x^3 - 4(x-1)^3]}{6} & \text{for } 1 \le x \le 2, \\ \dfrac{[x^3 - 4(x-1)^3 + 6(x-2)^3]}{6} & \text{for } 2 \le x \le 3, \\ \dfrac{[x^3 - 4(x-1)^3 + 6(x-2)^3 - 4(x-3)^3]}{6} & \text{for } 3 \le x \le 4, \\ 0 & \text{elsewhere,} \end{cases}$$

is a cubic spline w.r.t. the knots $\{0, 1, 2, 3, 4\}$. Note that ϕ, ϕ', and ϕ'' are continuous, but ϕ''' does not exist at the knots, and so $\phi \in \mathscr{L}^2$. The cubic PP functions

$$\phi_i(x) = \phi\left(\frac{(x-a)}{h-i+2}\right), \quad i = -1, \ldots, n+1 \tag{41}$$

are cubic splines w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$, and form a set of basis functions for the space of cubic splines w.r.t. these knots.

8.5.6 Cubic Hermite Piecewise Polynomials

The cubic PP model functions

$$\phi(x) = \begin{cases} x^2(3-2x) & \text{for } 0 \le x \le 1, \\ (2-x)^2(2x-1) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

$$\psi(x) = \begin{cases} x^2(x-1) & \text{for } 0 \le x \le 1, \\ (2-x)^2(x-1) & \text{for } 1 \le x \le 2, \\ 0 & \text{elsewhere,} \end{cases}$$

are in $\mathscr{L}^1$ since ϕ, ϕ', ψ, and ψ' are all continuous at the knots $\{0, 1, 2\}$, but neither ϕ'' nor ψ'' exists at the knots. The functions

$$\phi_i(x) = \begin{cases} \phi\left(\dfrac{(x-a)}{h - i/2 + 1}\right) \\ \qquad \text{for } i \text{ even,} \\ \psi\left(\dfrac{(x-a)}{h - (i-1)/2 + 1}\right) \\ \qquad \text{for } i \text{ odd,} \\ i = 0, \ldots, 2n+1 \end{cases} \tag{42}$$

are cubic PPs w.r.t. the evenly spaced knots $\{x_i = a + ih : i = 0, \ldots, n\}$, where $h = (b-a)/n$. Note that ϕ_i and ϕ_i' are continuous, but ϕ_i'' does not exist at the knots, and so $\phi_i \in \mathscr{L}^1$. These functions form a basis for the space of cubic PPs in $\mathscr{L}^1$ w.r.t. these knots.

8.6 Piecewise Polynomial Interpolation

Piecewise polynomials, including splines, are often used for interpolation, especially for large data sets. The main advantage in using PPs, instead of polynomials, to interpolate a function $g(x)$ at n points, where $n \gg 1$, is that the error of a PP interpolant does not depend on the nth derivative of $g(x)$, but rather on a low-order derivative of $g(x)$. Usually, if $g(x) \in \mathscr{L}^{N+1}[a, b]$ and $S_N(x)$ is a PP of degree N in $\mathscr{L}^K$ that interpolates $g(x)$ at $n(N-K) + K + 1$ properly chosen points, then the interpolation error at any point $x \in [a, b]$ satisfies

$$|g(x) - s_N(x)| \le Ch^{N+1} \max_{a \le \xi \le b} |g^{(N+1)}(\xi)|$$

for some constant C independent of $h = \max\{x_i - x_{i-1} : i = 1, \ldots, n\}$. Another advantage of PP interpolation is that it leads either to simple relations that define the PP interpolant or gives rise to banded systems that can be solved easily for the coefficients of the PP interpolant by the techniques described

in Sec. 2.7. In the following subsections, we give examples of PP interpolation. We assume throughout that $h = (b-a)/n$ and $x_i = a + ih$ for $i = 0, \ldots, n$, but these restrictions can be removed easily.

An introduction to PP interpolation is presented in many introductory numerical analysis textbooks, such as Johnson and Riess (1982). For a more detailed discussion, see an advanced text such as Prenter (1975) or de Boor (1978).

8.6.1 Linear Spline Interpolation

Let $\phi_i(x)$ be defined by (39). The function

$$S_1(x) = \sum_{i=0}^{n} c_i \phi_i(x)$$

with $c_i = g(x_i)$ is the unique linear spline that interpolates $g(x)$ at the knots $\{x_i : i = 0, \ldots, n\}$.

8.6.2 Quadratic Spline Interpolation

Let $\phi_i(x)$ be defined by (40). The function

$$s_2(x) = \sum_{i=0}^{n+1} c_i \phi_i(x)$$

is the unique quadratic spline that interpolates $g(x)$ at the $n+2$ points x_0, $(x_{i-1} + x_i)/2$, $i = 1, \ldots, n$, and x_n, if the coefficients c_i, $i = 0, \ldots, n+1$, are chosen so that they satisfy the $n+2$ conditions $s_2(x_0) = g(x_0)$, $s_2((x_{i-1} + x_i)/2) = g((x_{i-1} + x_i)/2)$, $i = 1, \ldots, n$, and $s_2(x_n) = g(x_n)$. These conditions give rise to a tridiagonal linear system of $n+2$ equations in the $n+2$ unknowns c_i, $i = 0, \ldots, n+1$:

$$\begin{pmatrix} 0.5 & 0.5 & & & & \\ 0.125 & 0.75 & 0.125 & & & \\ & 0.125 & 0.75 & 0.125 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0.125 & 0.75 & 0.125 \\ & & & & 0.5 & 0.5 \end{pmatrix} \times \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_n \\ c_{n+1} \end{pmatrix} = \begin{pmatrix} g(x_0) \\ g((x_0 + x_1)/2) \\ g((x_1 + x_2)/2) \\ \vdots \\ g((x_{n-1} + x_n)/2) \\ g(x_n) \end{pmatrix}.$$

Since this system is column diagonally dominant, it is nonsingular and has a unique solution. Moreover, if the first and last equations are multiplied by 0.25, the resulting system is symmetric positive-definite. So the coefficients $\{c_i\}$ of the associated quadratic spline interpolant can be computed easily by the techniques described in Sec. 2.7.

8.6.3 Cubic Spline Interpolation

Let $\phi_i(x)$ be defined by (41). The function

$$s_{32}(x) = \sum_{i=-1}^{n+1} c_i \phi_i(x)$$

is a cubic spline that interpolates $g(x)$ at the $n+1$ points x_i, $i = 0, \ldots, n$, if the coefficients c_i, $i = -1, \ldots, n+1$, are chosen so that they satisfy the $n+1$ conditions $s_{32}(x_i) = g(x_i)$, $i = 0, \ldots, n$. These conditions give rise to a linear system of $n+1$ equations in the $n+3$ unknowns c_i, $i = -1, \ldots, n+1$. To determine the unknown coefficients uniquely, two more appropriately chosen conditions (equations) are needed. The standard choices are

1. $s''_{32}(x_0) = 0$ and $s''_{32}(x_n) = 0$, giving rise to the *natural* or *minimum-curvature* cubic spline interpolant;

2. $s'_{32}(x_0) = g'(x_0)$ and $s'_{32}(x_n) = g'(x_n)$, giving rise to a more accurate cubic spline interpolant, but one that requires knowledge of $g'(x)$ (or a good approximation to it) at both end points;
3. the *not-a-knot* conditions, which force $s_{32}(x)$ to have a continuous third derivative at the knots x_1 and x_{n-1}.

In all three cases, the resulting linear system of $n+3$ equations in $n+3$ unknowns is almost tridiagonal. More specifically, it is tridiagonal with the exception of the two rows corresponding to the extra conditions. So the coefficients $\{c_i\}$ of the associated cubic spline interpolant can be computed easily by the techniques described in Sec. 2.7.

8.6.4 Cubic Hermite Piecewise Polynomial Interpolation

Let $\phi_i(x)$ be defined by (42). The function

$$s_{31}(x) = \sum_{i=0}^{2n+1} c_i \phi_i(x)$$

with $c_{2i} = g(x_i)$, $i = 0, \ldots, n$, and $c_{2i+1} = g'(x_i)$, $i = 0, \ldots, n$, is the unique Hermite PP that interpolates $g(x)$ and its derivative at the $n+1$ points x_i, $i = 0, \ldots, n$.

8.7 Least-Squares Approximation

It is possible to construct a function $f(x)$ that approximates the data $\{(x_i, y_i): i = 0, \ldots, m\}$ in the sense that $f(x_i) - y_i$ is "small" for $i = 0, \ldots, m$, but $f(x_i) \neq y_i$ in general. One way is to construct an f in some class of functions that minimizes

$$\sum_{i=0}^{m} w_i [f(x_i) - y_i]^2$$

for some positive *weights* w_i, $i = 0, \ldots, m$. Such an f is called a *discrete least-squares* approximation to the data.

Similarly, it is possible to construct a function $f(x)$ that approximates a continuous function $y(x)$ on an interval $[a, b]$ in the sense that $|f(x) - y(x)|$ is "small" for all $x \in [a, b]$. One way is to construct an f in some class of functions that minimizes

$$\int_a^b w(x)[f(x) - y(x)]^2 \, dx$$

for some positive and continuous *weight function* $w(x)$ on (a, b). Such an f is called a *continuous least-squares* approximation to y.

Often, f is chosen to be a polynomial of degree $n < m$. (For the discrete least-squares problem, if f is a polynomial of degree $n = m$, then f interpolates the data.)

Let f and $g \in \mathscr{L}[a, b]$. We denote the *discrete inner product* of f and g at distinct points x_i, $i = 0, \ldots, m$, w.r.t. the weights w_i, $i = 0, \ldots, m$, by

$$(f, g) = \sum_{i=0}^{m} w_i f(x_i) g(x_i)$$

and the *continuous inner product* of f and g on $[a, b]$ w.r.t. weight function $w(x)$ by

$$(f, g) = \int_a^b w(x) f(x) g(x) \, dx.$$

Given an inner product $(\cdot, \cdot)$, as above, we denote the *norm* of f by $\|f\| = (f, f)^{1/2}$. Thus we have the *discrete norm*

$$\|f\| = \left(\sum_{i=0}^{m} w_i f(x_i)^2 \right)^{1/2}$$

and the *continuous norm*

$$\|f\| = \left(\int_a^b w(x) f(x)^2 \, dx \right)^{1/2}.$$

Therefore, the problem of constructing a least-squares approximation $f(x)$ to a given set of data $\{(x_i, y_i)\}$ or to a given function $y(x)$ is to construct an $f(x)$ that minimizes the norm, discrete or continuous, respectively, of the error $\|f - y\|$.

We note that the "discrete inner product" is not strictly speaking an inner product in all cases, since it may fail to satisfy the property that $(f, f) = 0$ implies $f(x) = 0$ for all x. However, if we restrict the class of functions to which f belongs appropriately, then $(\cdot, \cdot)$, is a true inner product. For example, if we restrict f to the class of polynomials of degree at most n and if $n < m$, then $(f, f) = 0$ implies $f(x) = 0$ for all x. Similar remarks apply to the discrete norm.

Before giving the main theorem on how to construct the least-squares polynomial approximation to a given set of data or to a given function, we introduce orthogonal polynomials and the Gram–Schmidt process to construct them.

8.7.1 Orthogonal Polynomials

A set of $n+1$ polynomials $\{q_i(x): i = 0, \ldots, n\}$ is *orthogonal* w.r.t. the inner product $(\cdot, \cdot)$ if $(q_i, q_j) = 0$ for $i \neq j$. A set of $n+1$ orthogonal polynomials $\{q_i(x): i = 0, \ldots, n\}$ is *orthonormal* w.r.t. the inner product $(\cdot, \cdot)$ if in addition $(q_i, q_i) = 1$ for $i = 0, \ldots, n$.

8.7.2 The Gram–Schmidt Orthogonalization Algorithm

The Gram–Schmidt algorithm applied to a set of $n+1$ linearly independent polynomials $\{p_j : j = 0, \ldots, n\}$ generates a set of $n+1$ orthonormal polynomials $\{q_i(x) : i = 0, \ldots, n\}$ and a set of $n+1$ orthogonal polynomials $\{s_i(x) : i = 0, \ldots, n\}$. Often the set of $n+1$ linearly independent polynomials $\{p_j : j = 0, \ldots, n\}$ is chosen to be the set of monomials $\{x^j : j = 0, \ldots, n\}$.

The Gram–Schmidt algorithm for polynomials is similar to the Gram–Schmidt algorithm for matrices described in Sec. 4.6. The reader may refer to an introductory numerical methods text such as Johnson and Riess (1982) for more details. Here we note only that the role of the inner product of vectors in the algorithm in Sec. 4.6 is replaced by the inner product, discrete or continuous, of functions as defined in Sec. 8.7.1.

8.7.3 Constructing the Least-Squares Polynomial Approximation

The following result is proved in many introductory numerical methods books. See for example Johnson and Riess (1982).

THEOREM: Assume that we are given either a continuous function $y(x)$ on $[a, b]$ or a data set $\{(x_i, y_i) : i = 0, \ldots, m\}$. Let $\{q_j : j = 0, \ldots, n\}$ be a set of orthonormal polynomials w.r.t. an inner product $(\cdot, \cdot)$ appropriate for the given data and assume that $\{q_j : j = 0, \ldots, n\}$ spans $\{x^i : i = 0, \ldots, n\}$, where $n < m$ for the discrete problem. Then

$$p^*(x) = \sum_{j=0}^{n} (y, q_j) q_j(x)$$

is the least-squares polynomial approximate of degree at most n. It is optimal in the sense that, if $p(x)$ is any other polynomial of degree at most n, then $\|y - p^*\| < \|y - p\|$, where $\|\cdot\|$ is the norm associated with the inner product $(\cdot, \cdot)$.

As noted in Sec. 8.7.2, a set of orthonormal polynomials $\{q_j : j = 0, \ldots, n\}$ that spans $\{x^i : i = 0, \ldots, n\}$ can be constructed by the Gram–Schmidt algorithm applied to the monomial basis polynomials $\{x^i : i = 0, \ldots, n\}$.

9
Numerical Integration – Quadrature

In this section, we consider formulas for approximating integrals of the form

$$I(f) = \int_a^b f(x)\,dx.$$

Such formulas are often called quadrature rules. We assume that a and b are finite and that f is smooth in most cases, but we briefly discuss infinite integrals and singularities in Sec. 9.5.

In many practical problems, $f(x)$ is given either as a set of values $f(x_1), \ldots, f(x_n)$ or $f(x)$ is hard or impossible to integrate exactly. In these cases, the integral may be approximated by numerical techniques, which often take the form

$$Q(f) = \sum_{i=1}^{n} w_i f(x_i).$$

Such a formula is called a *quadrature rule*, the $\{w_i\}$ are called *weights*, and the $\{x_i\}$ are called *abscissae* or *nodes*.

Most quadrature rules are derived by first approximating f by a simpler function, frequently a polynomial, and then integrating the simpler function. Thus, the area under the curve f, which is the exact value of $I(f)$, is approximated by the area under the curve of the simpler function.

For a more detailed discussion of the topics in this section, see an introductory numerical methods text such as Conte and de Boor (1980), Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980).

9.1
Simple Quadrature Rules

Several simple quadrature rules follow:

- The *rectangle rule* approximates $I(f)$ by the area under the constant $y = f(a)$ or $y = f(b)$.
- The *midpoint rule* approximates $I(f)$ by the area under the constant $y = f((a+b)/2)$.
- The *trapezoidal rule* approximates $I(f)$ by the area under the straight line joining the points $(a, f(a))$ and $(b, f(b))$.
- *Simpson's rule* approximates $I(f)$ by the area under the quadratic that interpolates $(a, f(a))$, $(m, f(m))$ and $(b, f(b))$, where $m = (a+b)/2$.
- The *corrected trapezoidal rule* approximates $I(f)$ by the area under the cubic Hermite that interpolates $(a, f(a), f'(a))$ and $(b, f(b), f'(b))$.
- Newton–Cotes rules are discussed in Sec. 9.1.1 below.
- Gaussian rules are discussed in Sec. 9.1.2 below.

The formula $Q(f)$ and the associated error $I(f) - Q(f)$ for each quadrature rule listed above are given in Table 13, where n is the number of function and derivative evaluations, d is the polynomial degree of the quadrature rule (see Sec. 9.1.1 below), η is an unknown point in $[a, b]$ (generally different for each rule), $m = (a+b)/2$ is the midpoint of the interval $[a, b]$, and C and K are some constants. For the derivation of these quadrature rules and their associated error formulas, see an introductory numerical methods text such as Conte and de Boor (1980); Dahlquist and Björck (1974); Johnson and Riess (1982); Stoer and Bulirsch (1980).

9.1.1 Some Definitions

A quadrature rule that is based on integrating a polynomial interpolant is called an *interpolatory rule*. All the simple quadrature rules listed in Table 13 are interpolatory. Writing the polynomial interpolant in Lagrange form and integrating it, we see immediately that the weights w_i do not depend on the function f, but only on the abscissae $\{x_i : i = 1, \ldots, n\}$.

Tab. 13 Simple quadrature rules

Quadrature rule	***n***	***d***	***Interpolant***	$Q(f)$	$I(f) - Q(f)$
Rectangle	1	0	Constant	$(b-a)f(a)$	$\frac{(b-a)^2}{2}f'(\eta)$
Midpoint	1	1	Constant	$(b-a)f(m)$	$\frac{(b-a)^3}{24}f''(\eta)$
Trapezoidal	2	1	Linear	$\frac{b-a}{2}[f(a)+f(b)]$	$-\frac{(b-a)^3}{12}f''(\eta)$
Simpson's	3	3	Quadratic	$\frac{b-a}{6}[f(a)+4f(m)+f(b)]$	$-\frac{(b-a)^5}{2880}f^{(4)}(\eta)$
Corrected trapezoidal	4	3	Cubic	$\frac{b-a}{2}[f(a)+f(b)]+\frac{(b-a)^2}{12}[f'(a)-f'(b)]$	$\frac{(b-a)^5}{720}f^{(4)}(\eta)$
Newton–Cotes	n	$\geq n-1$	Deg. $n-1$	See Sec. 9.1.1	$\frac{(b-a)^{d+2}}{K}f^{(d+1)}(\eta)$
Gaussian	n	$2n-1$	Deg.$n-1$	See Sec. 9.1.2	$\frac{(b-a)^{d+2}}{C}f^{(d+1)}(\eta)$

Quadrature rules are, in general, not exact. An error formula for an interpolatory rule can often be derived by integrating the associated polynomial interpolant error. Error formulas for some simple quadrature rules are listed in Table 13.

A quadrature rule that is exact for all polynomials of degree d or less, but is not exact for all polynomials of degree $d+1$, is said to have *polynomial degree d*. An interpolatory rule based on n function and derivative values has polynomial degree at least $n-1$.

Quadrature rules that include the end points of the interval of integration $[a, b]$ as abscissae are called *closed rules*, while those that do not include the end points are called *open rules*. An advantage of open rules is that they can be applied to integrals with singularities at the end points, whereas closed rules usually can not.

Interpolatory rules based on equidistant abscissae are called *Newton–Cotes* rules. This class includes the rectangle, midpoint, trapezoidal, corrected trapezoidal, and Simpson's rules. Both open and closed Newton–Cotes quadrature rules exist.

9.1.2 Gaussian Quadrature Rules

A quadrature rule

$$Q(f) = \sum_{i=1}^{n} w_i f(x_i)$$

is fully determined by n, the abscissae $\{x_i : i = 1, \ldots, n\}$, and the weights $\{w_i : i = 1, \ldots, n\}$. Gauss showed that, given n and the end points a and b of the integral,

1. there exists a unique set of abscissae $\{x_i\}$ and weights $\{w_i\}$ that give a quadrature rule – called the Gaussian quadrature rule – that is exact for all polynomials of degree $2n-1$ or less;
2. no quadrature rule with n abscissae and n weights is exact for all polynomials of degree $2n$;
3. the weights $\{w_i\}$ of the Gaussian quadrature rule are all positive;
4. the Gaussian quadrature rule is open;
5. the abscissae $\{x_i\}$ of the Gaussian quadrature rule are the roots of the shifted Legendre polynomial $q_n(x)$ of degree n, which is the unique monic polynomial of degree n that is orthogonal to all polynomials of degree $n-1$ or less w.r.t. the continuous inner product

$$(f, g) = \int_a^b f(x)g(x)\,dx$$

(see Sec. 8.7.1);
6. the Gaussian quadrature rule is interpolatory, i.e., it can be derived by integrating the polynomial of degree $n-1$ that interpolates f at the abscissae $\{x_i\}$.

Thus, Gauss derived the class of open interpolatory quadrature rules of maximum polynomial degree $d = 2n-1$. As noted above, these formulas are called *Gaussian quadrature rules* or *Gauss–Legendre quadrature rules*.

9.1.3 Translating the Interval of Integration

The weights and abscissae of simple quadrature rules are usually given w.r.t. a specific interval of integration, such as [0,1] or $[-1, 1]$. However, the weights and abscissae can be transformed easily to obtain a related quadrature rule appropriate for some other interval.

One simple way to do this is based on the linear change of variables $\hat{x} = \beta(x-a)/(b-a) + \alpha(b-x)/(b-a)$, which leads to the relation

$$\int_\alpha^\beta f(\hat{x})d\hat{x} = \int_a^b f\left(\beta\frac{x-a}{b-a} + \alpha\frac{b-x}{b-a}\right) \times \frac{\beta-\alpha}{b-a}\,dx. \tag{43}$$

So, if we are given a quadrature rule on $[a, b]$ with weights $\{w_i : i = 1, \ldots, n\}$ and abscissae $\{x_i : i = 1, \ldots, n\}$, but we want to compute

$$\int_{\alpha}^{\beta} f(\hat{x}) d\hat{x},$$

we can apply the quadrature rule to the integral on the right side of (43). An equivalent way of viewing this is that we have developed a related quadrature rule for the interval of integration $[\alpha, \beta]$ with weights and abscissae

$$\hat{w}_i = \frac{\beta - \alpha}{b - a} w_i \quad \text{and}$$

$$\hat{x}_i = \beta \frac{x_i - a}{b - a} + \alpha \frac{b - x_i}{b - a}$$

for $i = 1, \ldots, n$. Note that, because we have used a linear change of variables, the original rule for $[a, b]$ and the related one for $[\alpha, \beta]$ have the same polynomial degree.

9.1.4 Comparison of Gaussian and Newton–Cotes Quadrature Rules

We list some similarities and differences between Gaussian and Newton–Cotes quadrature rules:

1. The weights of a Gaussian rule are all positive, which contributes to the stability of the formula. High-order Newton–Cotes rules typically have both positive and negative weights, which is less desirable, since it leads to poorer stability properties.
2. Gaussian rules are open, whereas there are both open Newton–Cotes rules and closed Newton–Cotes rules.
3. Gaussian rules attain the maximum possible polynomial degree $2n - 1$ for a formula with n weights and abscissae, whereas the polynomial degree d of a Newton–Cotes rules satisfies $n - 1 \leq d \leq 2n - 1$ and the upper bound can be obtained for $n = 1$ only.
4. The abscissae and weights of Gaussian rules are often irrational numbers and hard to remember, but they are not difficult to compute. The abscissae of a Newton–Cotes rule are easy to remember and the weights are simple to compute.
5. The set of abscissae for an n-point Gaussian rule and for an m-point Gaussian rule are almost disjoint for all $n \neq m$. Thus we can not reuse function evaluations performed for one Gaussian rule in another Gaussian rule. Appropriately chosen pairs of Newton–Cotes rules can share function evaluations effectively.
6. Both Gaussian and Newton–Cotes rules are interpolatory.

9.2 Composite (Compound) Quadrature Rules

To increase the accuracy of a numerical approximation to an integral, we could use a rule with more weights and abscissae. This often works with Gaussian rules, if f is sufficiently smooth, but it is not advisable with Newton–Cotes rules, for example, because of stability problems associated with high-order formulas in this class.

Another effective way to achieve high accuracy is to use *composite* quadrature rules, often also called *compound* quadrature rules. In these schemes, the interval of integration $[a, b]$ is subdivided into *panels* (or subintervals) and on each panel the same simple quadrature rule is applied. If the associated simple quadrature rule is interpolatory, then this approach leads to the integration of a piecewise polynomial

Tab. 14 Composite quadrature rules

Quadrature Rule	*n*	*d*	*PP interpolant*	*Formula*	*Error*
Rectangle	s	0	Constant	$h\sum_{i=0}^{s-1} f(a+ih)$	$\frac{h}{2}(b-a)f'(\eta)$
Midpoint	s	1	Constant	$h\sum_{i=1}^{s} f[a+(i-1/2)h]$	$\frac{h^2}{24}(b-a)f''(\eta)$
Trapezoidal	$s+1$	1	Linear	$\frac{h}{2}[f(a)+f(b)+2\sum_{i=1}^{s-1} f(a+ih)$	$-\frac{h^2}{12}(b-a)f''(\eta)$
Simpson's	$2s+1$	3	Quadratic	$\frac{h}{6}\left[f(a)+f(b)+2\sum_{i=1}^{s-1} f(a+ih)+4\sum_{i=1}^{s-1} f(a+(i-1/2)h)\right]$	$-\frac{h^4}{2880}(b-a)f^{(4)}(\eta)$
Corrected trapezoidal	$s+3$	3	Cubic Hermite	$\frac{h}{2}\left[f(a)+f(b)+2\sum_{i=1}^{s-1} f(a+ih)\right]+\frac{h^2}{12}[f'(a)-f'(b)]$	$\frac{h^4}{720}(b-a)f^{(4)}(\eta)$

(PP) interpolant. Thus, using a composite quadrature rule, instead of a simple one with high polynomial degree, leads to many of the same benefits that are obtained in using PP interpolants compared with high-degree polynomial interpolants (see Sec. 8.5).

Table 14 summarizes the composite quadrature rules and the associated error formulas. In the table, n is the number of function and derivative evaluations, d is the polynomial degree of the quadrature rule, η is an unknown point in $[a, b]$ (in general different for each rule), $h = (b - a)/s$ is the step size of each panel, s is the number of panels, and PP stands for piecewise polynomial. Composite quadrature rules based on Gaussian or Newton–Cotes rules can also be used, although they are not listed in Table 14.

9.3 Adaptive Quadrature

We see from Table 14 of composite quadrature rules that the smaller the step size h of a panel, the smaller the expected error. It is often the case that an approximation to the integral

$$I(f) = \int_a^b f(x)\,dx$$

is needed to within a specified accuracy $\in$. In this case, *adaptive quadrature* is often used. Such schemes refine the grid (or collection of panels) until an estimate of the total error in the integration is within the desired precision $\in$. Adaptive quadrature is particularly useful when the behavior of the function f varies significantly in the interval of integration $[a, b]$, since the scheme can use a fine grid where f is hard to integrate and a coarse grid where it is easy, leading to an efficient and accurate quadrature procedure.

Tab. 15 Adaptive quadrature procedure

```
subroutine AQ(a, b, ∈)
    (Q, E) = LQM(a, b)
    if (E ≤∈) then
      return (Q, E)
    else
      m = (a + b)/2
      return AQ(a, m, ∈ /2) + AQ(m, b, ∈ /2)
    end
end
```

Table 15 gives a simple general recursive procedure for adaptive quadrature. We assume that we can make use of a routine LQM (Local Quadrature Module) that implements a quadrature rule in some interval $[a, b]$ and returns Q, an approximation to the integral, and E, an estimate of the error. In the next section, we discuss how an error estimate may be obtained.

We note that the adaptive quadrature procedure shown in Table 15 does not illustrate how to reuse function evaluations where possible. An effective adaptive quadrature routine should do this, since function evaluations are often the most computationally expensive part of the procedure.

9.4 Romberg Integration and Error Estimation

As discussed in the last subsection, adaptive quadrature requires an error estimate. As an illustration, we consider how one may be obtained for the composite trapezoidal rule.

Let $T_s(f)$ denote the composite-trapezoidal-rule approximation to

$$I(f) = \int_a^b f(x)\,dx$$

using s panels and let $E_s = I(f) - T_s(f)$ be the associated error. On the basis of

the error formula in Table 13, we expect E_{2s} to be about four times smaller than E_s, assuming that $f''(x)$ does not vary too much. That is,

$$I(f) - T_s(f) = E_s,$$
$$I(f) - T_{2s}(f) = E_{2s} \approx \tfrac{1}{4} E_s.$$

Subtracting these two equations, we get

$$T_{2s}(f) - T_s(f) = E_s - E_{2s} \approx \tfrac{3}{4} E_s.$$

So $E_s \approx 4[T_{2s}(f) - T_s(f)]/3$ and $E_{2s} \approx [T_{2s}(f) - T_s(f)]/3$. Thus, by applying the composite trapezoidal rule first with s and then with $2s$ panels, we obtain estimates of the error in both $T_s(f)$ and $T_{2s}(f)$.

If we add the estimate of the error $E_s = I(f) - T_s(f)$ to $T_s(f)$ we often obtain a better approximation to $I(f)$ than either $T_s(f)$ or $T_{2s}(f)$. That is, $\hat{T}_s(f) = T_s(f) + 4[T_{2s}(f) - T_s(f)]/3 = [4T_{2s}(f) - T_s(f)]/3$ is often a better approximation to $I(f)$ than either $T_s(f)$ or $T_{2s}(f)$, particularly if the function f is smooth and the grid is fine. Thus, by applying the compound trapezoidal rule with s and $2s$ panels and taking an appropriate linear combination of the two approximations, we construct a better approximation to $I(f)$. To be more specific, it can be shown that this eliminates the lowest-order term in the error E_s or E_{2s}. This process can be repeated to eliminate the next higher-order term in the error and so on. In addition, it can be generalized easily to other quadrature rules.

This is the basic idea behind *Romberg integration*. By applying a quadrature rule repeatedly with more panels each time, we can eliminate the leading terms of the error expansion, and thereby obtain better and better approximations to $I(f)$.

9.5 Infinite Integrals and Singularities

If one or both end points of the interval of integration $[a, b]$ are infinite, the integral is called infinite. We restrict the discussion of infinite integrals to the case

$$I(f) = \int_a^\infty f(x)\,dx,$$

sometimes called a semi-infinite integral, since only one end point is infinite. Other cases can be handled similarly.

Under the assumption that

$$I(f) = \int_a^\infty f(x)\,dx$$

exists, one way to approximate the integral is to *truncate* $I(f)$, and compute instead

$$\hat{I}(f) = \int_a^b f(x)\,dx,$$

for some sufficiently large b, by a standard quadrature rule $Q(f)$. The error in this approach is $I(f) - Q(f) = [I(f) - \hat{I}(f)] + [\hat{I}(f) - Q(f)]$. It is often possible to choose b so that

$$I(f) - \hat{I}(f) = \int_b^\infty f(x)\,dx$$

is small and to choose Q so that $\hat{I}(f) - Q(f)$ is also small.

$I(f)$ can also be approximated by first performing a change of variables to *transform* the infinite integral to a standard one. More specifically, let $x = g(t)$. Then

$$\begin{aligned} I(f) &= \int_a^\infty f(x)dx \\ &= \int_{g^{-1}(a)}^{g^{-1}(\infty)} f(g(t)) \cdot g'(t)\,dt, \end{aligned}$$

where $g^{-1}(x)$ is the inverse of $g(x)$. If we can choose g so that $g^{-1}(a)$ and $g^{-1}(\infty)$ are both finite, then $I(f)$ is transformed to a finite integral that can be evaluated

by a standard quadrature rule. However, this procedure may introduce singularities (discussed below). If so, it might not lead to a computationally easier problem to solve.

Infinite integrals can also be approximated by special quadrature formulas that are directly applicable to infinite intervals of integration. For further details concerning this approach, see an introductory numerical methods text such as Conte and de Boor (1980), Dahlquist and Björck (1974), Johnson and Riess (1982), or Stoer and Bulirsch (1980).

A singular integral

$$I(f) = \int_a^b f(x)\,dx$$

is one in which f is singular (i.e., becomes infinite) at some point in $[a, b]$. Singular and infinite integrals are closely related: A change of variables often transforms one into the other.

In the computing of singular integrals by a quadrature rule, the value of f might be required at or close to a point of singularity, and so the quadrature rule may be either inapplicable or inaccurate. It often happens that the singularity in f occurs at the end point a or b, in which case, an open formula, such as a Gaussian rule, may be effective.

The performance of a quadrature rule applied to a singular integral might be improved by a change of variables. A transformation $x = g(t)$ that often helps to remove or lessen the effect of a singularity is $g(t) = b - (b - a)u^2(2u + 3)$ for $u = (t - b)/(b - a)$.

9.6 Monte-Carlo Methods

Monte-Carlo methods (*q.v.*) are of a statistical nature. For the sake of simplicity, we briefly present them for one-dimensional integrals only, but they can be extended easily to multidimensional integrals and are most useful in this context.

To begin, choose n random points $\{U_i : i = 1, \ldots, n\} \subset [0, 1]$ and scale each U_i to $[a, b]$ by $u_i = a + U_i(b - a)$. Then

$$Q_n(f) = \frac{(b-a)}{n} \sum_{i=1}^{n} f(u_i)$$

is a Monte-Carlo approximation to

$$I(f) = \int_a^b f(x)\,dx.$$

If we consider $Q_n(f)$ to be a random variable, then it can be shown that its mean is $I(f)$ and its standard deviation is $|b - a| \times \rho(f)/\sqrt{n}$, where $\rho(f)$ is a constant that depends on f, but not n. Assuming that $Q_n(f)$ is close to being normally distributed, we are led to statistical statements about the error, such as that $|Q_n(f) - I(f)| \leq 2|b - a|\rho(f)/\sqrt{n}$ nineteen times out of twenty.

Note that the error bound above decreases like $1/\sqrt{n}$, much more slowly than the bounds for the standard compound quadrature rules given in Table 14. This suggests that Monte-Carlo methods are not very effective for one-dimensional integrals of smooth functions. However, an error formula similar to that given above continues to hold for multidimensional integrals, while extensions of standard methods become increasingly less efficient as the dimension of the integral to be approximated increases. As a result, Monte-Carlo methods are among the best schemes available for approximating high-dimensional integrals.

10 Ordinary Differential Equations

In this section, we consider numerical methods for the solution of ordinary differential equations (ODEs). We begin by introducing some simple schemes for the initial-value problem (IVP)

$$y'(x) = f(x, y(x)) \quad x \in [a, b],$$
$$y(a) = y_0, \tag{44}$$

where $y : \mathbb{R} \to \mathbb{R}^m$ and $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$. We assume throughout that the IVP (44) has a unique solution for $x \in [a, b]$. We also discuss more sophisticated adaptive methods and explain the terms *stiff* and *nonstiff* problems and how to choose methods appropriate for these two classes of problems. We then briefly consider the boundary-value problem (BVP)

$$y'(x) = f(x, y(x)) \quad x \in [a, b],$$
$$g(y(a), y(b)) = 0, \tag{45}$$

which we assume has a locally unique solution.

For both IVPs and BVPs, we consider systems of first-order ODEs only, since most commonly available codes are for first-order systems and any higher-order ODE can be reduced to a system of first-order equations. However, using a method designed for higher-order equations directly may lead to a more efficient solution of the problem.

Because of space constraints, we do not discuss many important related problems such as differential-algebraic equations or delay differential equations. For a more comprehensive discussion of these and other related topics, see an advanced text such as Ascher et al. (1988), Butcher (1987), Hairer et al. (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994).

10.1 Initial-Value Problems (IVPs)

10.1.1 Two Simple Formulas

Most standard numerical methods for the IVP (44) start with the initial value y_0 at $x_0 = a$ and then compute approximations $y_n \approx y(x_n)$ for $n = 1, \ldots, N$ on a discrete grid $a = x_0 < x_1 < \cdots < x_N = b$. The distance between adjacent grid points, $h_n = x_{n+1} - x_n$, $n = 0, \ldots, N-1$, is referred to as the *step size* at step $n+1$. Schemes are often presented with a constant step size $h = h_n$ for all n, but this is generally not required. Moreover, as discussed below, variable–step-size methods are often much more efficient.

Possibly the simplest numerical scheme for (44) is *Euler's method*, sometimes called the *forward Euler method*:

$$y_{n+1} = y_n + h_n f(x_n, y_n). \tag{46}$$

This formula is motivated from the observation that the true solution of a scalar IVP of the form (44) satisfies

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + h_n y'(x_n) + \frac{h_n^2}{2} y''(\eta_n) \\ &= y(x_n) + h_n f(x_n, y(x_n)) \\ &\quad + \frac{h_n^2}{2} y''(\eta_n) \end{aligned} \tag{47}$$

for some point $\eta_n \in [x_n, x_{n+1}]$, which follows from standard Taylor-series theory. Thus, the true solution of the IVP (44) satisfies an equation that is very similar to (46). This argument can be extended easily to systems.

The approximations $y_n \approx y(x_n)$ are computed in the order $n = 1, \ldots, N$ using the formula (46). To be more specific, on

the first step of Euler's method from x_0 to $x_1 = x_0 + h_0$, we substitute the initial value (x_0, y_0) into the right side of (46) to compute $y_1 \approx y(x_1)$. Thus, at the end of the first step, we have (x_1, y_1). On the second step of Euler's method from x_1 to $x_2 = x_1 + h_1$, we substitute (x_1, y_1) into the right side of (46) to compute $y_2 \approx y(x_2)$. The procedure continues in a similar way for $n = 2, \ldots, N$. Note that this evaluation process applies equally well to systems of equations (i.e., $m > 1$). In this case, h_n is a scalar, but y_{n+1}, y_n and $f(x_n, y_n)$ are all m-vectors.

Euler's method is an *explicit formula* in the sense that the evaluation process described above does not require the solution of any linear or nonlinear algebraic equations. The backward Euler formula

$$y_{n+1} = y_n + h_n f(x_{n+1}, y_{n+1}) \tag{48}$$

is a typical example of an *implicit formula*. It can be motivated from a Taylor-series expansion of the true solution $y(x)$ of the IVP (44) about x_{n+1} similar to the expansion (47) above for $y(x)$ about x_n. Moreover, we again compute the approximations $y_n \approx y(x_n)$ in the order $n = 1, \ldots, N$. However, note that, on step $n + 1$ from x_n to $x_{n+1} = x_n + h_n$, we start with (x_n, y_n) and must solve Eq. (48) for y_{n+1}.

We will return to the question of how to solve for y_{n+1} shortly, but first we explain briefly in the next subsection why we may wish to use an implicit scheme (such as the backward Euler formula) rather than an explicit one (such as the forward Euler formula) even though the former clearly requires more work per step than the latter.

10.1.2 Stiff IVPs

Roughly speaking, a stiff IVP is one in which some terms in the solution decay rapidly with respect to the length of the integration, while others vary slowly on this time scale. To illustrate this concept, consider the linear constant-coefficient problem $y' = Ay$, $y(0) = y_0$, for $x \in [0, 1]$, where

$$A = \tfrac{1}{2}\begin{pmatrix} -10^6 - 1 & 10^6 - 1 \\ 10^6 - 1 & -10^6 - 1 \end{pmatrix},$$

$$y_0 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

If we let $z = Py$, where

$$P = \tfrac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

$$D = PAP^{-1} = \begin{pmatrix} -10^6 & 0 \\ 0 & -1 \end{pmatrix},$$

then $z' = Py' = PAy = PAP^{-1}z = Dz$ and $z(0) = Py(0) = (1, 1)^{\mathrm{T}}$. Therefore, $z(x) = (e^{-10^6 x}, e^{-x})^{\mathrm{T}}$ and so $y(x) = P^{-1}z(x) = (e^{-x} + e^{-10^6 x}, e^{-x} - e^{-10^6 x})^{\mathrm{T}}$. The term $e^{-10^6 x}$ that occurs in both $z(x)$ and $y(x)$ gives rise to an initial transient that decays rapidly with respect to the interval of integration [0,1], while the term e^{-x} is associated with a slowly varying smooth term on this scale. The fast and slow terms occur in different components of $z(x)$, but they are mixed in $y(x)$, which is often the case in practice. After the $e^{-10^6 x}$ term dies out, both components of $y(x)$ vary smoothly like e^{-x}.

If we apply the forward Euler formula to $y' = Ay$, we get $y_{n+1} = y_n + h_n A y_n = (I + h_n A) y_n$. If we multiply the last equation through by P and perform the change of variables $z_n = Py_n$, we get $z_{n+1} = (I + h_n D) z_n$. The two components of z_n satisfy $z_{n+1}^{(1)} = (1 - 10^6 h_n) z_n^{(1)}$ and $z_{n+1}^{(2)} = (1 - h_n) z_n^{(2)}$. If we let $h_n > 2 \times 10^{-6}$ at any point during the integration, then $(1 - 10^6 h_n) <$

-1 and $z_n^{(1)}$ will grow in magnitude and oscillate in sign as n increases. Since $y_n = P^{-1}z_n$, this will cause both components of y_n to oscillate about e^{-x_n} with growing amplitude as n increases. Since this instability is undesirable, we must restrict $h_n < 2 \times 10^{-6}$ throughout the integration, even though, after the initial transient, this is likely a much smaller step size than would be required to integrate the slowly varying e^{-x} component of the solution accurately.

On the other hand, if we apply the backward Euler formula to $y' = Ay$, we get $y_{n+1} = y_n + h_n A y_{n+1}$ and so $y_{n+1} = (I - hA)^{-1}y_n$. If we multiply this equation through by P and perform the change of variables $z_n = Py_n$, we get $z_{n+1} = (I - h_n D)^{-1}z_n$. Therefore, $z_{n+1}^{(1)} = z_n^{(1)}/(1 + 10^6 h_n)$ and $z_{n+1}^{(2)} = z_n^{(2)}/(1 + h_n)$. Hence, no matter how large $h_n > 0$ is, $z_n^{(1)}$ decays as n increases. Consequently, after the initial transient, we can choose $h_n > 0$ to integrate $z_n^{(2)}$ accurately without fear of $z_n^{(1)}$ becoming unstable. Since $y_n = P^{-1}z_n$, the same conclusion applies to y_n.

The example above can be generalized to larger systems of equations $y' = Ay$. If A is diagonalizable, then the performance of the method on $y' = Ay$ can be deduced from its performance on the scalar test problems $y' = \lambda y$, where the λ's range over the eigenvalues of A. If the real part of each λ is negative, then $y(x) \to 0$ as $x \to \infty$. We would like the numerical solution to have the same behavior without having to restrict h_n outside the transient region. Methods with this property are called *A-stable*. Generalizing the example above, it is easy to see that the backward Euler formula is A-stable, while the forward Euler formula is not.

The importance of the example above and the scalar test problem $y' = \lambda y$ in particular is that the performance of methods on these simple problems is indicative of their behavior on more general nonlinear stiff problems. A nonrigorous, but intuitive, justification of this follows from the local linearization of $y' = f(x, y)$ at (x_n, y_n):

$$y'(x) \approx f(x_n, y_n) + f_x(x_n, y_n)(x - x_n) + f_y(x_n, y_n)(y - y_n),$$

where $f_y(x, y) = \partial f(x, y)/\partial y \in \mathbb{R}^{m \times m}$ is the Jacobian of f. This problem is usually stiff if

1. some eigenvalue of $f_y(x_n, y_n)$ has a large negative real part with respect to the interval of integration,
2. no eigenvalue of $f_y(x_n, y_n)$ has a large positive real part with respect to the interval of integration, and
3. no eigenvalue of $f_y(x_n, y_n)$ has a large imaginary part unless it also has a relatively large negative real part.

An IVP that is not stiff is called *nonstiff*.

Stiff IVPs arise in many applications, such as chemical kinetics and electrical circuits. As noted earlier, they are characterized by components that vary on vastly different time scales: Some terms in the solution decay rapidly to steady state while others vary slowly.

The observation above that the explicit forward Euler formula is not appropriate for a stiff problem, while the implicit backward Euler formula is, can be generalized. All commonly used formulas that are suitable for stiff problems are implicit in some sense.

See the survey article of Shampine and Gear (1979) or an advanced text such as Butcher (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994) for a more detailed discussion of stiffness.

10.1.3 Solving Implicit Equations

We return now to methods for solving for y_{n+1} in an implicit formula such as (48). One common approach is the predictor–corrector technique, which is just a fixed-point iteration as described in Sec. 6.1. For this scheme (and most others), we need in initial approximation $y_{n+1}^{(0)}$ to y_{n+1}. This could be computed, for example, from the forward Euler formula or some other explicit scheme, or simply by taking $y_{n+1}^{(0)} = y_n$. In the terminology of predictor–corrector techniques, the formula for computing $y_{n+1}^{(0)}$ is referred to as the *predictor* formula. For a predictor–corrector method based on the backward Euler formula (48), the corrector formula would be

$$y_{n+1}^{(l+1)} = y_n + h_n f(x_{n+1}, y_{n+1}^{(l)}), \quad l = 0, 1, \ldots. \tag{49}$$

We substitute $y_{n+1}^{(0)}$ into the right side of (49) to compute $y_{n+1}^{(1)}$, which we in turn substitute into the right side of (49) to compute $y_{n+1}^{(2)}$, and so on. It is easy to show that $y_{n+1}^{(l)} \to y_{n+1}$ as $l \to \infty$ if f satisfies the Lipschitz condition

$$\|f(x_{n+1}, y) - f(x_{n+1}, z)\| \le L\|y - z\|$$

for some constant L and all y, z in a convex domain containing y_{n+1} and $y_{n+1}^{(l)}$ for $l = 0, 1, \ldots$ and $h_n L < 1$. In most codes, one or two corrections only are needed, since the initial guess $y_{n+1}^{(0)}$ is normally a good approximation to y_{n+1}. Consequently, this scheme is not much more expensive to implement than the explicit forward Euler formula. However, a predictor–corrector implementation of an A-stable method (such as the backward Euler formula) is not A-stable. On the contrary, because of the requirement $h_n L < 1$, it will suffer a step size restriction on a stiff problem similar to that of an explicit formula.

Alternatively, we could rewrite the backward Euler formula (48) as

$$F(y) = y - y_n - h_n f(x_{n+1}, y) = 0, \tag{50}$$

where we have replaced the unknown y_{n+1} by y, and then apply one of the other techniques described in Sec. 6 for finding roots of equations to compute the solution $y = y_{n+1}$ of (50). The most commonly used root-finding technique in this context is Newton's method – or a variant of it. As noted in Sec. 6.7, for systems of equations, Newton's method takes the form

$$[I - h_n f_y(x_{n+1}, y_{n+1}^{(l)})]\boldsymbol{\Delta}_l = y_n + h_n f(x_{n+1}, y_{n+1}^{(l)}) - y_{n+1}^{(l)}, \tag{51}$$

$$y_{n+1}^{(l+1)} = y_{n+1}^{(l)} + \boldsymbol{\Delta}_l, \tag{52}$$

where $f_y(x, y) = \partial f(x, y)/\partial y \in \mathbb{R}^{m\times m}$ is the Jacobian of f. Note that we must solve a linear system of m equations in m unknowns to compute the Newton update vector $\boldsymbol{\Delta}_l$ in (51). Typically, Gaussian elimination with partial pivoting (see Sec. 2.5) is used to solve such linear systems. A band or sparse solver (see Sec. 2.7) may dramatically decrease the cost of solving (51) if $I - h_n f_y(x_{n+1}, y_{n+1}^{(l)})$ is large and sparse. Similarly, iterative methods, such as the preconditioned conjugate-gradient method (see Sec. 3.2), may significantly reduce the cost of solving some large sparse problems. See Sec. 13 for a discussion of sources of high-quality numerical software, including stiff-ODE solvers that incorporate sparse and iterative linear equation solvers.

The computational work required to solve (51) can be decreased significantly by using a chord Newton method, often called a simplified Newton method,

that holds the Newton iteration matrix $I - h_n f_y(x_{n+1}, y_{n+1}^{(l)})$ constant over several iterations and possibly several steps of the integration, thus avoiding the necessity to factor the Newton iteration matrix on each iteration (see Sec. 6.8). However, even with this savings, the cost per step of the Newton iteration may be much larger than a predictor–corrector method. However, it has the advantage for formulas appropriate for stiff problems that it avoids the step size restriction associated with the predictor–corrector technique or explicit formulas. Thus, even though the Newton iteration might make the scheme much more expensive per step, the step sizes that can be used might be so much larger that the total cost of the integration is significantly less. Finally note that, as for predictor–corrector methods, an initial guess for $y_{n+1}^{(0)}$ is required. It can be computed by the techniques described above.

10.1.4 Higher-order Formulas

A numerical method for ODEs is said to be *of order p* or *p*th *order* or *p*th-*order convergent* if $y_n = y(x_n) + O(h^p)$ for some integer $p \geq 1$, where $O(h^p)$ is any quantity (in this case, the global error) that can be bounded by h^p times a constant that is independent of h, but that may depend on the IVP and the numerical method. Most standard texts on the numerical solution of ODEs show that both the forward and backward Euler methods are first-order convergent.

Higher-order methods are frequently used in practice because they offer the potential of significantly reducing the computational work required to generate an accurate solution to the IVP (44). To get an intuitive feeling for this, suppose that the length of integration $b - a = 1$, that we use a constant step size h throughout the numerical integration, that the global error for a pth-order method satisfies $y_n - y(x_n) = h^p$, and that we require this error to be of size 10^{-10}. Under these assumptions, the optimal step size for the method is $h = 10^{-10/p}$, and the resulting number of steps needed to integrate from a to b is $N = 10^{+10/p}$. To be more specific, for $p = 1, 2, 5, 10$, the number of steps required is $N = 10^{10}, 10^5, 10^2, 10^1$, respectively. Thus, even though a higher-order method may require more work per step than a lower-order scheme, the dramatic reduction in the number of steps required frequently makes a higher-order method much more efficient than a lower-order one – particularly for problems with stringent error tolerances.

Two common second-order formulas are the *trapezoidal rule*

$$y_{n+1} = y_n + \tfrac{1}{2}h_n[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \tag{53}$$

and the *implicit midpoint rule*

$$y_{n+1} = y_n + f\left(x_n + \frac{h_n}{2}, (y_n, +y_{n+1})/2\right), \tag{54}$$

each of which is implicit, since one clearly needs to solve for y_{n+1}. Note that neither formula requires much more work per step than the backward Euler formula. Moreover, both formulas are A-stable and effective for solving stiff problems at relaxed error tolerances.

10.1.5 Runge–Kutta Formulas

Runge–Kutta (RK) formulas are a general class of methods containing many higher-order schemes. The general form of an s-stage RK formula is

$$k_i = f\left(x_n + c_i h_n, y_n + h_n \sum_{j=1}^{s} a_{ij} k_j\right),$$

$$i = 1, \ldots, s,$$

$$y_{n+1} = y_n + h_n \sum_{i=1}^{s} b_i k_i. \tag{55}$$

That is, we must first compute the s function values k_i and then form a weighted average of the k_i's to compute y_{n+1} from y_n.

RK formulas are one-step schemes in the sense that all the information required to compute y_{n+1} from y_n is generated on the current step from x_n to x_{n+1}. That is, unlike multistep formulas discussed in the next subsection, a RK formula does not require any information from past steps.

If the stages of the RK formula can be ordered so that $a_{ij} = 0$ for all $j \geq i$, then the formula (55) is explicit in the sense that the k_i's can be computed in the order $i = 1, \ldots, s$ without having to solve any linear or nonlinear equations. In what follows, we assume that the formula has been so ordered if possible. If the RK formula (55) is not explicit, then it is implicit and at least one linear or nonlinear equation must be solved to compute the k_i's.

The coefficients of a RK formula are frequently displayed in a tableau as

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{22} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss} \\ \hline & b_1 & b_2 & \ldots & b_s \end{array}$$

If all the elements in the tableau on the diagonal and above it are zero, then the associated formula is explicit.

All the methods considered so far are in fact RK formulas. The RK tableaux for the forward Euler formula, backward Euler formula, implicit midpoint rule, and trapezoidal rule are listed below in that order:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}, \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \quad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array},$$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Note that the first three are one-stage RK formulas and the final one, the trapezoidal rule, is a two-stage RK formula. The tableau for the classical four-stage fourth-order explicit RK formula is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

This formula has been widely used since it was published by Kutta in 1901. In the days of hand calculation, the zero coefficients below the diagonal were a distinct benefit, but this is no longer a significant advantage on a modern computer. Moreover, there are now many better formulas, both of order four and of higher order. The interested reader should consult an advanced text such as Butcher (1987), Hairer et al. (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994) for further details.

As the sample formulas above suggest, a high-order RK formula requires more stages than a low-order one. The minimum number of stages that an explicit RK formula requires to attain orders 1 to 8 are listed below.

Order	1	2	3	4	5	6	7	8
Stages	1	2	3	4	6	7	9	11

On the other hand, implicit s-stage RK formulas of order $2s$ exist for all $s \geq 1$. Moreover, it can be shown that this is the maximal order possible.

Explicit RK formulas are frequently used to solve nonstiff IVPs. Some implicit RK formulas are A-stable (or nearly so) and are suitable for solving stiff IVPs. Formulas with four or fewer stages are quite effective for problems with relaxed error tolerances, while formulas with five or more stages are suitable for problems with more stringent accuracy requirements. Since RK formulas are one-step schemes, unlike linear multistep formulas (LMFs) discussed in the next subsection, RK formulas are more suitable than LMFs for problems that require rapid changes in step size, such as problems with discontinuities. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on RK formulas.

10.1.6 Linear Multistep Formulas

Linear multistep formulas (LMFs) can be written in the form

$$y_{n+1} = \sum_{i=1}^{k} \alpha_i y_{n+1-i} + h \sum_{i=0}^{k} \beta_i f(x_{n+1-i}, y_{n+1-i}), \tag{56}$$

where we assume that at least one of α_k or β_k is nonzero [otherwise we can reduce k in (56)]. Formula (56) is in fact a k-step method, since it uses values over k steps to compute y_{n+1}. Therefore, we assume that, at the start of step $n+1$, we have $y_{n+1-k}, \ldots, y_n$ and our task is to compute y_{n+1} by (56). In this case, if $\beta_0 = 0$, then the $f(x_{n+1}, y_{n+1})$ term can be dropped from the right side of (56), and so we can evaluate the right side of (56) to compute y_{n+1} without having to solve any linear or nonlinear equations. That is, if $\beta_0 = 0$, the formula is explicit. On the other hand, if $\beta_0 \neq 0$, then y_{n+1} occurs on both sides of (56), and so a linear or nonlinear equation must be solved to compute y_{n+1}. Therefore, the formula is implicit. Depending on the context, a predictor–corrector method or some variant of Newton's method is typically used to solve for y_{n+1}.

Of course, at the start of the integration, $n = 0$ and $y_{1-k}, \ldots, y_{-1}$ are typically not available for $k > 1$. One solution to this problem is to compute $y_1, \ldots, y_{k-1}$ by a one-step method (such as a RK formula) of the same order as the k-step LMF and then start using the k-step LMF at step k. Alternatively, we could use a one-step LMF on step 1, a two-step LMF on step 2, and so on, until we reach step k, after which we can use (56) on that step and all subsequent steps. Most LMF codes employ the latter strategy and adjust the step size so that the accuracy obtained by the formulas with smaller k is comparable to that obtained by formulas with larger k.

If we ignore stability, then it is possible to obtain a k-step LMF of order $2k$ for any $k \geq 1$. Moreover, it is easy to show that this is the maximal order possible. However, these maximal-order formulas are unstable for $k \geq 3$ and so are not useful in practice. It can be shown that, for any $k \geq 1$, the maximal order of a stable k-step LMF is $k+1$ if k is odd and $k+2$ if k is even.

We have presented the LMFs in this section for fixed step size only, since variable–step-size formulas are considerably more complicated. However, a variable–step-size variable-order scheme may be far more efficient in practice. Such schemes are discussed in advanced texts on the numerical solution of ODEs, such as Hairer et al. (1987), Hairer and Wanner (1991), Lambert (1991), and Shampine

(1994), but not usually in introductory numerical methods books.

10.1.7 Adams Formulas

Adams formulas are a subclass of LMFs that have the form

$$y_{n+1} = y_n + h\sum_{i=0}^{k} \beta_i f(x_{n+1-i}, y_{n+1-i}). \tag{57}$$

In the explicit Adams–Bashforth formulas, $\beta_0 = 0$ and the remaining k β_i's are chosen to obtain the maximal possible order k. In the implicit Adams–Moulton formulas, $\beta_0 \neq 0$ and the $k+1\beta_i$'s are chosen to obtain the maximal possible order $k+1$. These coefficients are listed in most advanced texts on numerical methods for ODEs and in many introductory numerical methods books. It turns out that the forward Euler formula is the one-step Adams–Bashforth formula, and the trapezoidal rule is the one-step Adams–Moulton formula. Moreover, note that the order of the Adams–Moulton formulas is optimal for k odd and nearly optimal for k even.

The implicit Adams–Moulton formulas have somewhat better numerical characteristics than the explicit Adams–Bashforth formulas. Consequently, Adams formulas are usually implemented in a predictor–corrector fashion, with the k- or $(k+1)$-step Adams–Bashforth formula used for the predictor and a k-step Adams–Moulton formula used for the corrector. Adams predictor–corrector schemes are the basis for several very effective variable–step-size variable-order codes for nonstiff IVPs. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on Adams formulas.

10.1.8 Backward Differentiation Formulas

Backward differentiation formulas (BDFs), sometimes called *Gear formulas*, are another subclass of LMFs that have the form

$$y_{n+1} = \sum_{i=1}^{k} \alpha_i y_{n+1-i} + h\beta_0 f(x_{n+1}, y_{n+1}), \tag{58}$$

where $\beta_0 \neq 0$ and so the BDFs are implicit. The $k+1$ coefficients of a k-step BDF are chosen to obtain the maximal possible order k. However, the BDFs are stable for $1 \leq k \leq 6$ only. They are A-stable for $k = 1$ and 2 and nearly A-stable for $k = 3$, 4, and 5, with the loss of A-stability increasing with k. For $k = 6$ the loss of A-stability increases to such an extent that this formula is frequently excluded from use.

The coefficients for the BDFs are listed in most advanced texts on numerical methods for ODEs and in many introductory numerical methods books. It turns out that the backward Euler formula is the one-step BDF.

Because the BDFs are usually used to solve stiff problems, the implicit equation is normally solved by Newton's method or some variant of this root-finding scheme.

BDFs are the basis for several very effective variable–step-size variable-order codes for stiff IVPs. See Sec. 13 for a discussion of sources of high-quality numerical software, including routines based on BDFs.

10.1.9 Other Methods

Taylor-series methods and extrapolation schemes are two other classes of formulas that are sometimes used in practice, but much less frequently than Runge–Kutta or linear multistep formulas. See an advanced text such as Butcher (1987), Hairer et al. (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine

(1994) for a discussion of these and other classes of methods.

10.1.10 Adaptive Methods

Most good programs for the numerical solution of ODEs vary their step size – and possibly their order – in an attempt to solve the problem as efficiently as possible subject to a user-specified error tolerance, *tol*. The error that is controlled is usually the local error on each step, rather than the global error $y_n - y(x_n)$ that the user might at first expect. However, in most good programs the global error is at least roughly proportional to *tol*, so that reducing *tol* usually reduces the global error. A few codes report an estimate of the global error as well. If such an estimate is available, it is often optional, since estimating the global error frequently increases the cost of the integration significantly.

A useful way to interpret *tol* and the associated local error is in the *backward error* sense. (See Sec. 2.8 for a discussion of backward error analysis in the context of solving linear algebraic systems $Ax = b$.) When called to solve the IVP (44), many good programs generate a numerical solution that is the exact solution of the slightly perturbed problem

$$z'(x) = f(x, z(x)) + \delta(x), \quad z(a) = y_0, \tag{59}$$

where $\|\delta(x)\| \lesssim tol$. A few codes compute $\delta(x)$ explicitly and attempt to ensure that it is bounded by *tol*, but most satisfy (59) indirectly (some less reliably than others) by controlling some measure of the local error. For a more complete discussion of global errors, local errors, and their relationship to the perturbed equation (59), see an advanced text such as Butcher (1987), Hairer et al. (1987), Hairer and Wanner (1991), Lambert (1991), or Shampine (1994).

We believe that this backward error approach is often the most natural way to view the error in the numerical integration of an IVP. In many practical problems, we know $f(x, y)$ approximately only, possibly because of measurement errors or neglected terms in the model. Therefore, the true solution of the system satisfies an equation of the form (59), where in this case $\delta(x)$ is the error in the model. So, any solution to an IVP of the form (59) may be equally good provided $\|\delta(x)\|$ is less than the error in the model.

Programs for the numerical solution of ODEs often contain many other useful features. For example, some routines for nonstiff IVPs warn the user if the problem is stiff, while others automatically switch between stiff and nonstiff methods depending on the characteristics of the problem. Some programs contain sophisticated strategies to integrate problems with discontinuities in f or its derivatives much more efficiently and reliably than programs that do not attempt to detect discontinuities. Also, some programs return an interpolant for the numerical solution or allow the user to evaluate the numerical solution at very closely spaced points more efficiently than if the integration method itself produced all these output points. This can greatly increase the efficiency of codes when used to produce graphical output or to detect when the numerical solution satisfies some condition [such as $y(x) = c$ for some constant c].

10.2 Boundary-Value Problems (BVPs)

10.2.1 Shooting Methods

Shooting is conceptionally one of the simplest numerical techniques for solving

the boundary-value problem (BVP) (45). In its simplest form, often called *simple shooting*, we guess an initial condition $y(a) = y_0$ for the IVP (44) for the same ODE as the BVP (45), solve the IVP (44), and test whether the boundary condition $g(y_0, y(b; a, y_0)) = 0$ is satisfied, or nearly so, where $y(b; a, y_0)$ is the solution at $x = b$ of the IVP (44) with the initial condition $y(a) = y_0$. In most cases, the first guess for the initial condition $y(a) = y_0$ does not yield a $g(y_0, y(b; a, y_0))$ that is close enough to 0. So we must apply some root-finding technique to adjust the initial condition $y(a) = y_0$ until $g(y_0, y(b; a, y_0)) = 0$ is satisfied, or nearly so, assuming that there is a solution to the BVP.

Each time we adjust the initial condition, we must solve the IVP (44) again with the new initial condition $y(a) = y_0^{(l)}$ to compute $y(b; a, y_0^{(l)})$ and then $g(y_0^{(l)}, (b; a, y_0^{(l)}))$. For a scalar ODE ($m = 1$), we could try a simple technique such as bisection (see Sec. 6.4) to solve $g(y_0, y(b; a, y_0)) = 0$, but this converges slowly and so requires many solutions of the IVP (44) with different initial conditions $y_0^{(l)}$. Moreover, bisection is not applicable to systems of ODEs ($m > 1$).

The usual approach is to apply a variant of Newton's method (see Sec. 6.7) to solve $g(y_0, y(b; a, y_0)) = 0$. However, this requires that we compute an approximation to the Newton iteration matrix

$$\frac{dg(y_0^{(l)}, y(b; a, y_0^{(l)}))}{dy_0} = \frac{\partial g(y_0^{(l)}, y(b; a, y_0^{(l)})}{\partial y_a} + \frac{\partial g(y_0^{(l)}, y(b; a, y_0^{(l)}))}{\partial y_b} \frac{\partial y(b; a, y_0^{(l)})}{\partial y_0},$$

where $\partial g(y_0^{(l)}, y(b; a, y_0^{(l)}))/\partial y_a$ is the partial derivative of g with respect to its first argument, $\partial g(y_0^{(l)}, y(b; a, y_0^{(l)}))/\partial y_b$ is the partial derivative of g with respect to its second argument, and $\partial y(b; a, y_0^{(l)})/\partial y_0$ is the partial derivative of $y(b; a, y_0^{(l)})$ with respect to the initial condition $y(a) = y_0^{(l)}$. It can be shown that $\partial y(b; a, y_0^{(l)})/\partial y_0 = Y_l(b)$ for $Y_l : \mathbb{R} \to \mathbb{R}^{m\times m}$ the solution of the *variational equation*

$$Y_l'(x) = f_y(x, y_l(x))\, Y_l(x), \;\; Y_l(a) = I, \quad (60)$$

where $y_l(x)$ is the solution of the associated IVP (44) with initial condition $y(a) = y_0^{(l)}$ and $f_y(x, y) = \partial f(x, y)/\partial y \in \mathbb{R}^{m\times m}$ is the Jacobian of f. Therefore, on each iteration of Newton's method, we must solve the IVP (44) with initial condition $y(a) = y_0^{(l)}$ for $y_l(x)$ as well as the variational equation (60) associated with $y_l(x)$. Since it may take many iterations before we find a $y_0^{(l)}$ for which $g(y_0^{(l)}, y(b; a, y_0^{(l)}))$ is sufficiently close to 0, this is often a computationally expensive process.

Moreover, the associated IVP (44) may be unstable even though the BVP (45) is stable. As a result, simple shooting may break down or perform poorly. One way around this difficulty is to employ *multiple shooting*. In this scheme, we choose $N + 1$ shooting points $\{x_i : i = 0, \ldots, N\}$ satisfying $a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$, guess at N initial conditions s_i, $i = 0, \ldots, N - 1$, and solve the N IVPs

$$y_i' = f(x, y_i), \;\; x \in [x_i, x_{i+1}] \; i = 0, \ldots, N - 1, \; y_i(x_i) = s_i. \quad (61)$$

These IVPs are completely independent and so could be integrated simultaneously. Hence, this scheme is often called *parallel shooting*.

We need to adjust the initial conditions s_i, $i = 0, \ldots, N - 1$, so that

$$y_i(x_{i+1}) = s_{i+1}, \;\; i = 0, \ldots, N - 2, \quad (62)$$

$$g(s_0, y_N(b)) = 0, \quad (63)$$

where the first set of conditions (62) ensures $y_i(x_{i+1}) = y_{i+1}(x_{i+1})$ at the $N-1$ interior shooting points $x_1, \ldots, x_{N-1}$, thus allowing us to patch the functions $y_i(x)$ together into a continuous function $y(x)$ on $[a, b]$, and the second condition (63) enforces the boundary condition for the BVP (45).

A variant of Newton's method (see Sec. 6.7) is usually used to solve (62)–(63). The solution process is similar to, but somewhat more complicated than, that described above for simple shooting. It should be noted that the linear systems associated with Newton's method for (62)–(63) have a very special structure that can be exploited to great computational advantage. See an advanced text such as Ascher et al. (1988) for details.

Both simple and multiple shooting simplify significantly when applied to a linear ODE $y' = A(x)y + b(x)$. Newton's method converges in one iteration and the resulting scheme is equivalent to what is frequently called the method of *superposition*. If the boundary conditions are separated, this scheme simplifies still further. See an advanced text such as Ascher et al. (1988) for details.

Good shooting programs contain heuristics for choosing the shooting points and adjusting the tolerance for the IVP solver in an attempt to solve the BVP to within a user-specified tolerance. They also contain many other components, similar to those described in Sec. 10.1.10 for IVPs.

10.2.2 One-Step Methods

It is common to apply a one-step method, such as a Runge–Kutta (RK) formula, to solve the BVP (45). Since a collocation method applied to an ODE often reduces to a RK formula, this class of methods is broader than it might at first appear.

To simplify the discussion, assume that the one-step method can be written in the form

$$y_{n+1} = y_n + h_n \phi(x_n, y_n, h_n), \tag{64}$$

where $y_n \approx y(x_n)$, $h_n = x_{n+1} - x_n$, and $a = x_0 < x_1 < \ldots < x_N = b$. Note that the RK formula (55) is of this form with

$$\phi(x_n, y_n, h_n) = \sum_{i=1}^{s} b_i k_i,$$

$$k_i = f\left(x_n + c_i h_n, y_n + h_n \sum_{j=1}^{s} a_{ij} k_j\right). \tag{65}$$

To apply the one-step formula (64) to the BVP (45), we simply combine the equations (64) together with the boundary conditions to get a large system of equations

$$\Phi(y_0, \ldots, y_N) = \left\{ \begin{array}{c} y_{n+1} - y_n - h_n \phi(x_n, y_n, h_n), \\ n = 0, \ldots, N-1, \\ g(y_0, y_N) \end{array} \right\} = 0. \tag{66}$$

It is usual to apply a variant of Newton's method (see Sec. 6.7) to solve (66). As for shooting, the main difficulty here is to compute the $(N+1)m \times (N+1)m$ Newton iteration matrix $\partial\Phi(y_0, \ldots, y_N)/\partial(y_0, \ldots, y_N)$ and solve the associated linear system for the update to the approximate solution $y_0^{(l)}, \ldots, y_N^{(l)}$ to (66). See an advanced text such as Ascher et al. (1988) for a more complete discussion of this important point.

Good BVP codes contain heuristics for choosing the grid points to solve the BVP to within a user-specified tolerance. They also contain many other components, similar to those described in Sec. 10.1.10 for IVPs.

10.2.3 Other Methods

There are several other classes of numerical methods for BVPs for ODEs. Some of

these are discussed in Sec. 11 as numerical methods for BVPs for partial differential equations. An important class of methods, not discussed there, consists of *defect correction* schemes, including *deferred correction* as a special case. The basic idea behind these schemes is to apply a simple technique, possibly in the class discussed in the last subsection, and then estimate the *defect* or truncation error in the discretization and solve a related problem again with the same simple technique in an attempt to eliminate the error. See an advanced text such as Ascher et al. (1988) for further details.

11 Partial Differential Equations (PDEs)

A partial differential equation (PDE) is an equation in which the partial derivative of some order of the unknown function w.r.t. some independent variable occurs. For example,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = g(x, y) \tag{67}$$

is a PDE, where $u(x, y)$ is an unknown function, $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ denote the partial second derivatives of u w.r.t. x and y, respectively (often denoted by u_{xx} and u_{yy}, respectively), and $g(x, y)$ is a given function. The terms that involve u and its derivatives define the (partial differential) *operator* L, where, for example, $L = \partial^2/\partial x^2 + \partial^2/\partial y^2$ in (67), and the rest of the terms (usually the right side of the equation) form the *source term*.

11.1 Classes of Problems and PDEs

PDEs describe many important physical and technological phenomena. These phenomena can be divided into two basic types, which in turn are associated with two basic classes of problems for PDEs.

1. Equilibrium phenomena, elliptic PDEs, boundary-value problems. In steady-state phenomena, the equilibrium configuration u often satisfies

$$Lu = g \text{ in } \Omega, \tag{68}$$

$$Bu = \gamma \text{ on } \partial\Omega, \tag{69}$$

where Ω is a spatial N-dimensional domain, $\partial\Omega$ is the boundary of Ω, u is the unknown function of N variables, g and γ are known functions of N variables, and L and B are partial differential operators. Such problems are called *boundary-value problems* (BVPs). Often L is an elliptic operator. Equation (69) is frequently referred to as the *boundary condition* (BC). The definition of an elliptic operator in the general case is beyond the scope of this article, but some typical examples are given below.

2. Propagation phenomena, parabolic and hyperbolic PDEs, initial-value problems. In phenomena of a transient nature, the initial state is often given and we wish to predict the subsequent behavior. The function u at some point $t \in (0, T)$ frequently satisfies

$$Lu = g \text{ in } \Omega \times (0, T), \tag{70}$$

$$Bu = \gamma \text{ on } \partial\Omega \times (0, T), \tag{71}$$

while the initial configuration satisfies

$$Iu = g_0 \text{ in } \Omega \cup \partial\Omega, \tag{72}$$

where $(0,T)$ is the time interval of interest, Ω is a spatial N-dimensional domain, $\partial\Omega$ is the boundary of Ω, u is the unknown function of N spatial variables and one-time variable t, g and γ are known functions of N spatial variables and t, g_0 is a known

function of N spatial variables, and L, B, and I are partial differential operators. Such problems are called *initial-value problems* (IVPs). L is often either a parabolic or a hyperbolic operator (see below). Equation (72) is often referred to as an *initial condition* (IC).

11.1.1 Some Definitions

The *dimension* of a PDE is the number of independent variables in the PDE. The *order* of a PDE is the order of the highest derivative of the unknown function occurring in the PDE. A PDE is called *linear* if there are no nonlinear terms in the equation involving the unknown function or its derivatives; otherwise it is called *nonlinear*.

For example,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{j=1}^{N} b_j(x)\frac{\partial u}{\partial x_j} + c(x)u = d(x) \quad (73)$$

is N-dimensional, second order, and linear, where $x = (x_1, \ldots, x_N)$ is an N-dimensional vector of independent variables, $u(x)$ is the unknown function, and $\{a_{ij} : i, j = 1, \ldots, N\}$, $\{b_j : j = 1, \ldots, N\}$, c, and d are given functions of x. If any of the functions a_{ij}, b_j, or c depends on u or its derivatives, or if d is nonlinear in u or its derivatives, then the PDE is nonlinear.

A linear operator L is called *positive-definite* if $(Lu, u) > 0$ for all $u \neq 0$, where $(\cdot, \cdot)$ denotes an inner product (see Sec. 8.7). In addition, L is called *self-adjoint* if $(Lu, v) = (u, Lv)$ for all u and v in the associated function space. For example, the two-dimensional second-order linear PDE

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + fu = g \quad (74)$$

is self-adjoint if $d(x, y) = \partial a/\partial x$, $e(x, y) = \partial c/\partial y$ and $b(x, y) = 0$.

Consider the linear differential equation $Lu = g$, with L self-adjoint and positive-definite. Consider also the quadratic functional $F(u)$ defined by $F(u) = (Lu, u) - 2(u, g)$. The *minimum functional theorem* states that the solution of the differential equation $Lu = g$ coincides with the function u that minimizes $F(u)$. In general, when a solution to a differential equation corresponds to an extremum of a related functional, we have a *variational principle*. Numerical methods that construct approximations to the solution of a differential equation by using such a relationship are called *variational methods, Ritz methods, Rayleigh–Ritz methods*, or *energy methods*. The latter term comes from the observation that many variational methods are based on the physical principle of energy minimization.

The two-dimensional second-order linear PDE (74) is *elliptic, parabolic*, or *hyperbolic* if $D = b^2 - 4ac < 0, = 0$, or >0, respectively. For the definitions of these terms in the general case, the reader is referred to any introductory PDE book, such as Ames (1992), Celia and Gray (1992), or Hall and Porsching (1990). Typical examples of elliptic, parabolic, and hyperbolic PDEs are

- Laplace's equation, $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} + u_{x_Nx_N} = 0$, which is elliptic;
- the heat equation, $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} - u_{x_N} = 0$, which is parabolic; and
- the wave equation $u_{x_1x_1} + u_{x_2x_2} + \cdots + u_{x_{N-1}x_{N-1}} - u_{x_Nx_N} = 0$, which is hyperbolic.

Two other classical elliptic PDEs (given in two dimensions) are

- Poisson's equation, $u_{xx} + u_{yy} = f(x, y)$, and
- the Helmholtz equation, $u_{xx} + u_{yy} + \kappa u = f(x, y)$, where κ is a constant.

The *normal derivative* of a surface $u(x, y)$ is the rate of change of u along the direction of the outward normal (i.e., the direction perpendicular to the surface). Let α be the angle that the direction of the outward normal makes with the x axis at a point (x, y) on u. Then the normal derivative $\partial u/\partial n$ of u (often denoted by u_n) at the point (x, y) is $u_n = u_x \times \cos\alpha + u_y \sin\alpha$. The normal derivative can also be written as the inner product of the gradient of u, ∇u, with the unit outward normal vector, n. That is, $u_n = \nabla u \cdot n$. This definition of the normal derivative for two dimensions can be generalized easily to N dimensions.

11.1.2 Boundary Conditions

We now list some common types of boundary conditions (BCs) corresponding to the partial differential operator B in (69) or (71).

- *Dirichlet:* $Bu = u$.
- *Neumann:* $Bu = u_n$.
- *General* (for second-order PDEs): $Bu = \alpha(x)u + \beta(x)u_n$.
- *Mixed* (for second-order PDEs): Often, on parts of the boundary we have Dirichlet BCs and on the other parts Neumann ones. (Also, the term "mixed" may sometimes refer to more general types of BCs.)
- *Essential:* For PDEs of order $2m$, essential BCs involve u and its derivatives of order up to $m - 1$.
- *Natural:* For PDEs of order $2m$, natural BCs involve the derivatives of u from order m to $2m - 1$.

For further reading on the classification of PDE problems, operators, and boundary conditions, see Ames (1992), Celia and Gray (1992), or Hall and Porsching (1990).

11.2 Classes of Numerical Methods for PDEs

The two most commonly used methods for approximating the solution of PDEs are next described briefly.

Finite-Difference Methods (FDMs) have the following main steps:

1. Choose a finite-difference (FD) approximation of the derivatives involved in the PDE, BCs, and ICs. The result is a discretized PDE, BCs, and ICs.
2. Choose a set of n data points in the domain and on the boundary, on which the discretized PDE, BCs, and ICs must be satisfied. The result is a set of n equations w.r.t. the approximate values of u at the n data points.
3. Write the n equations of step **2** as a system and solve the system (discrete model). (If the PDE is linear, the system will usually be linear.) The solution is the approximate value of u at each of the n data points.
4. Evaluate the approximation to u at some point(s) of the domain (if needed).

Finite-Element Methods (FEMs) have the following main steps:

1. Choose a finite-element (FE) space, say n-dimensional, in which the approximation $u_{\boldsymbol{\Delta}}$ is constrained to belong, and a set of basis functions that span the space, say $\{\phi_i : i = 1, \ldots, n\}$. Then write

$$u_{\boldsymbol{\Delta}}(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

The unknown scalars α_i, $i = 1, \ldots, n$, are often called the *degrees of freedom* (DOF), or *coefficients*, of the FE representation of $u_{\boldsymbol{\Delta}}$.

2. Choose a set of n conditions that the approximation $u_{\boldsymbol{\Delta}}$ must satisfy. The result is a set of n equations w.r.t. the n coefficients of $u_{\boldsymbol{\Delta}}$.
3. Write the n equations of step **2** as a system and solve the system (discrete model). (If the PDE is linear, the system will usually be linear.) The solution is the vector of coefficients of $u_{\boldsymbol{\Delta}}$.
4. Evaluate the approximation to u at some point(s) of the domain.

11.2.1 Analysis of Numerical Methods for PDEs

Some common techniques used to analyze numerical methods for PDEs are discussed below. The analysis can be used to evaluate a method w.r.t. some chosen criteria or measures. In the following discussion, we use $u_{\boldsymbol{\Delta}}$ to denote the approximation to u computed by the method.

11.2.1.1 **Convergence Analysis (for BVPs and IVPs)** We study the behavior of the error $u - u_{\boldsymbol{\Delta}}$ as n increases. Assuming $\|u - u_{\boldsymbol{\Delta}}\| \to 0$ as $n \to \infty$, we can write $\|u - u_{\boldsymbol{\Delta}}\| \leq C(1/n)^{\alpha}$ for some constants C and α. The largest constant α for which this inequality holds is called the *order of convergence* of the method. As a first rough measure, the larger the α the better the method, as $(1/n)^{\alpha}$ will converge to 0 faster as $n \to \infty$ for larger α. To estimate the order of convergence of a method experimentally, we often devise PDE problems with known solutions and then solve them using the PDE method under investigation, first using n DOF, then $2n$ DOF, etc. We then plot $\|u - u_{\boldsymbol{\Delta}}\|$ versus n on a log–log scale. The slope of the plotted line is an approximation to the order of convergence of the method.

11.2.1.2 **Stability Analysis (for IVPs)** We study the behavior of the error $u - u_{\boldsymbol{\Delta}}$ as a function of t for increasing t. We often say that a method is *stable* if $\|u - u_{\boldsymbol{\Delta}}\|$ remains bounded as $t \to \infty$. Otherwise, it is called *unstable*. Or, we study how the error at some point in time propagates to the next point in time. In a stable method, the error is not amplified.

11.2.1.3 **Time (Computational) Complexity Analysis** We study the time that the method takes to compute the approximate solution to the PDE as a function of the n DOF. The time is usually proportional to the number of floating-point operations, although it also depends on the implementation and the hardware (computer) used. The most time-consuming part of a FDM or FEM is usually the third step (solution of the system), while the second most time-consuming part is usually the second step (generation of the system). For FDMs, the fourth step can also be time consuming, particularly if the value of the approximation at arbitrary points of the domain is required, since this computation requires interpolation, often using piecewise polynomials (PPs) or splines. The data to be interpolated are the approximate values of u at the grid points. Interpolation is not required in step **4** of a FEM, since the approximate solution can be evaluated at any point of the domain by the formula

$$u_{\boldsymbol{\Delta}}(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

By studying the particular implementation of a method, we are usually able to derive an approximate formula, such

as time $\approx Kn^{\beta}$, for some constant K, or time $= O(n^{\beta})$, relating the computational complexity to n. The smaller the β, the faster the method, and, among methods with the same β, the smaller the K, the faster the method.

11.2.1.4 **Memory Complexity Analysis** We study the memory (storage) requirements of a method as a function of the n DOF. These requirements depend on the storage scheme used for the matrix arising in step **2** and the solver used in step **3**.

11.2.1.5 **Overall Efficiency Analysis** Often, the most practical way of comparing two methods is to ask

1. if the methods were to run for the same length of time, which one would give the least error, or
2. given a certain error tolerance, which method satisfies that tolerance faster.

To test the overall efficiency of methods, we usually plot the error versus the time required to compute the approximate solution on a log–log scale. The method with the steepest slope is the most efficient.

11.3 Finite-Difference Methods for BVPs

A FD approximation to a derivative of a function u at a point x is a linear combination of values of u at points near x (often including x). Usually, a FD approximation is first derived for some derivative of a function of one variable, and then it is extended to partial derivatives of functions of several variables.

Let x be the point of interest and h, h_E, h_W small step sizes. The following are several examples of FD approximations in one dimension:

$$u_x(x) = \frac{u(x+h) - u(x)}{h} + O(h), \quad (75)$$

$$u_x(x) = \frac{u(x) - u(x-h)}{h} + O(h), \quad (76)$$

$$u_x(x) = \frac{u(x+h) - u(x-h)}{2h} + O(h^2), \quad (77)$$

$$u_x(x) = \frac{h_W^2 u(x+h_E) + (h_E^2 - h_W^2)u(x) - h_E^2 u(x-h_W)}{h_E(h_E+h_W)h_W} + O(h_E h_W), \quad (78)$$

$$u_{xx}(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2), \quad (79)$$

$$u_{xx}(x) = \frac{2h_W u(x+h_E) - 2(h_E+h_W)u(x) + 2h_E u(x-h_W)}{h_E(h_E+h_W)h_W} + O(h_E - h_W) + O([\max(h_E, h_W)]^2). \quad (80)$$

Let (x, y) be the point of interest and h, h_E, h_W, h_N, h_S small step sizes. The following are several examples of FD approximations in two dimensions.

$$u_{xx}(x, y) = \frac{u(x+h, y)-2u(x, y)+u(x-h, y)}{h^2}+O(h^2), \tag{81}$$

$$u_{xx}(x, y)+u_{yy}(x, y) = \frac{u(x+h, y)+u(x, y+h)-4u(x, y)+u(x, y-h)+u(x-h, y)}{h^2} + O(h^2), \tag{82}$$

$$u_{xy}(x) = \frac{u(x+h, y+h)-u(x-h, y+h)-u(x+h, y-h)+u(x-h, y-h)}{4h^2} + O(h^2), \tag{83}$$

$$u_{xy}(x) = \frac{u(x+h_E, y+h_N)-u(x-h_W, y+h_N)-u(x+h_E, y-h_S)+u(x-h_W, y-h_S)}{(h_E+h_W)(h_S+h_N)} + O(\max(h_E, h_W, h_S, h_N)). \tag{84}$$

Note the following.

1. Each FD approximation listed above includes an error term. The actual FD approximation is the right side of the equation excluding the error term.
2. Approximations (75)–(79) and (81)–(83) use *uniform* step sizes, while the rest use *nonuniform* step sizes.
3. Approximations (75), (76), (80), and (84) are of *first order*, while the rest are of *second order*. The order refers to the (lowest) exponent of the step size(s) in the error term.
4. All FD approximation formulas are derived by using appropriate Taylor-series expansions around the point of approximation.
5. Approximations (81)–(84) are derived by using combinations of one-dimensional Taylor series and make use of values of u at points on a *rectangular grid*.
6. It is possible to derive two-dimensional FD approximations that make use of values of u at points on a *triangular, quadrilateral* (but not rectangular), *polygonal*, or *irregular grid* with points positioned arbitrarily. Such approximations can be derived by using two-dimensional Taylor's series.

11.3.1 An Example of a Finite-Difference Method in One Dimension

Consider the problem

$$u_{xx} = g(x) \quad \text{in } (0, 1), \tag{85}$$

$$u = \gamma(x) \quad \text{at } x = 0 \quad \text{and } x = 1. \tag{86}$$

Using the FD approximation (79), we transform (85) to

$$[u(x+h) - 2u(x) + u(x-h)]/h^2 = g(x) + O(h^2). \tag{87}$$

Let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of *grid points* and let $U_i \approx u(x_i)$ for $i = 1, \ldots, n$. Without the $O(h^2)$ error term, the discretized PDE (87) at the grid point x_i becomes

$$\frac{(U_{i+1} - 2U_i + U_{i-1})}{h^2} = g(x_i). \tag{88}$$

From (86), we have for the point x_1

$$\frac{(U_2 - 2U_1)}{h^2} = g(x_1) - \frac{\gamma(x_0)}{h^2}. \qquad (89)$$

By writing equation (89) first, then (88) for $i = 2, \ldots, n-2$, and finally a relation similar to (89) for the point x_{n-1}, and then multiplying each equation by h^2, we get the following linear system:

$$\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ \vdots \\ U_{n-2} \\ U_{n-1} \end{pmatrix}$$

$$= h^2 \begin{pmatrix} g(x_1) \\ g(x_2) \\ g(x_3) \\ \vdots \\ g(x_{n-2}) \\ g(x_{n-1}) \end{pmatrix} - \begin{pmatrix} \gamma(x_0) \\ 0 \\ 0 \\ \vdots \\ 0 \\ \gamma(x_n) \end{pmatrix}. \qquad (90)$$

Note that this system is symmetric, diagonally dominant in all rows, and strictly diagonally dominant in the first and last rows. Therefore it is also positive-definite and has a unique solution. By solving it, we obtain $U_i \approx u(x_i)$, $i = 1, \ldots, n-1$. Using interpolation, we can approximate u at any other point of the domain.

It can be proved that $\max\{|u(x_i) - U_i| : i = 1, \ldots, n-1\} = O(h^2)$. That is, the approximation is second order at the grid points.

The computational complexity of the method described is $O(n)$, since the linear system (90) is tridiagonal and of size $n-1$ (see Sec. 2.7).

11.3.2 An Example of a Finite-Difference Method in Two Dimensions

Consider the problem

$$u_{xx} + u_{yy} = g(x, y) \quad \text{in } (0, 1) \times (0, 1), \qquad (91)$$

$$u = \gamma(x, y) \quad \text{on } x = 0, \quad x = 1, \quad y = 0, \quad y = 1. \qquad (92)$$

Using the FD approximation (82), we transform (91) to

$$\frac{u(x+h, y) + u(x, y+h) - 4u(x, y) + u(x, y-h) + u(x-h, y)}{h^2} = g(x, y) + O(h^2). \qquad (93)$$

Let $\{(x_i, y_j) : x_i = ih, y_j = jh, i, j = 0, \ldots, n\}$ with $h = 1/n$ be the set of grid points and let $U_{ij} \approx u(x_i, y_j)$. Without the $O(h^2)$ error term, the discretized PDE (93) at the grid point (x_i, y_j), $i = 1, \ldots, n-1, j = 1, \ldots, n-1$, becomes

$$\frac{(U_{i+1,j} + U_{i,j+1} - 4U_{i,j} + U_{i,j-1} + U_{i-1,j})}{h^2} = g(x_i, y_j). \qquad (94)$$

From (92), we have for the point (x_1, y_1)

$$\frac{(U_{1,2} + U_{2,1} - 4U_{1,1})}{h^2} = g(x_1, y_1) - \frac{[\gamma(x_0, y_1) + \gamma(x_1, y_0)]}{h^2}. \qquad (95)$$

Similar relations hold at the three other corners of the domain. Also, for the points $(x_1, y_j), j = 2, \ldots, n-2$, we have

$$\frac{(U_{1,j+1} - 4U_{1,j} + U_{1,j-1} + U_{2,j})}{h^2} = \frac{g(x_1, y_j) - \gamma(x_0, y_j)}{h^2}. \qquad (96)$$

Similar relations hold for other grid points one grid line away from the boundary.

One way to number the grid points, and also the equations and unknowns, is bottom up then left to right: $(1, 1), (1, 2), (1, 3), \ldots, (1, n-2), \quad (1, n-1), \quad (2, 1), (2, 2), (2, 3), \ldots, (2, n-2), (2, n-1), \ldots$.

That is, first (95), then (96) for $j = 2, \ldots, n-2$, then relations similar to (95) for the points (x_1, y_{n-1}), (x_2, y_1), then (94) for $i = 2, j = 2, \ldots, n-2$, etc. Using this ordering, we get a linear system $AU = \mathbf{g}$, where, after multiplying each equation by h^2, the matrix A has the form

$$\begin{bmatrix}
-4 & 1 & & & & 1 & & & & & & & & & & & & & & & & & & \\
1 & -4 & 1 & & & & 1 & & & & & & & & & & & & & & & & & \\
 & \ddots & & \ddots & & & & \ddots & & & & & & & & & & & & & & & & \\
 & & 1 & -4 & 1 & & & & 1 & & & & & & & & & & & & & & & \\
 & & & 1 & -4 & & & & & 1 & & & & & & & & & & & & & & \\
1 & & & & & -4 & 1 & & & & 1 & & & & & & & & & & & & & \\
 & 1 & & & & 1 & -4 & 1 & & & & 1 & & & & & & & & & & & & \\
 & & \ddots & & & & \ddots & \ddots & \ddots & & & & \ddots & & & & & & & & & & & \\
 & & & \ddots & & & & \ddots & \ddots & \ddots & & & & \ddots & & & & & & & & & & \\
 & & & & 1 & & & & 1 & -4 & 1 & & & & 1 & & & & & & & & & \\
 & & & & & 1 & & & & 1 & -4 & & & & & 1 & & & & & & & & \\
 & & & & & & 1 & & & & & -4 & 1 & & & & & & & & & & & \\
 & & & & & & & 1 & & & & 1 & -4 & 1 & & & & & & & & & & \\
 & & & & & & & & \ddots & & & & \ddots & \ddots & \ddots & & & & & & & & & \\
 & & & & & & & & & 1 & & & & 1 & -4 & 1 & & & & & & & & \\
 & & & & & & & & & & 1 & & & & 1 & -4 & & & & & & & &
\end{bmatrix}.$$

This can be rewritten as

$$\begin{pmatrix}
T & I & & & \\
I & T & I & & \\
 & \ddots & \ddots & \ddots & \\
 & & I & T & I \\
 & & & I & T
\end{pmatrix},$$

where T is a tridiagonal matrix of size $n-1$ with -4's on the diagonal and 1's on the super- and subdiagonal and I is the identity matrix of size $n-1$. Thus the matrix is *block-tridiagonal* of size $(n-1)^2$. The vector of unknown is

$$\begin{aligned}
U = (&U_{1,1}, U_{1,2}, \ldots, U_{1,n-2}, U_{1,n-1}, \\
&U_{2,1}, U_{2,2}, \ldots, U_{2,n-2}, U_{2,n-1}, \ldots \\
&U_{n-2,1}, U_{n-2,2}, \ldots, U_{n-2,n-2}, \\
&U_{n-2,n-1}, \\
&U_{n-1,1}, U_{n-1,2}, \ldots, U_{n-1,n-2}, \\
&U_{n-1,n-1})^{\mathrm{T}}.
\end{aligned}$$

The right-side vector $\mathbf{g}$ is:

$$\begin{aligned}
\mathbf{g} = (&h^2 g_{1,1} - \gamma_{0,1} - \gamma_{1,0}, \\
&h^2 g_{1,2} - \gamma_{0,2}, \ldots, \\
&h^2 g_{1,n-2} - \gamma_{0,n-2}, \\
&h^2 g_{1,n-1} - \gamma_{0,n-1} - \gamma_{1,n}, \\
&h^2 g_{2,1} - \gamma_{2,0}, h^2 g_{2,2}, \ldots, h^2 g_{2,n-2}, \\
&h^2 g_{2,n-1} - \gamma_{2,n}, \ldots, \\
&h^2 g_{n-2,1} - \gamma_{n-2,0}, \\
&h^2 g_{n-2,2}, \ldots, h^2 g_{n-2,n-2}, \\
&h^2 g_{n-2,n-1} - \gamma_{n-2,n}, \\
&h^2 g_{n-1,1} - \gamma_{n,1} - \gamma_{n-1,0}, \\
&h^2 g_{n-1,2} - \gamma_{n,2}, \ldots,
\end{aligned}$$

$$h^2 g_{n-1,n-2} - \gamma_{n,n-2},$$
$$h^2 g_{n-1,n-1} - \gamma_{n,n-1} - \gamma_{n-1,n})^{\mathrm{T}},$$

where $g_{ij} = g(x_i, y_j)$ and $\gamma_{ij} = \gamma(x_i, y_j)$. Note that this system is symmetric, diagonally dominant in all rows, and strictly diagonally dominant in all rows corresponding to grid points one grid line away from the boundary. Therefore, it is positive-definite and has a unique solution. By solving the system $AU = \mathbf{g}$, we obtain $U_{ij} \approx u(x_i, y_j)$ for $i = 1, \ldots, n-1$ and $j = 1, \ldots, n-1$. Using interpolation, we can approximate the value of u at any other point of the domain.

It can be proved that $\max\{|u(x_i, y_j) - U_{i,j}| : i, j = 1, \ldots, n-1\} = O(h^2)$. That is, the approximation is second order at the grid points.

The computational complexity of the method described above depends on the method used to solve the linear system $AU = \mathbf{g}$. Note that A has at most five nonzero entries per row, it is banded with lower and upper bandwidth $n-1$, and its size is $(n-1)^2$. If a direct band solver is used to solve $AU = \mathbf{g}$, then the computational complexity of the method is $O(n^4)$, but sparse direct solvers are more efficient (see Sec. 2.7). In addition, there exist iterative methods (e.g., multigrid; see Sec. 11.8 and Briggs, 1987) that can solve this system much more efficiently, reducing the computational complexity of the method to almost $O(n^2)$.

Note that the properties of the matrix A, such as symmetry, diagonal dominance, positive-definiteness, and the sparsity pattern (block-tridiagonal with at most five nonzero entries per row), are highly dependent on the simplicity of the differential operator associated with (91) and boundary conditions (92), the choice of uniform and rectangular grid, and the FD approximation (93). For a differential operator with first-order derivative terms and/or Neumann BCs, symmetry is lost. Symmetry may also be lost if a nonuniform grid is chosen, even if it is rectangular. Diagonal dominance depends on the coefficients of the differential operator and on the absence of first-order derivative terms. The block-tridiagonal form will most likely be affected if an irregular grid is chosen. Fast linear solvers, such as multigrid and FFT (fast Fourier transform) solvers, work well on the matrix A, but may not perform as well on more general systems. The development of fast linear solvers for such matrices is an open and active area of research. See, for example, Van Loan (1992) or Hackbusch (1994) and the references therein. For further reading on FDMs, see Strikwerda (1989).

11.4 Finite-Element Methods for BVPs

The first step in a FEM is to choose a FE approximation space and a basis for it. The most commonly used spaces are piecewise polynomials (PPs) or splines (see Sec. 8.5). Let n be the dimension of the approximation space and let $\{\phi_j(x) : j = 1, \ldots, n\}$ be a set of basis functions for the space.

Consider the problem (68)–(69). Let

$$u_{\Delta}(x) = \sum_{j=1}^{n} \alpha_j \phi_j(x)$$

be the approximation to u. The next step in a FEM is to choose n conditions that the approximation must satisfy. A FEM is characterized by these conditions. The most common FEMs are the *Galerkin method* and the *collocation method*.

11.4.1 The Galerkin Method

Given an inner product $(\cdot,\cdot)$, usually defined by

$$(f,g) = \int_{\Omega} f(x)g(x)\,dx,$$

we require that u_{Δ} satisfies

$$(\phi_i,\ Lu_{\Delta} - g) = 0, \quad i = 1,\ldots,n, \quad (97)$$

forcing the *residual* $Lu_{\Delta} - g$ to be orthogonal to the approximation space, and making it, in a sense, as "small" as possible. If L is a linear operator, then the relations (97) are equivalent to

$$\sum_{j=1}^{n} \alpha_j(\phi_i, L\phi_j) = (\phi_i, g), \quad i = 1,\ldots,n, \quad (98)$$

which can be written in the form $A\alpha = \mathbf{g}$, where A is an $n \times n$ matrix with entries $A_{ij} = (\phi_i, L\phi_j)$, $i = 1,\ldots,n$, $j = 1,\ldots,n$, $\alpha = (\alpha_1,\ldots,\alpha_n)^{\mathrm{T}}$ is the vector of coefficients, and $\mathbf{g}$ is a vector with entries $g_i = (\phi_i, g)$, $i = 1,\ldots,n$. We usually use numerical integration to compute the entries of A and $\mathbf{g}$ (see Sec. 9).

As an example of the Galerkin method in one dimension, consider the problem (85)–(86) and the set of grid points $\{x_i = ih : i = 0,\ldots,n\}$ with $h = 1/n$. Let $\{\phi_i : i = 0,\ldots,n\}$ be the set of linear spline basis functions w.r.t. the knots (grid points) $\{x_i\}$, as defined by (39). Then

$$u_{\Delta}(x) = \sum_{i=0}^{n} \alpha_i\phi_i(x)$$

is the linear spline approximation to u. From the BC (86), we get $u_{\Delta}(x_0) = \gamma(x_0)$, which implies $\alpha_0 = \gamma(x_0)$. Similarly, $\alpha_n = \gamma(x_n)$. The remaining unknowns $\{\alpha_i : i = 1,\ldots,n-1\}$ are determined by the $n-1$ Galerkin conditions

$$\sum_{j=0}^{n} \alpha_j(\phi_i, L\phi_j) = (\phi_i, g), \quad i = 1,\ldots,n-1,$$

which, for the particular L associated with (85), are equivalent to

$$\sum_{j=0}^{n} \alpha_j(\phi_i, \phi_j'') = (\phi_i, g), \quad i = 1,\ldots,n-1.$$

Writing the inner product (ϕ_i, ϕ_j'') as an integral and applying integration by parts, these conditions reduce to

$$\begin{aligned}\sum_{j=1}^{n-1} \alpha_j \int_0^1 \phi_i'\phi_j' dx &= \left[\phi_i \sum_{j=0}^{n} \alpha_j\phi_j'\right]_0^1 \\ &- \int_0^1 \phi_i g dx - \alpha_0 \int_0^1 \phi_i'\phi_0' dx \\ &- \alpha_n \int_0^1 \phi_i'\phi_n'\,dx, \quad i = 1,\ldots,n-1.\end{aligned} \quad (99)$$

Note that the term

$$\left[\phi_i \sum_{j=0}^{n} \alpha_j\phi_j'\right]_0^1 = 0,$$

since $\phi_i(0) = \phi_i(1) = 0$ for $i = 1,\ldots,n-1$. Relations (99) form a linear system of size $n-1$; the associated matrix A has elements

$$A_{i,j} = \int_0^1 \phi_i'\phi_j'\,dx.$$

Since the basis functions $\{\phi_i\}$ are nonzero on at most two subintervals, A is tridiagonal with elements

$$A_{i,i} = \int_{x_{i-1}}^{x_{i+1}} \phi_i'\phi_i'\,dx, \quad i = 1,\ldots,n-1,$$

$$A_{i,i-1} = \int_{x_{i-1}}^{x_i} \phi_i' \phi_{i-1}' \, dx, \quad i = 2, \ldots, n-1,$$

$$A_{i,i+1} = \int_{x_i}^{x_{i+1}} \phi_i' \phi_{i+1}' \, dx, \quad i = 1, \ldots, n-2.$$

It can be proved that this matrix is also symmetric positive-definite. Thus, the associated system has a unique solution. By solving the system, we obtain the coefficients $\{\alpha_i : i = 0, \ldots, n\}$ of $u_{\boldsymbol{\Delta}}$, which we can evaluate at any point of the domain $(0, 1)$.

It can be proved that $\max\{|u(x) - u_{\boldsymbol{\Delta}}(x)| : x \in [0, 1]\} = O(h^2)$. That is, the approximation is second order on the whole domain.

The computational complexity of the method described is $O(n)$, since the linear system that has to be solved is tridiagonal and of size $n - 1$ (see Sec. 2.7).

Relations (99) can also be derived using a variational method (see Sec. 11.1.1). Thus, for problem (85)–(86), there is a variational method equivalent to the Galerkin method. This is true for every differential equation problem with a self-adjoint positive-definite operator. There exist differential equation problems, though, that are not characterized by variational principles. In such cases, the Galerkin method is applicable, while the variational method is not.

As an example in two dimensions, consider problem (91)–(92) with the grid points $\{(x_i, y_j) : x_i = ih, y_j = jh, i, j = 0, \ldots, n\}$ for $h = 1/n$. A common way to define an approximation space for two-dimensional problems is to choose a tensor product of approximation spaces in each dimension. Let $\{\phi_i(x) : i = 0, \ldots, n\}$ be the linear spline basis functions w.r.t. the knots $\{x_i : i = 0, \ldots, n\}$ and let $\{\phi_j(y) : j = 0, \ldots, n\}$ be the linear spline basis functions w.r.t. the knots $\{y_j : j = 0, \ldots, n\}$, as defined in (39). Then

$$u_{\boldsymbol{\Delta}}(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_{ij} \phi_i(x) \phi_j(y)$$

is the bilinear spline approximation to u. Continuing as in the one-dimensional case, we derive a system of $(n+1)^2$ equations in $(n+1)^2$ unknowns. The associated matrix A is block-tridiagonal, with at most nine non-zero entries per row and bandwidth $n + 2$. It is also symmetric positive-definite. Thus, the associated system has a unique solution. Moreover, it can be proved that the approximation $u_{\boldsymbol{\Delta}}$ is second order.

Note that, if instead of a rectangular subdivision of the domain and bilinear elements, we choose a triangular subdivision and linear elements (w.r.t. x and y), we would get a system similar to that of Sec. 11.3.2.

An important property of the Galerkin method is that, for any self-adjoint positive-definite differential operator, the resulting matrix is symmetric positive-definite, even if the grid is irregular. As stated before, for every differential equation problem with a self-adjoint and positive-definite operator, there is a variational method equivalent to the Galerkin method. This holds for higher-dimension problems too. Thus, large, sparse, symmetric, positive-definite matrices arise from the application of variational methods.

For an introduction to the FEM, including its computer implementation, see Becker et al. (1981). An error analysis is carried out in Strang and Fix (1973).

11.4.2 The Collocation Method

We first pick n *collocation points* $\{t_i : i = 1, \ldots, n\}$ in Ω and on $\partial\Omega$. We then require

that $u_{\boldsymbol{\Delta}}$ satisfies

$$Lu_{\boldsymbol{\Delta}}(t_i) - g(t_i) = 0, \quad \text{if } t_i \in \Omega, \quad (100)$$

$$Bu_{\boldsymbol{\Delta}}(t_i) - \gamma(t_i) = 0, \quad \text{if } t_i \in \partial\Omega, \quad (101)$$

forcing the residuals $Lu_{\boldsymbol{\Delta}} - g$ and $Bu_{\boldsymbol{\Delta}} - \gamma$ to be zero at the collocation points, and making them, in a sense, as "small" as possible. If L and B are linear, relations (100)–(101) are equivalent to

$$\sum_{j=1}^{n} \alpha_j L\phi_j(t_i) = g(t_i), \quad \text{if } t_i \in \Omega, \quad (102)$$

$$\sum_{j=1}^{n} \alpha_j B\phi_j(t_i) = \gamma(t_i), \quad \text{if } t_i \in \partial\Omega, \quad (103)$$

which can be written in the form $A\alpha = \mathbf{g}$, where A is an $n \times n$ matrix with entries $A_{ij} = L\phi_j(t_i), j = 1, \ldots, n$, for all $t_i \in \Omega$, and $A_{ij} = B\phi_j(t_i), j = 1, \ldots, n$, for all $t_i \in \partial\Omega$; $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ is the vector coefficients; and $\mathbf{g}$ is a vector with entries $g_i = g(t_i)$ for all $t_i \in \Omega$ and $g_i = \gamma(t_i)$ for all $t_i \in \partial\Omega$.

The choice of collocation points is critical to the success of the method. It affects not only the solvability and other properties (such as symmetry, diagonal dominance, bandedness) of the matrix A but also the accuracy of the approximation $u_{\boldsymbol{\Delta}}$. Depending on the FE approximation space that $u_{\boldsymbol{\Delta}}$ belongs to, some standard choices of collocation points in one dimension are listed below.

1. If the FE approximation space is the space of quadratic splines (quadratic PPs in $\mathscr{C}^1$), the collocation points are chosen to be the midpoints of the subintervals (x_{i-1}, x_i), $i = 1, \ldots, n$, and the two boundary points. The same choice of collocation points is effective if the FE approximation space is composed of any other even-degree splines, with the exception that some additional collocation conditions may be required at boundary points or points close to the boundary.
2. If the FE approximation space is the space of cubic splines (cubic PPs in $\mathscr{C}^2$), the collocation points are chosen to be the grid points $\{x_i : i = 0, \ldots, n\}$. At each of the boundary points, x_0 and x_n, both conditions (100) and (101) are imposed. The same choice of collocation points is effective if the FE approximation space is composed of any other odd-degree splines, with the exception that some additional collocation conditions may be required at boundary points or points close to the boundary.
3. If the FE approximation space is the space of cubic PPs in $\mathscr{C}^1$ (cubic Hermite PPs), the collocation points are chosen to be the two Gauss points $x_{i-1} + (3 \pm \sqrt{3}) \times (x_i - x_{i-1})/6$ in each subinterval (x_{i-1}, x_i), $i = 1, \ldots, n$, and the two boundary grid points.

As an example in one dimension, consider problem (85)–(86) and the set of grid points $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$. Let the collocation points be the midpoints $t_i = (x_{i-1} + x_i)/2$, $i = 1, \ldots, n$, and the end points $t_0 = x_0$ and $t_{n+1} = x_n$. Let $\{\phi_i : i = 0, \ldots, n+1\}$ be the quadratic spline basis functions w.r.t. the knots (grid points) $\{x_i\}$, as defined in (40). Then

$$u_{\boldsymbol{\Delta}}(x) = \sum_{i=0}^{n+1} \alpha_i \phi_i(x)$$

is the quadratic spline approximation to u. Relation (100) for the PDE (85) becomes $u''_{\boldsymbol{\Delta}}(t_i) = g(t_i)$, and so relation (102) becomes

$$\alpha_{i-1}\phi''_{i-1}(t_i) + \alpha_i\phi''_i(t_i) + \alpha_{i+1}\phi''_{i+1}(t_i) = g(t_i),$$

which reduces to

$$(\alpha_{i-1} - 2\alpha_i + \alpha_{i+1})/h^2 = g(t_i),$$
$$i = 1, \ldots, n. \tag{104}$$

Relation (101) for the BC (86) becomes $u_{\boldsymbol{\Delta}}(t_0) = \gamma(t_0)$, and so relation (103) becomes

$$\alpha_0\phi_0(t_0) + \alpha_1\phi_1(t_0) = \gamma(t_0),$$

which reduces to

$$\frac{(\alpha_0 + \alpha_1)}{2} = \gamma(t_0). \tag{105}$$

Similarly, the collocation condition at $t_{n+1} = 1$ reduces to

$$\frac{(\alpha_n + \alpha_{n+1})}{2} = \gamma(t_{n+1}). \tag{106}$$

Writing (105) first, then (104) for $i = 1, \ldots, n$, and finally (106), we get a tridiagonal system of equations w.r.t. the coefficients $\{\alpha_i : i = 0, \ldots, n+1\}$. The system is diagonally dominant and it can be proved that it has a unique solution. It can also be scaled so that it is symmetric positive-definite.

It can be proved that $\max\{|u(x) - u_{\boldsymbol{\Delta}}(x)| : x \in [0, 1]\} = O(h^2)$. That is, the approximation is second order on the whole domain. There exists a variant of this method, though, that is fourth order at the grid points and midpoints and third order on the whole domain (Houstis et al., 1988).

The computational complexity of the method described above is $O(n)$, since the linear system that must be solved is tridiagonal and of size $n + 1$ (see Sec. 2.7).

As an example in two dimensions, consider problem (91)–(92) and the set of grid points $\{(x_i, y_j) : x_i = ih, y_j = jh, i, j = 0, \ldots, n\}$ with $h = 1/n$. A common approximation space for two-dimensional problems is a tensor product of approximation spaces in each dimension. Let $\{\phi_i(x) : i = 0, \ldots, n+1\}$ be the quadratic spline basis functions w.r.t. the knots (grid points) $\{x_i : i = 0, \ldots, n\}$ and let $\{\phi_j(y) : j = 0, \ldots, n+1\}$ be the quadratic spline basis functions w.r.t. the knots $\{y_j : j = 0, \ldots, n\}$, as defined in (40). Then

$$u_{\boldsymbol{\Delta}}(x, y) = \sum_{i=0}^{n+1}\sum_{j=0}^{n+1} \alpha_{ij}\phi_i(x)\phi_j(y)$$

is the biquadratic spline approximation to u. Continuing as in the one-dimensional case, we derive a system of $(n+2)^2$ equations and unknowns. The associated matrix is block-tridiagonal, with at most nine nonzero entries per row, and has bandwidth $n + 3$. It can be proved that this system has a unique solution and that the approximation $u_{\boldsymbol{\Delta}}$ is second order. With appropriate modifications, though, the order can be improved as in the one-dimensional case (Christara, 1994).

For a general introduction to collocation methods, see Prenter (1975).

11.5 Finite-Difference Methods for IVPs

Consider the problem (70)–(72). Let the temporal grid points be $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$. Starting with the initial values of u at t_0, given by (72), most FDMs for IVPs compute approximate values of u at each subsequent temporal grid point t_j, in the order $j = 1, \ldots, m$, using previous and/or current approximate values of u at neighboring space points.

If at each temporal grid point t_j a method uses only approximations from previous temporal grid points, it is called *explicit*, as it does not require the solution of a system of equations to proceed from one temporal

grid point to the next. If at some temporal grid point t_j a method uses approximations from the current temporal grid point t_j, it is called *implicit*, as it requires the solution of a system of equations to proceed from one temporal grid point to the next. If at the time step from t_{j-1} to t_j a method uses approximations from t_{j-1} and t_j only, it is called *one-step*. Likewise, we can define *two-step* methods, etc. These definitions for PDEs are similar to those given in Sec. 10 for ODEs.

11.5.1 An Example of an Explicit One-Step Method for a Parabolic IVP

Consider the problem

$$u_t = u_{xx} \text{ in } (0, 1) \times (0, T), \tag{107}$$

$$u = \gamma_0(t) \text{ on } x = 0, t \in (0, T), \tag{108}$$

$$u = \gamma_1(t) \text{ on } x = 1, t \in (0, T), \tag{109}$$

$$u = g(x) \text{ on } t = 0, x \in [0, 1]. \tag{110}$$

Using the FD approximations (79) for u_{xx} and (75) for u_t, we transform (107) to

$$\frac{[u(x, t + h_t) - u(x, t)]}{h_t} = \frac{[u(x + h, t) - 2u(x, t) + u(x - h, t)]}{h^2} + O(h_t + h^2). \tag{111}$$

Let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i, t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then the discretized PDE (111) at the point (x_i, t_j), $i = 1, \ldots, n-1$, $j = 1, \ldots, m$, becomes

$$\frac{(U_{i,j+1} - U_{i,j})}{h_t} = \frac{(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})}{h^2}.$$

Letting $r = h_t/h^2$, we can rewrite this relation as

$$U_{i,j+1} = rU_{i+1,j} + (1 - 2r)U_{i,j} + rU_{i-1,j} \tag{112}$$

for $i = 2, \ldots, n-2$. For $i = 1$, we have from (108)

$$U_{1,j+1} = rU_{2,j} + (1 - 2r)U_{1,j} + r\gamma_0(t)_j. \tag{113}$$

Similarly, for $i = n - 1$, we have from (109)

$$U_{n-1,j+1} = r\gamma_1(t_j) + (1 - 2r)U_{n-1,j} + rU_{n-2,j}. \tag{114}$$

For $j = 1$, we have from (110)

$$U_{i,1} = rg(x_{i+1}) + (1 - 2r)g(x_i) + rg(x_{i-1}). \tag{115}$$

Thus, we can compute $U_{i,j} \approx u(x_i, t_j)$ from a linear combination of three neighboring spatial approximations at t_{j-1}.

It can be proved that if $r < \frac{1}{2}$, then $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n, j = 1, \ldots, m\} = O(h^2 + h_t)$; thus the order of convergence is one w.r.t. to h_t and two w.r.t. h. It can also be proved that if $r < \frac{1}{2}$, the method is stable. However, the restriction $r < \frac{1}{2}$ may be impractical for many problems, since it forces h_t to be very small if h is small and so the method must take many steps to integrate the problem.

The computational complexity of the method is $O(nm)$, since for each grid point (x_i, t_j) a constant number of floating-point operations must be performed.

11.5.2 An Example of an Implicit One-Step Method for a Parabolic IVP

Consider the problem (107)–(110) once more. Using the FD approximations (79) for u_{xx} and (76) for u_t, we transform

(107) to

$$\frac{[u(x,t+h_t)-u(x,t)]}{h_t} = \frac{[u(x+h,t+h_t)-2u(x,t+h_t) + u(x-h,t+h_t)]}{h^2} + O(h_t+h^2). \quad (116)$$

Again let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i, t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then, the discretized PDE (116) at the point (x_i, t_j), $i = 1, \ldots, n-1$, $j = 1, \ldots, m$, becomes

$$\frac{(U_{i,j+1} - U_{i,j})}{h_t} = \frac{(U_{i+1,j+1} - 2U_{i,j+1} + U_{i-1,j+1})}{h^2}.$$

Letting $r = h_t/h^2$ again, we can rewrite this relation as

$$-rU_{i-1,j+1} + (1+2r)U_{i,j+1} - rU_{i+1,j+1} = U_{i,j} \quad (117)$$

for $i = 2, \ldots, n-2$. For $i = 1$, we have from (108)

$$(1+2r)U_{1,j+1} - rU_{2,j+1} = U_{1,j} + r\gamma_0(t_{j+1}). \quad (118)$$

Similarly, for $i = n-1$, we have from (109)

$$-rU_{n-2,j+1} + (1+2r)U_{n-1,j+1} = U_{n-1,j} + r\gamma_1(t_{j+1}). \quad (119)$$

For $j = 1$, we have from (110)

$$-rU_{i-1,1} + (1+2r)U_{i,1} - rU_{i+1,1} = g(x_i). \quad (120)$$

Thus, at the jth time step, a tridiagonal linear system must be solved to compute $U_{i,j} \approx u(x_i, t_j)$. The diagonal entries of the associated matrix are all equal to $1+2r$, while the off-diagonal entries are all equal to $-r$. The system is symmetric positive-definite and strictly diagonally dominant; thus it has a unique solution.

It can be proved that $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n,\ j = 1, \ldots, m\} = O(h^2 + h_t)$; thus the order of convergence is one w.r.t. to h_t and two w.r.t. h. It can also be proved that the method is stable without any restrictions on r (except $r > 0$). The computational complexity of the method is $O(nm)$, since at each time step we must solve a tridiagonal linear system of size $n-1$ (see Sec. 2.7).

Note that for the problem (107)–(110), which is one-dimensional w.r.t. to space, both the explicit and implicit methods have the same computational complexity. This is not true for problems in more space dimensions. For such problems, the solution of a linear system at each time step can be very time consuming, making an implicit method much more expensive per step than an explicit one. However, because there is no restriction on r for some implicit schemes, while there always is for an explicit one, some implicit schemes may be able to take far fewer time steps than an implicit one. As a result, an implicit method may be computationally more efficient than an explicit one.

11.5.3 An Example of an Explicit Two-Step Method for a Hyperbolic IVP

Consider the problem

$$u_{tt} = u_{xx} \text{ in } (0,1) \times (0,T), \quad (121)$$

$$u = \gamma_0(t) \text{ on } x = 0,\ t \in (0,T), \quad (122)$$

$$u = \gamma_1(t) \text{ on } x = 1,\ t \in (0,T), \quad (123)$$

$$u = \gamma_0(x) \text{ on } t = 0,\ x \in (0,1), \quad (124)$$

$$u_t = \gamma_1(x) \text{ on } t = 0,\ x \in (0,1). \quad (125)$$

Using the FD approximation (79) for u_{xx} and u_{tt}, we transform (121) to

$$\frac{[u(x, t+h_t) - 2u(x, t) + u(x, t-h_t)]}{h_t^2} = \frac{[u(x+h, t) - 2u(x, t) + u(x-h, t)]}{h^2} + O(h_t^2 + h^2). \quad (126)$$

Again let $\{x_i = ih : i = 0, \ldots, n\}$ with $h = 1/n$ be the set of spatial grid points and $\{t_j = jh_t : j = 0, \ldots, m\}$ with $h_t = T/m$ be the set of temporal grid points. Also let $U_{i,j} \approx u(x_i, t_j)$ for $i = 0, \ldots, n$ and $j = 0, \ldots, m$. Then, the discretized PDE (126) at the point (x_i, t_j), $i = 1, \ldots, n-1$, $j = 1, \ldots, m$, becomes

$$\frac{(U_{i,j+1} - 2U_{i,j} + U_{i,j-1})}{h_t^2} = \frac{(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})}{h^2}.$$

Letting $r = h_t/h$, we can rewrite this relation as

$$U_{i,j+1} = r^2 U_{i-1,j} + 2(1 - r^2) U_{i,j} + r^2 U_{i+1,j} - U_{i,j-1} \quad (127)$$

for $i = 2, \ldots, n-1$. For grid points close to the boundary, the approximate values of U are replaced by the values of the functions γ_0 and γ_1 at the appropriate points, as in Secs. 11.5.1 and 11.5.2. Thus, we can compute $U_{i,j+1} \approx u(x_i, t_{j+1})$ from a linear combination of three neighboring spatial approximations at time t_j and one approximation at time t_{j-1}.

Since (127) is a two-step formula, at the initial time point t_0 it cannot be applied as is. At that point, we use the ICs (124)–(125) and the FD approximation (75) to get

$$U_{i,1} = g_0(x_i) + h_t g_1(x_i). \quad (128)$$

It can be proved that, if $r < 1$, then the method is stable. It can also be shown that $\max\{|u(x_i, t_j) - U_{i,j}| : i = 1, \ldots, n, j = 1, \ldots, m\} = O(h^2 + h_t^2)$; thus the order of convergence is two w.r.t. to both h_t and h. Note that the restriction $r < 1$ is not impractical in this case, since it requires only that $h_t < h$. The computational complexity of the method is $O(nm)$, since for each grid point (x_i, t_j) we apply a formula with a constant number of floating-point operations.

11.6 The Method of Lines

The general idea behind the *method of lines* (MOL) is to use an ODE solver along one of the dimensions of the PDE, while using a PDE discretization across the other dimensions. In its most common form for the solution of IVPs for PDEs, an ODE solver is used along the temporal dimension, while a PDE discretization is employed across the spatial dimensions, transforming an IVP for a PDE into a system of IVPs for ODEs.

To see how this is done, consider the problem (107)–(110) again. Let

$$u_{\boldsymbol{\Delta}}(x, t) = \sum_{j=1}^{n} \alpha_j(t)\phi_j(x)$$

be a FE approximation to the true solution $u(x, t)$. Now apply a FEM condition to $u_{\boldsymbol{\Delta}}$ to discretize the PDE (107) w.r.t. the spatial dimension. For example, collocation at the points x_i, $i = 1, \ldots, n$, yields

$$\sum_{j=1}^{n} \alpha_j'(t)\phi_j(x_i) = \sum_{j=1}^{n} \alpha_j(t)\phi_j''(x_i).$$

Let $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_n(t))^{\mathrm{T}}$, Φ be the matrix with entries $\Phi_{ij} = \phi_j(x_i)$, and A be the matrix with entries $A_{ij} = \phi_j''(x_i)$.

Then the PDE (107) is approximated by the system of ODEs $\Phi\boldsymbol{\alpha}'(t) = A\boldsymbol{\alpha}(t)$.

To obtain an IC for the ODE, we construct an interpolant $g_{\boldsymbol{\Delta}}$ of g in the same space as that spanned by $\{\phi_j : j = 1, \ldots, n\}$. Let

$$g_{\boldsymbol{\Delta}}(x) = \sum_{j=1}^{n} \beta_i \phi_i(x)$$

be the FE representation of $g_{\boldsymbol{\Delta}}$ in that space and set $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$. Then

$$\Phi\boldsymbol{\alpha}'(t) = A\boldsymbol{\alpha}(t), \tag{129}$$

$$\boldsymbol{\alpha}(0) = \boldsymbol{\beta} \tag{130}$$

is a well-defined IVP for ODEs. Thus, the PDE problem (107)–(110) is converted to an IVP for a system of n ODEs. The latter can be solved by the techniques described in Sec. 10.

Note that applying an ODE method to discretize the IVP (129)–(130) results in a discretization for the PDE (107)–(110). That is, the MOL produces a discretization for a PDE. However, it is generally agreed that standard software for ODEs is more highly developed than for PDEs. Thus, using the MOL to decouple the discretization of the spatial and temporal variables allows us to exploit easily sophisticated time-stepping techniques. As a result, the MOL is often the simplest effective method to solve a PDE.

11.7 Boundary-Element Methods

The general idea behind *boundary-element methods* (BEMs) is to transform the PDE to an integral equation in which the integrations take place along the boundary only of the PDE domain, thus eliminating the need for domain discretization and reducing the dimension of the PDE by one. For example, a one-dimensional integral equation is solved instead of an equivalent two-dimensional PDE. The BEM is applicable to BVPs for Laplace's or Poisson's equation, and many other simple PDEs. If applicable, this approach is often very effective, especially when the PDE domain is highly irregular.

11.8 The Multigrid Method

The multigrid method (MM) exploits the connection between a physical problem and its matrix analog to accelerate the convergence of an iterative method (see Sec. 3). For simplicity, we describe the MM for the one-dimensional problem (85)–(86), although the merits of the scheme become apparent for two- and higher-dimensional problems (see Sec. 11.3.2). Also, we illustrate the technique using Jacobi's method as the basic iterative scheme. The MM, though, can be used with many other iterative methods as a preconditioning technique.

Let A be the matrix in (90). Apply Jacobi's method (see Sec. 3.1) with an extra damping factor of 2 to the linear system (90). The associated iteration matrix is $G = I - A/4$. It can be shown that the eigenvalues of G are $\mu_i = \cos^2[i\pi/(2n)]$, $i = 1, \ldots, n-1$, and that the components of the eigenvector v_i associated with μ_i are $\sin[i\pi(j/n)]$, $j = 1, \ldots, n-1$. These are also the eigenvectors of A. Since $\{v_i : i = 1, \ldots, n-1\}$ spans $\mathbb{R}^{n-1}$, we can write the error e_0 associated with the initial guess for the damped Jacobi iteration as

$$e_0 = \sum_{i=1}^{n-1} \alpha_i v_i$$

for some scalars $\{\alpha_i : i = 1, \ldots, n-1\}$. It then follows easily from the discussion in

Sec. 3.1 that the error at iteration k is

$$e_k = \sum_{i=1}^{n-1} \alpha_i \mu_i^k v_i.$$

The terms of the sum corresponding to small values of i are called *low-frequency components*, while those corresponding to large values of i are called *high-frequency components*. Note that $0 < \mu_{n-1} < \mu_{n-2} < \cdots < \mu_2 < \mu_1 < 1$. Moreover, $\mu_1 \approx 1 - (\pi/2n)^2$, while $\mu_{n-1} \approx (\pi/2n)^2$. Consequently, $e_k \to 0$ as $k \to \infty$, but the low-frequency components of the error converge slowly, while the high-frequency components converge rapidly.

To accelerate the convergence of the low-frequency components of the error, consider solving the problem (90) on a coarse grid with $\hat{n} = n/2$ subintervals and $\hat{n} + 1$ grid points, assuming for simplicity that n is even. Although the coarse grid has about half the number of grid points of the fine one, the $\hat{n} - 1$ eigenvectors of the matrix $\hat{A}$ for the coarse grid provide a good representation of the low- to middle-frequency eigenvectors of A. As a result, the solution to the problem (90) on the coarse grid provides a good approximation to the low- to middle-frequency components of the fine-grid solution. This suggests that the coarse-grid solution can be used to provide good approximations to the low- to middle-frequency components of the fine-grid solution, while the damped Jacobi iteration on the fine grid can be used to provide good approximations to the middle- to high-frequency components of the fine-grid solution.

This is the motivation behind the MM. The term "multigrid" refers to the use of several levels of grids (possibly a fine grid, several intermediate-level grids, and a coarse grid), so that each level damps certain components of the error fast.

To view the MM as a preconditioning technique, consider the linear system $Au = g$ corresponding to the discretization of problem (85)–(86) on some fine grid. Apply one (or a few) damped Jacobi iteration(s) to $Au = g$ to obtain an approximate solution vector $\tilde{u}$. Let $r = g - A\tilde{u}$ be the *residual* vector. Project r to a coarse grid to obtain the *coarse-grid residual* vector $\hat{r}$. This can be done by appropriately interpolating the components of r and evaluating the interpolant at the points of the coarse grid (see Sec. 8). Let $\hat{A}$ be the matrix corresponding to the discretization of problem (85)–(86) on the coarse grid. Solve (or approximately solve) $\hat{A}\tilde{r} = \hat{r}$. This can be done by applying a few damped Jacobi iterations, or by recursively applying the MM to $\hat{A}\tilde{r} = \hat{r}$, or by using a direct solver (see Sec. 2), since $\hat{A}$ is a smaller matrix than A. Now $\tilde{r}$ is the *preconditioned coarse-grid residual* vector. Extend $\tilde{r}$ to the fine grid to obtain the *preconditioned fine-grid residual* vector $\bar{r}$. This can be done by appropriately interpolating the components of $\tilde{r}$ and evaluating the interpolant at the points of the fine grid. Add $\bar{r}$ to $\tilde{u}$ to obtain a new approximate solution vector. Repeat the process until convergence. Usually, only a few iterations are needed. Note that this scheme has some similarities to iterative improvement, as described in Sec. 2.9.

The power of the MM lies in the fact that the coarse grid, which acts as a preconditioner, allows the information to pass from a point of the problem domain to another point in a few steps, while the fine grid maintains the accuracy required. Note that the interpolation and the evaluation of the interpolant needed for the projection of a fine-grid vector to a coarse-grid vector and for the extension of a coarse-grid vector to a fine-grid

vector often reduce to simple relations, such as averaging neighboring vector components. A practical introduction to the MM, including an error analysis, can be found in Briggs (1987).

12 Parallel Computation

The increasing demand by scientists and engineers to solve larger and larger problems constitutes the primary motivation for parallel computation. Another impetus is the cost effectiveness of computers consisting of many standard CPUs compared with those based on one very fast CPU.

A parallel computer has the ability to execute simultaneously many different processes by having several independent processors concurrently perform operations on different data. A parallel computer in which all processors perform the same operation on different data is called a *single-instruction, multiple-data* (SIMD) machine, while one in which each processor has the ability to perform different operation(s) on data is called a *multiple-instruction, multiple-data* (MIMD) machine. Sometimes a processor may use data computed by another processor, either by exchanging messages with it or by sharing some common memory area. The former is characteristic of *distributed-memory* or *message-passing* machines, while the latter is characteristic of *shared-memory* machines.

A vector computer may be viewed as a restricted form of a parallel machine. On such a computer, one arithmetic operation is performed by several (usually a few) processors, which cooperate in a pipelined manner. Each processor receives an input operand from another processor, performs a part of the operation on it, and passes the result on to the next processor. While a processor performs its part of the operation on an operand, the next processor performs its part of the operation on the previous operand, etc. This pipelining technique is very effective when the same operation is performed on many data – for example, on each component of a long vector.

To utilize parallel computers effectively, we must be able to split the computation into *parallel processes* to be assigned to different processors. To this end, it is desirable to have fully independent computations in each process. For many problems, though, this is impossible, but there are techniques to minimize the dependence of the computation of one process on another. It is also desirable to have almost equal amounts of computation in each process for *load balancing*. This is also hard to accomplish in many cases, but again there exist techniques to achieve reasonable load balancing.

Many computational scientists study the parallelization of existing numerical methods, as well as the development of new methods appropriate for parallel machines. In the next subsection, we discuss, as a simple example, a numerical method that is effective for solving tridiagonal linear systems on parallel computers.

12.1 Cyclic Reduction

Consider solving the tridiagonal system

$$\begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & b_3 & a_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & & b_n & a_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}$$

$$= \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

by the following LU factorization algorithm for tridiagonal matrices:

for $k = 1, \ldots, n-1$ do

$b_{k+1} = b_{k+1}/a_k$

$a_{k+1} = a_{k+1} - b_{k+1}c_k$

end

At the end of the computation, the modified b's form the subdiagonal of the unit lower-triangular matrix L and the modified a's and c's form the diagonal and superdiagonal, respectively, of the upper-triangular matrix U (see Secs. 2.3 and 2.7).

Note that the computation proceeds in the order $b_2, a_2, b_3, a_3, \ldots$. Each value computed depends on the previous one. Therefore, it seems that the computation is purely sequential and that there is no easy way to parallelize it. However, there are other ways to solve tridiagonal linear systems, and some of them can be implemented effectively on a parallel machine.

Assume, for simplicity, that $n = 2^q - 1$, where q is a positive integer. Multiply row 1 by b_2/a_1 and subtract it from row 2, eliminating x_1 from row 2. Also multiply row 3 by c_2/a_3 and subtract it from row 2, eliminating x_3 from row 2. The new row 2 involves variables x_2 and x_4 only.

Repeat the process described above for the $(n-1)/2$ groups of rows (3, 4, 5), (5, 6, 7), etc. This eliminates the odd unknowns from the even equations. The even equations form a new tridiagonal linear system of about half the size, $(n-1)/2 = 2^{q-1} - 1$, called the *reduced* system. This technique is often called *odd–even reduction*.

Now apply odd–even reduction to the reduced system to obtain another reduced system that is again about half as big as the first reduced system. The recursive application of odd–even reduction continues for $q = \log_2(n+1)$ steps. At each step, the even equations of the previous step form a reduced tridiagonal system of about half the size of the system from the previous step. At the end of step q, one equation in one unknown remains, so that unknown can be computed easily. This recursive technique is often called *cyclic reduction*.

Then the computation continues in the reverse order with a process called *back substitution*. At each step of back substitution, the even variables are known from the solution of the associated reduced system of about half the size. Substituting these values back into the odd equations of the larger system, we can easily compute all the odd variables.

Both the cyclic-reduction algorithm and the back-substitution algorithm require $q = \log_2(n+1)$ steps each. In cyclic reduction, the number of floating-point operations is divided by 2 at each step, starting with $O(n)$ floating-point operations in the first step. Thus, it requires $O(n \log n)$ arithmetic operations. Similarly, back substitution also requires $O(n \log n)$ arithmetic operations. Therefore, the computational complexity of the full solution is $O(n \log n)$.

Observe that the algorithm described above is highly parallel. The elimination operations applied to a group of three rows to obtain the reduced system at each step are independent of the elimination operations applied to any other group of three rows, and so can be

carried out in parallel. Similarly, the substitution operations to compute the odd unknowns in a reduced system given the even ones are independent of each other. Thus, the unknowns of each back-substitution step can also be computed in parallel.

Assume that we have $p = (n-1)/2$ processors. Initially, processor 1 is assigned rows (1, 2, 3), processor 2 is assigned rows (3, 4, 5), processor 3 is assigned rows (5, 6, 7), and so on. After the first odd–even reduction step, processor 2 will use equation 2 from processor 1, equation 4 from itself, and equation 6 from processor 3. Similarly, processor 4 will use equation 6 from processor 3, equation 8 from itself, and equation 10 from processor 5, and so on. Only the even processors will continue. The procedure is repeated. The final reduced system is solved by one processor. For the back substitution, one processor works first, then two, then four, and so on.

Thus, the algorithm requires $2\log_2(n+1)$ steps with a constant amount of computation done on each processor per step, So the parallel computational complexity of the algorithm is $O(\log n)$, which is a factor of $O(n) = O(p)$ improvement over the $O(n \times \log n)$ computational complexity of the serial version of the algorithm, and a little less than $O(n)$ improvement over the $O(n)$ serial computational complexity of the standard LU factorization algorithm for tridiagonal systems. So we can say that, asymptotically, the algorithm has perfect *speedup*.

Note that when a processor uses rows computed by another processor, some communication and/or synchronization must take place between processors. This may degrade the parallel performance of the algorithm from the perfect asymptotic performance. The time spent in communication and/or synchronization depends heavily on the way the processors cooperate. More specifically, it depends on the interconnection network between processors and on the implementation of specific hardware instructions.

For an introduction to parallel numerical methods, see Bertsekas and Tsitsiklis (1989), Ortega (1988), or Van de Velde (1994).

13
Sources of numerical software

Although most of this article has dealt with elementary numerical methods, we strongly recommend that readers do not program these schemes themselves. High-quality software incorporating these – or more sophisticated – numerical methods is readily available. In addition, good library routines often contain many additional strategies and heuristics (not discussed here) to improve their efficiency and reliability. Using such routines, rather than attempting to reprogram them, will likely save readers a significant amount of time as well as produce superior numerical results.

We highly recommend that readers familiarize themselves with the *Guide to Available Mathematical Software* (GAMS) recently developed by the National Institute of Standards and Technology (NIST). GAMS is both an on-line cross-index of available mathematical software as well as a repository of some 9000 high-quality problem-solving modules from more than 80 software packages. It provides centralized access to such items as abstracts, documentation, and source code of the software modules that it catalogs. Most of this software represents FORTRAN subprograms for mathematical problems that

commonly occur in computational science and engineering, such as solution of systems of linear algebraic equations, computing matrix eigenvalues, solving nonlinear systems of differential equations, finding minima of nonlinear functions of several variables, evaluating the special functions of applied mathematics, and performing nonlinear regression. Among the packages cataloged in GAMS are

- the IMSL, NAG, PORT, and SLATEC libraries;
- the BLAS, EISPACK, FISHPAK, FNLIB, FFTPACK, LAPACK, LINPACK, and STARPAC packages;
- the DATAPLOT and SAS statistical analysis systems;
- the netlib routines, including the Collected Algorithms of the ACM (see below).

Note that although GAMS catalogs both public-domain and proprietary software, source code of proprietary software is not available through GAMS, although related items such as documentation and example programs often are. Software can be found either by browsing through a decision tree or performing a key-word search. GAMS can be accessed in several ways: telnet gams.nist.gov; gopher gams.nist.gov; or ⟨www browser⟩ http://gams.nist.gov, where ⟨www browser⟩ is a World Wide Web browser such as Mosaic or Netscape. Report any questions or problems to gams@cam.nist.gov. For more details, log in to the system or see Boisvert et al. (1985) and Boisvert (1990).

Included in the software cataloged by GAMS are many high-quality public-domain routines available by electronic mail (e-mail) from *netlib*. These routines are now also available through Xnetlib, a more sophisticated X interface to netlib and the NA-Net Whitepages, or through the World Wide Web at the address http://www.netlib.org/index.html. For more information on netlib, see Dongarra and Grosse (1987) or Dongarra et al. (1995), send the message "send index" by e-mail to either netlib@ornl.gov or netlib@research.att.com, or access http://www.netlib.org/index.html through the World Wide Web.

The ACM Transactions on Mathematical Software publishes refereed public-domain software. These high-quality routines, covering a broad range of problem areas, are included in the Collected Algorithms of the ACM, available through both GAMS and netlib.

Not mentioned above are the commercial interactive packages MATLAB, Maple, and Mathematica. MATLAB is built upon a foundation of sophisticated matrix software and includes routines for solving many standard mathematical and statistical problems. In addition, "toolboxes" for several application areas, such as control theory, are available. Both Maple and Mathematica are primarily symbolic algebra packages, but contain many high-quality numerical routines as well. For more information on MATLAB, contact The MathWorks Inc., 24 Prime Park Way, Natick, MA 01760; phone: (508) 653-1415; FAX: (508) 653-2997; e-mail: info@mathworks.com. For more information on Maple, contact Waterloo Maple Software, 450 Phillip St., Waterloo, Ontario, Canada, N2L 5J2; phone: (519) 747-2373; FAX: (519) 747-5284; e-mail: info@maplesoft.on.ca. For more information on Mathematica, contact Wolfram Research Inc., 100 Trade Center Dr., Champaign, IL 61820-7237; phone: (217) 398-0700; FAX: (217) 398-0747; e-mail: info@wri.com.

Glossary

Banded Matrix: An $m \times n$ matrix all of whose nonzero elements occur in a band around its main diagonal.

Block Diagonal: See **Diagonal**.

Characteristic Equation of a Matrix: Defined only for an $n \times n$ matrix A, the equation $\det(a - \lambda I)$.

Characteristic Polynomial of a Matrix: Defined only for an $n \times n$ matrix A, the polynomial $p(\lambda) = \det(A - \lambda I)$, of degree n.

Column-Diagonally Dominant: Describing a matrix A if A^{T} is **row-diagonally dominant**.

Diagonal: Describing an $m \times n$ matrix $D = [d_{ij}]$ for which $d_{ij} = 0$ for $i \neq j$. The matrix D is *block-diagonal* if each d_{ij} in the preceding definition is a submatrix rather than a single number.

Diagonally Dominant: Describing a matrix A if either A or A^{T} is **row-diagonally dominant**.

Eigenvector: Defined only for an $n \times n$ matrix A, a (possibly complex) number λ such that for some nonzero vector x, $Ax = \lambda x$. Any vector satisfying this equation is an *eigenvector* associated with the eigenvalue λ.

Euclidean Norm (of a vector *x*): The quantity $\|x\|_2 = (x^{\mathrm{H}}x)^{1/2}$, where x^{H} is the complex-conjugate transpose of the vector x. Often called the *2-norm*.

Flop: A floating-point operation on a computer; a multiplication and either an addition or a subtraction.

Hermitian: Describing an $n \times n$ matrix A for which $A^{\mathrm{H}} = A$, where A^{H} is the complex-conjugate transpose of A.

Hessenberg: Either **upper** or **lower Hessenberg**. A symmetric Hessenberg matrix is tridiagonal.

Leading Principal Minor of a Matrix *A*: The $k \times k$ submatrix in the top left corner of A.

Lower Hessenberg: Describing a matrix A for which $a_{ij} = 0$ for $i < j - 1$. That is, it is lower-triangular except for a single non-zero superdiagonal.

Lower Triangular: Describing an $n \times n$ matrix $L = [l_{ij}]$ for which $l_{ij} = 0$ for $1 \leq i < j \leq n$. It is *strictly lower-triangular* if $l_{ij} = 0$ for $1 \leq i \leq j \leq n$, and *block lower-triangular* or *block strictly lower-triangular*, respectively, if each l_{ij} in the preceding definitions is a submatrix rather than a single number.

Orthogonal: Describing an $m \times n$ matrix Q for which $Q^{\mathrm{T}}Q = I$.

Permutation Matrix: An $n \times n$ matrix with exactly one element in each row and column equal to 1 and all other elements equal to 0.

Rank of a Matrix: The maximal number of independent rows (or columns) of the matrix.

Right Triangular: Describing an $m \times n$ matrix $R = [r_{ij}]$ such that $r_{ij} = 0$ for $i > j$. If $m = n$, the terms *right triangular* and *upper triangular* are equivalent.

Row-Diagonally Dominant: Describing an $m \times n$ matrix, with $m \leq n$, for which

$$\sum_{j=1, j \neq i}^{n} |a_{ij}| < |a_{ii}|$$

for $i = 1, 2, \ldots, m$.

Similar Matrices: Any two matrices A and B such that $B = WAW^{-1}$ for some nonsingular matrix W, which is called the *associated similarity transformation*.

Sparse: Describing a matrix in which the number of nonzero elements is much less than the total number of elements in the matrix.

Spectral Radius of a (Square) Matrix *A*: The quantity $\rho(A) = \max\{|\lambda| : \lambda \text{an eigenvalue of } A\}$.

Strictly Lower Triangular: See **Lower Triangular**.

Strictly Upper Triangular: See **Upper Triangular**.

Symmetric: Describing an $n \times n$ matrix for which $A^{\mathrm{T}} = A$, where A^{T} is the transpose of A.

Symmetric Indefinite: Describing a real symmetric matrix A for which $x^{\mathrm{T}}Ax > 0$ for some real n-vector x and $y^{\mathrm{T}}Ay < 0$ for some real n-vector y.

Symmetric Positive (Negative) Definite: Describing a real symmetric matrix A for which $x^{\mathrm{T}}Ax > 0 (x^{\mathrm{T}}Ax < 0)$ for all real n-vectors $x \neq 0$.

Symmetric Positive (Negative) Semidefinite: Describing a real symmetric matrix A for which $x^{\mathrm{T}}Ax \geq 0 (x^{\mathrm{T}}Ax \leq 0)$ for all real n-vectors $x \neq 0$.

Transpose: For an $m \times n$ matrix $A = [a_{ij}]$, the $n \times m$ matrix $A^{\mathrm{T}} = [a^{\mathrm{t}}_{ij}]$ where $a^{\mathrm{t}}_{ij} = a_{ji}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. The transpose of a column (row) vector is a row (column) vector.

Unit Lower Triangular: Describing a lower-triangular matrix $L = [l_{ij}]$ for which $l_{ii} = 1$ for $i = 1, 2, \ldots, n$.

Upper Hessenberg: Describing a matrix $A = [a_{ij}]$ such that $a_{ij} = 0$ for $i > j + 1$. That is, it is upper-triangular except for a single nonzero subdiagonal.

Upper Triangular: An $n \times n$ matrix $U = [u_{ij}]$ such that $u_{ij} = 0$ for $1 \leq j < i \leq n$. It is *strictly upper-triangular* if $u_{ij} = 0$ for $1 \leq j \leq i \leq n$, and *block upper-triangular* or *block strictly upper-triangular*, respectively, if each u_{ij} in the preceding definitions is a submatrix rather than a single number.

Tridiagonal: Describing an $m \times n$ matrix $A = [a_{ij}]$ such that $a_{ij} = 0$ for $|i - j| > 1$.

Mathematical Symbols Used

A^{H}: The complex-conjugate transpose of the matrix A.

A^{T}: The transpose of the matrix A.

$\mathbb{C}$: The set of complex numbers.

$\mathbb{C}^n$: The set of complex vectors with n components.

$\mathbb{C}^{m \times n}$: The set of complex $m \times n$ matrices.

$\mathscr{C}$: The set of continuous functions.

$\mathscr{C}[a, b]$: The set of continuous functions on the interval $[a, b]$.

$\mathscr{C}^p$: The set of continuous functions with p continuous derivatives.

$\mathscr{C}^p[a, b]$: The set of continuous functions with p continuous derivatives on the interval $[a, b]$.

$O(h^p)$: Any quantity that depends on h that can be bounded above by Ch^p for some constant C and all $h \in (0, H]$ for some $H > 0$.

$O(n^p)$: Any quantity that depends on n that can be bounded above by Cn^p for some constant C and all positive integers n.

$\mathbb{R}$: The set of real numbers.

$\mathbb{R}^n$: The set of real vectors with n components.

$\mathbb{R}^{m\times n}$: The set of real $m \times n$ matrices.

x^{T}: The transpose of the vector x.

$\|x\|$ and $\|A\|$: Norms of the vector x and the matrix A, respectively.

$\|x\|_2$ and $\|A\|_2$: Euclidean norms (also called two-norms) of the vector x and the matrix A, respectively.

z^{H}: The complex-conjugate transpose of the vector z.

$\rho(A)$: The spectral radius of a square matrix A.

Abbreviations Used

ACM: Association for Computing Machinery.

ADI: Alternating direction implicit.

BC: Boundary condition.

BVP: Boundary-value problem.

CD: Conjugate direction.

CG: Conjugate gradient.

DOF: Degrees of freedom.

FD: Finite difference.

FDM: Finite-difference method.

FE: Finite element.

FEM: Finite-element method.

GAMS: Guide to Available Mathematical Software.

GE: Gaussian elimination.

IC: Initial condition.

ICF: Incomplete Cholesky factorization.

IEEE: Institute of Electrical and Electronics Engineers.

IVP: Initial-value problem.

LMF: Linear multistep formula.

MM: Multigrid method.

ODE: Ordinary differential equation.

PCG: Preconditioned conjugate gradient.

PDE: Partial differential equation.

PP: Piecewise polynomial.

RK: Runge–Kutta.

SD: Steepest descent.

SOR: Successive overrelaxation.

SPD: Symmetric positive-definite.

SSOR: Symmetric successive overrelaxation.

w.r.t.: With respect to.

List of Works Cited

Ames, W. F. (1992), *Numerical Methods for Partial Differential Equations*, New York: Academic.

Ascher, U. M., Mattheij, R. M. M., Russell, R. D. (1988), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Englewood Cliffs, NJ: Prentice-Hall.

Atkinson, K. E. (1989), *An Introduction to Numerical Analysis*, 2nd ed., New York: Wiley.

Axelsson, O. (1994), *Iterative Solution Methods*, Cambridge, U.K.: Cambridge Univ. Press.

Becker, E. B., Carey, G. F., Oden, J. T. (1981), Finite Elements, Vol. I, Englewood Cliffs, NJ: Prentice-Hall.

Bertsekas, D. P., Tsitsiklis, J. N. (1989), *Parallel and Distributed Computation: Numerical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Boisvert, R. (1990), "The Guide to Available Mathematical Software Advisory System," in: E. Houstis, J. Rice, R. Vichnevetsky (Eds.), *Intelligent Mathematical Software Systems*, Amsterdam: North-Holland, pp. 167–178.

Boisvert, R. F., Howe, S. E., Kahaner, D. K. (1985), "GAMS: A Framework for the Management of Scientific Software," *ACM Trans. Math. Software* **11** (4), 313–355.

Briggs, W. L. (1987), *A Multigrid Tutorial*, Philadelphia: SIAM.

Buchanan, J. L., Turner, P. R. (1992), *Numerical Methods and Analysis*, New York: McGraw-Hill.

Butcher, J. C. (1987), *The Numerical Analysis of Ordinary Differential Equations*, New York: Wiley.

Celia, M. A., Gray, W. G. (1992), *Numerical Methods for Differential Equations*, Englewood Cliffs, NJ: Prentice-Hall.

Christara, C. C. (1994), "Quadratic Spline Collocation Methods for Elliptic Partial Differential Equations," *BIT* **34** (1), 33–61.

Conte, S. D., de Boor, C. (1980), *Elementary Numerical Analysis*, 3rd ed., New York: McGraw-Hill.

Cullum, J., Willoughby, R. (1985), *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. 1, Theory, and Vol. 2, Programs, Boston: Birkhäuser.

Dahlquist, G., Björck, Å. (1974), *Numerical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Davis, P. J. (1975), *Interpolation and Approximation*, New York: Dover.

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

Dennis, J. E., Schnabel, R. B. (1983), *Numerical Methods for Unconstrained Optimisation and Nonlinear Equations*, Englewood Cliffs, NJ: Prentice-Hall.

Dongarra, J., Grosse, E. (1987), "Distribution of Mathematical Software via Electronic Mail," *Commun. ACM* **30** (5), 403–407.

Dongarra, J., Rowan, T., Wade, R. (1995), "Software Distribution Using XNETLIB," *ACM Trans. Math. Software* **21** (1), 79–88.

Duff, I., Erisman, A., Reid, J. (1986), *Direct Methods for Sparse Matrices*, Oxford, U.K.: Oxford Univ. Press.

George, A., Liu, J. (1981), *Computer Solution of Large Sparse Positive Definite Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Goldberg, D. (1991), "What Every Computer Scientist Should Know about Floating-Point Arithmetic," *ACM Comput. Surveys* **23**, 5–48.

Golub, G. H., Van Loan, C. F. (1989), *Matrix Computations*, 2nd ed., Baltimore: John Hopkins Univ. Press.

Hackbusch, W. (1994), *Iterative Solution of Large Sparse Systems of Equations*, New York: Springer-Verlag.

Hageman, L. A., Young, D. M. (1981), *Applied Iterative Methods*, New York: Academic.

Hager, W. W. (1988), *Applied Numerical Linear Algebra*, Englewood Cliffs, NJ: Prentice-Hall.

Hairer, E., Nørsett, S. P., Wanner, G. (1987), *Solving Ordinary Differential Equations I: Nonstiff Problems*, New York: Springer-Verlag.

Hairer, E., Wanner, G. (1991), *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, New York: Springer-Verlag.

Hall, C. A., Porsching, T. A. (1990), *Numerical Analysis of Partial Differential Equations*, Englewood Cliffs, NJ: Prentice-Hall.

Householder, A. (1970), *The Numerical Treatment of a Single Nonlinear Equation*, New York: McGraw-Hill.

Houstis, E. N., Christara, C. C., Rice, J. R. (1988), "Quadratic Spline Collocation Methods for Two-Point Boundary Value Problems," *Int. J. Numer. Methods Eng.* **26**, 935–952.

IEEE (1985), *IEEE Standard for Binary Floating-Point Arithmetic*, New York: American National Standards Institute, Std. 754-1985.

Johnson, L. W., Riess, R. D. (1982), *Numerical Analysis*, Reading, MA: Addison-Wesley.

Kahaner, D., Moler, C., Nash, S. (1989), *Numerical Methods and Software*, Englewood Cliffs, NJ: Prentice-Hall.

Lambert, J. D. (1991), *Numerical Methods for Ordinary Differential Equations*, New York: Wiley.

Ortega, J. M. (1988), *Introduction to Parallel and Vector Solution of Linear Systems*, New York: Plenum.

Parlett, B. N. (1968), "Global Convergence of the Basic QR Algorithm on Hessenberg Matrices," *Math. Comput.* **22**, 803–817.

Parlett, B. N. (1980), *The Symmetric Eigenvalue Problem*, Englewood Cliffs, NJ: Prentice-Hall.

Prenter, P. M. (1975), *Splines and Variational Methods*, New York: Wiley.

Saad, Y. (1992), *Numerical Methods for Large Eigenvalue Problems*, New York: Manchester Univ. Press (Wiley).

Scott, D. (1981) "The Lanczos Algorithm," in: I. S. Duff (Ed.), *Sparse Matrices and Their Uses*, London: Academic, pp. 139–160.

Shampine, L. F. (1994), *Numerical Solution of Ordinary Differential Equations*, New York: Chapman & Hall.

Shampine, L. F., Gear, C. W. (1979), "A User's View of Solving Stiff Ordinary Differential Equations," *SIAM Rev.* **21**, 1–17.

Stoer, J., Bulirsch, R. (1980), *Introduction to Numerical Analysis*, New York: Springer-Verlag.

Strang, G., Fix, G. J. (1973), *An Analysis of the Finite Element Method*, Englewood Cliffs, NJ: Prentice-Hall.

Strikwerda, J. C. (1989), *Finite Difference Schemes and Partial Differential Equations*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Van de Velde, E. F.. (1994), *Concurrent Scientific Computing*, New York: Springer-Verlag.

Van Loan, C. F. (1992), *Computational Frameworks for the Fast Fourier Transform*, Philadelphia: SIAM.

Varga, R. S. (1962), *Matrix Iterative Analysis*, Englewood Cliffs, NJ: Prentice-Hall.

Wilkinson, J. H. (1965), *The Algebraic Eigenvalue Problem*, Oxford, U.K.: Oxford Univ. Press.

Young, D. M. (1971), *Iterative Solution of Large Linear Systems*, New York: Academic.

Further Reading

Ciarlet, P. G. (1989), *Introduction to Numerical Linear Algebra and Optimization*, Cambridge, U.K.: Cambridge Univ. Press.

Forsythe, G. E., Malcolm, M. A., Moler, C. B. (1977), *Computer Methods for Mathematical Computations*, Englewood Cliffs, NJ: Prentice-Hall.

Forsythe, G. E., Moler, C. B. (1967), *Computer Solution of Linear Algebraic Systems*, Englewood Cliffs, NJ: Prentice-Hall.

Golub, G. H., Ortega, J. M. (1992), *Scientific Computing and Differential Equations*, New York: Academic.

Golub, G. H., Ortega, J. M. (1993), *Scientific Computing: An Introduction with Parallel Computing*, New York: Academic.

Isaacson, E., Keller, H. B. (1966), *Analysis of Numerical Methods*, New York: Wiley.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, U.K.: Cambridge Univ. Press.

Schultz, M. H. (1973), *Spline Analysis*, Englewood Cliffs, NJ: Prentice-Hall.

Stewart, G. W. (1973), *Introduction to Matrix Computations*, New York: Academic.

Wilkinson, J. H. (1963), *Rounding Errors in Algebraic Processes*, Englewood Cliffs, NJ: Prentice-Hall.

Perturbation Methods

James Murdock
Iowa State University, Ames, USA

Introduction ***386***
1 Basic Concepts ***387***
1.1 Perturbation Methods Versus Numerical Methods *387*
1.2 Perturbation Parameters *387*
1.3 Perturbation Series *389*
1.4 Uniformity *390*
2 Nonlinear Oscillations and Dynamical Systems ***392***
2.1 Rest Points and Regular Perturbations *393*
2.2 Simple Nonlinear Oscillators and Lindstedt's Method *394*
2.3 Averaging Method for Single-Frequency Systems *397*
2.4 Multifrequency Systems and Hamiltonian Systems *399*
2.5 Multiple-Scale Method *400*
2.6 Normal Forms *402*
2.7 Perturbation of Stable Manifolds; Melnikov Functions *403*
3 Initial and Boundary Layers ***404***
3.1 Multiple-Scale Method for Initial Layer Problems *404*
3.2 Matching for Initial Layer Problems *405*
3.3 Slow–Fast Systems *407*
3.4 Boundary Layer Problems *407*
3.5 WKB Method *408*
3.6 Fluid Flow *409*
4 Perturbations of Matrices and Spectra ***410***
Glossary ***412***
List of Works Cited ***414***
Further Reading ***415***

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Introduction

Perturbation theory and arises when a situation is given that admits of a mathematical description, and one asks how this description changes when the situation is varied slightly, or ''perturbed.'' This could result either in a continuation of the original situation with only small quantitative changes, or in an abrupt qualitative change in the nature of the situation. Among the possible ''abrupt'' changes are the formation of a transition layer, and the creation of various types of bifurcations; although bifurcation theory is usually treated as a separate subject from perturbation theory, the two areas are closely related. The specific subject matter of the present article will be the following two topics.

1. A system of ordinary or partial differential equations is given, together with initial or boundary conditions. The system contains a small parameter, and is explicitly solvable when the parameter is zero. One asks how to construct approximate solutions (in explicit analytic form) when the parameter is small but nonzero; one asks for error estimates for these approximate solutions, and whether the approximate solutions exhibit the same qualitative behavior as the unknown exact solutions.
2. A matrix or linear transformation is given, depending on a small parameter. The eigenvalues and eigenvectors (or the Jordan normal form) are known when the parameter is zero, and one asks for approximate calculations of the eigenvalues or normal form when the parameter is small but nonzero.

The origins of perturbation theory lie in three classical problems, planetary motion, viscous fluid flow past a wall, and changes in the spectrum as a matrix or linear operator is varied. The present article is structured in the same threefold way: after an initial section presenting basic concepts common to the three areas, we take up in turn dynamical systems (Sec. 2), transition layer problems (Sec. 3), and spectra (Sec. 4). To conclude this introduction, we briefly describe the three classical problems.

Isaac Newton showed that the inverse square law of gravitational force implies that a single planet will move around the sun in an ellipse, satisfying Kepler's laws of planetary motion. The same law of gravity, however, implies that the several planets will exert attractive forces on each other, which will ''perturb'' their orbits. Laplace computed the perturbations, to a certain degree of approximation, and found that they were periodic, so that the solar system was ''stable'' (in Laplace's sense) and would not destroy itself. His techniques were laborious, and can be much simplified today by Hamiltonian mechanics; furthermore, he did not entirely prove that the solar system is stable, since his method, if carried to higher orders, does not necessarily converge. However, a great many of the ideas of modern perturbation theory originated here: variation of parameters, averaging, multiple scales, and the problems that in the twentieth century led to the Kolmogorov–Arnol'd–Moser (KAM) theorem, the Nekhoroshev theorem, and other topics in dynamical systems theory.

In theory, a fluid that is viscous (even to the smallest degree) will adhere to the walls of any container (such as a pipe) in which it is flowing. However, at any reasonable distance from the walls the fluid flows almost as if the wall were not there. In order to resolve this apparent paradox, L. Prandtl introduced the idea of a ''boundary layer,'' a thin layer of fluid against the

wall in which the fluid passes from rest to rapid motion. Here, the ''unperturbed'' problem is the inviscid flow (which does not adhere to the wall), the ''perturbation'' is the small viscosity, and the effect of the perturbation is not a small correction of the motion but a quite drastic correction confined to a small region (the boundary layer). This example leads to the ideas of stretched or scaled coordinates, inner and outer solutions, and matching.

According to quantum mechanics, all observable quantities are the eigenvalues of operators on a Hilbert space. In simple problems, the Hilbert space will be finite dimensional and the operators are representable as matrices; in other cases, they are partial differential operators. In modeling an atom, for instance, the eigenvalues will be related to the frequencies of light (the spectrum) emitted by the atom when electrons jump from one shell to another. These frequencies can be perturbed, for instance, if the atom is placed in a weak magnetic field. Mathematically, the resulting problem is to determine the changes in the ''spectrum'' (the set of eigenvalues) of a matrix or other operator when the operator is slightly perturbed. One striking difference between this problem and the first two is that quantum mechanics is entirely a linear theory, whereas in both the dynamical systems problems and the boundary layer problems, nonlinearities can play a crucial role.

1 Basic Concepts

1.1 Perturbation Methods Versus Numerical Methods

A system of differential equations that is not explicitly solvable calls for an approximate solution method of some type. The most useful approximate methods are numerical methods (implemented on a digital computer) and perturbation methods. Perturbation methods are usable only if the system is ''close'' to an explicitly solvable system, in the sense that the system would become solvable if certain small changes were made, such as deleting small terms or averaging some term over a rapidly rotating angle. In such a case, perturbation theory takes the solution of the simplified problem as a ''zeroth order approximation'' that can be successively improved, giving higher-order approximations having explicit formulas. Numerical methods operate without the restriction that the problem be nearly solvable, but give a solution in the form of numerical tables or graphs. The advantage of a formula is that one can see by inspection the manner in which each variable affects the solution. However, since both types of solution are approximate, it is often best to verify a perturbation solution by comparing it with numerical ones (or directly with experimental data), especially when a mathematically valid error estimate for the perturbation solution is missing.

1.2 Perturbation Parameters

Mathematical models of physical phenomena typically contain several variables, which are divided into ''coordinates'' (of space or time) and ''parameters.'' A spring/mass system with a cubic nonlinearity, for instance, will contain position and time coordinates as well as parameters for the mass and for the coefficients of the linear and cubic terms in the restoring force. The first step in preparing such a system for perturbation analysis is to nondimensionalize these variables. The

second step is to look for solvable special cases that can serve as the starting point for approximating the solution of nearby cases. Most often, these solvable cases will be obtained by setting some of the parameters equal to zero. For instance, the forced and damped nonlinear oscillator given by

$$\ddot{y} + C\dot{y} + k^2 y + Ay^3 = B\cos\omega t \qquad (1)$$

becomes easily solvable if $A = B = C = 0$; it is still solvable if only $A = 0$, but not quite as simply. It is therefore plausible to look for approximate solutions by perturbation theory if A, B, and C are small, and it may also be possible if only A is small.

Suppose that we choose to investigate the case when A, B, and C are small. Ideally, we could treat these as three small independent parameters, and considerable work is now being devoted to the investigation of such "multi-parameter" perturbation methods, especially in the context of bifurcation theory (see Golubitsky and Schaeffer, 1985). However, most classical perturbation methods are developed only for single-parameter problems. Therefore, it is necessary to *make a choice* as to how to reduce Eq. (1) to a single-parameter problem. The simplest way is to write $A = a\varepsilon$, $B = b\varepsilon$, and $C = c\varepsilon$, obtaining

$$\ddot{y} + \varepsilon c\dot{y} + k^2 y + \varepsilon a y^3 = \varepsilon b\cos\omega t. \qquad (2)$$

We appear to have added a parameter instead of reducing the number, but now a, b, and c are regarded as constants, whereas ε, the *perturbation parameter*, is taken as a small, but variable, quantity; typically, we expect the perturbation solution to have the form of a power series in ε. We have in effect chosen to investigate a particular *path* leading from the origin in the space of variables A, B, C. It is at once clear that other paths are possible, for instance

$$\ddot{y} + \varepsilon^2 c\dot{y} + k^2 y + \varepsilon a y^3 = \varepsilon b\cos\omega t. \qquad (3)$$

One might choose Eq. (3) over Eq. (2) if the goal is to investigate systems in which the damping is *extremely* small, small even when compared to the cubic term and the forcing. But it is not clear, without experience, what the best formulation will be for a given problem. As an example of the role of experience, one might expect (knowing something about resonance in the linear case) that the results of studying Eq. (2) will be different if ω is close to k than if it is far away. But how do you express mathematically the idea that "ω is close to k"? Recalling that the only parameter available to express "smallness" is ε, the best answer turns out to be

$$\omega^2 = k^2 + \varepsilon d \qquad (4)$$

where d is another constant. Substituting Eq. (4) into Eq. (2) leads to the "correct" formulation of the near-resonance problem. One can see that the setting up of perturbation problems is sometimes an art rather than a science. In the present article we will for the most part assume that a perturbation problem has been chosen. Mathematical analysis of the problem may then suggest the use of *stretched* or otherwise *rescaled* variables, which in fact amount to a modification of the initial perturbation problem.

In recent years, a further consideration has come to the fore regarding the choice of parameters in a mathematical model. Physical considerations may have led to a model such as Eq. (1), and yet we know that this model is not exactly correct; there may, for instance, be very small nonlinearities other than the cubic term in the restoring force, or nonlinearities in the damping, or additional harmonics in the forcing. How

many such effects should be included in the model? In the past, it was simply a matter of trial and error. But in certain cases, there now exists a mathematical theory that is able to determine just which additional small terms might make a *qualitative* (rather than just a tiny quantitative) difference in the solution. In these cases, it is sometimes best to *add all such significant small terms to the equation before attempting the solution, even if there is no evident physical reason for doing so.* The process of adding these additional terms is called finding the *universal unfolding* of the system. The advantage is that the universal unfolding will account for all possible qualitative behaviors that may be observed as a result of unknown perturbations. For instance, in the past many bifurcations were not observed to occur exactly as predicted. They were called *imperfect bifurcations,* and each situation required that the specific perturbation responsible for the imperfection be discovered and incorporated into the model. Now it is often possible to determine all possible imperfections in advance by examining the universal unfolding of the original model. See Golubitsky and Schaeffer (1985).

In addition to the types of problems already discussed, there exist problems that do not contain a perturbation parameter but nonetheless allow treatment by perturbation methods. For instance, a system of differential equations may have a particular solution (often an equilibrium solution or a periodic solution) that can be computed exactly, and one may wish to study the solutions lying in a neighborhood of this one. One way to treat such problems is called *coordinate perturbations*; the coordinates themselves (or more precisely, the differences between the coordinates of the known and unknown solutions) are treated as small quantities, in place of a perturbation parameter. Another approach is to introduce a parameter ε as a scale factor multiplying these coordinate differences. Both ideas will be illustrated in the discussion of *normal forms* in Sec. 2.6 below.

1.3 Perturbation Series

Let us suppose that a perturbation problem has been posed, and let the exact solution (which we wish to approximate) be denoted $u(x, \varepsilon)$. Here ε is the (scalar) perturbation parameter, x is a vector consisting of all other variables in the problem including coordinates and other parameters, and u is the quantity being solved for. (In the case of Eq. (2), $u = y$ and $x = (t, a, b, c, \omega)$, or if Eq. (4) is used then $x = (t, a, b, c, d)$.) The simplest form in which to seek an approximation is that of a (truncated) *power series* in ε:

$$u(x, \varepsilon) \cong u_0(x) + \varepsilon u_1(x) + \cdots + \varepsilon^k u_k(x). \tag{5}$$

There are times when this is insufficient and we require a *Poincaré series*

$$u(x, \varepsilon) \cong \delta_0(\varepsilon)u_0(x) + \delta_1(\varepsilon)u_1(x) + \cdots + \delta_k(\varepsilon)u_k(x) \tag{6}$$

where each δ_i is a monotone function of ε defined for $\varepsilon > 0$ satisfying

$$\lim_{\varepsilon \to 0} \frac{\delta_{i+1}(\varepsilon)}{\delta_i(\varepsilon)} = 0; \tag{7}$$

such functions δ_i are called *gauges.* (Of course, a Poincaré series reduces to a power series if $\delta_i(\varepsilon) = \varepsilon^i$.) Finally, there are times when not even a Poincaré series is sufficient and a *generalized series* is needed:

$$u(x, \varepsilon) \cong \delta_0(\varepsilon)u_0(x, \varepsilon) + \delta_1(\varepsilon)u_1(x, \varepsilon) + \cdots + \delta_k(\varepsilon)u_k(x, \varepsilon). \tag{8}$$

With such a series, it might appear that we could delete the gauges, or rather assimilate them into the u_i since these are now allowed to depend upon ε; but the intention is that the dependence of u_i on ε should not affect its order of magnitude. An example is the following two-term generalized series where the vector x consists only of the time t:

$$u(t, \varepsilon) \cong \sin(1+\varepsilon)t + \varepsilon \cos(1+\varepsilon)t. \quad (9)$$

Here, $u_0(t, \varepsilon) = \sin(1+\varepsilon)t$ and $u_1(t, \varepsilon) = \cos(1+\varepsilon)t$; the dependence of these coefficients upon ε modifies their period but not their amplitude, and the second term still has the order of magnitude of its gauge ε.

Notice that we have written only truncated series in the previous paragraph. While most perturbation methods allow in principle for the computation of infinite series, these series very often do not converge, and in practice it is impossible to calculate more than a few terms. The type of accuracy that we hope for in a perturbation solution is not convergence (improvement in accuracy as the number of terms increases), but rather *asymptotic validity*, which means improvement in accuracy as ε approaches zero. To explain this concept, let us consider a generalized series

$$u(\varepsilon) \cong \delta_0(\varepsilon)u_0(\varepsilon) + \delta_1(\varepsilon)u_1(\varepsilon) + \cdots + \delta_k(\varepsilon)u_k(\varepsilon) \quad (10)$$

that contains no variables other than ε. We will say that this series is an *asymptotic approximation* to $u(\varepsilon)$ if

$$u(\varepsilon) = \delta_0(\varepsilon)u_0(\varepsilon) + \delta_1(\varepsilon)u_1(\varepsilon) + \cdots + \delta_k(\varepsilon)u_k(\varepsilon) + R(\varepsilon) \quad (11)$$

where the remainder or error $R(\varepsilon)$ satisfies a bound of the form

$$|R(\varepsilon)| \leq c\delta_{k+1}(\varepsilon) \quad (12)$$

for some constant $c > 0$ and some gauge δ_{k+1} that approaches zero more rapidly than δ_k as $\varepsilon \to 0$. (If the series is vector-valued, then $|R(\varepsilon)|$ denotes a vector norm.) Equation (12) is abbreviated with the "big-oh" notation

$$R(\varepsilon) = \mathcal{O}(\delta_{k+1}(\varepsilon)). \quad (13)$$

The series (10) is called *asymptotically valid*, or *an asymptotic series*, if it is an asymptotic approximation (in the above sense) and in addition, every truncation of Eq. (10) is also an asymptotic approximation, with the error being "big-oh" of the first omitted gauge. The case in which $u(x, \varepsilon)$ depends upon variables x in addition to ε will be discussed in the following section on "Uniformity."

As a technical matter, any bound such as Eq. (12) is not intended to hold for all ε, but only for ε in some interval $0 \leq \varepsilon \leq \varepsilon_0$. A perturbation solution is never expected to be valid for large values of the perturbation parameter, and the meaning of "large" is relative. In this article, we will not continue to mention ε_0, but it is always lurking in the background.

To conclude this discussion of types of series in perturbation theory, it should be mentioned that Fourier series arise frequently in dealing with oscillatory problems, but a Fourier series is never a perturbation series. Rather, if a perturbation series such as $u(t, \varepsilon) \cong u_0(t) + \varepsilon u_1(t)$ depends periodically on time t, then the coefficients $u_k(t)$ may be expressed as Fourier series in t.

1.4 Uniformity

In the previous section, we have defined asymptotic validity of a perturbation series

for a function $u(\varepsilon)$ depending only on ε. This is adequate for a problem such as finding a root of a polynomial (supposing that the polynomial contains a perturbation parameter ε) because the root is a single number. But for most perturbation problems (such as differential equations), the solution is a function of space and/or time coordinates, and possibly various parameters, in addition to ε. For such problems, the previous definition of asymptotic validity is insufficient.

Let us return to the generalized series (8), and denote the error by $R(x, \varepsilon)$. Now we may require that *for each fixed* x this error is of order $\delta_{k+1}(\varepsilon)$:

$$|R(x, \varepsilon)| \leq c(x)\delta_{k+1}(\varepsilon). \tag{14}$$

Notice that the "constant" c here may change when we move to a new point x. In this case, we say that the error is *pointwise* of order δ_{k+1}.

Alternatively, we can choose a domain D of the variable x (remember that x may have several components, so D may be a subset of a vector space) and require that the error be *uniformly* of order δ_{k+1} for all x in D:

$$|R(x, \varepsilon)| \leq c\delta_{k+1}(\varepsilon). \tag{15}$$

Here, the constant c is truly constant. In this case, we say $R(x, \varepsilon) = \mathcal{O}(\delta_{k+1}(\varepsilon))$ *uniformly in* x for x in D. If every truncation of a perturbation series is uniformly of the order of the first omitted gauge, we say that the series is *uniformly valid*. (In the last sentence, we neglected to say "in D", but it is important to remember that such an expression has no meaning unless the domain D is understood.) Obviously, uniform asymptotic validity is stronger than pointwise validity, and it is safe to say that every method used in perturbation theory has been introduced in order to gain uniform validity for a problem for which previous methods only gave pointwise validity.

Now the definition of uniform validity calls for an estimate of the error of an approximation, and such an estimate is a difficult thing to come by. It would be convenient if there were an easier test for uniformity. In actuality there is not. However, it is possible to obtain a simple *necessary* (but not sufficient) condition for a series to be uniformly valid. Namely, it is not difficult to show that if the series (8) is uniformly valid in a domain D, then each coefficient $u_k(x)$ with $k \geq 1$ is bounded on D, that is, there exist constants c_k such that

$$|u_k(x, \varepsilon)| \leq c_k \tag{16}$$

for x in D. If this is true, we say that the series (8) is *uniformly ordered*. We have already encountered the concept of a uniformly ordered series when discussing Eq. (9). A uniformly ordered series is one in which each term after the first is of no greater order than is indicated by its gauge. It is easy to inspect a perturbation series, once it has been constructed, and determine whether it is uniformly ordered. If not, the series is called *disordered*, and it cannot be uniformly valid. On the other hand, if it is uniformly ordered, it does not follow that it is uniformly valid, since one has done nothing toward estimating the error. Almost all textbooks are misleading on this point, since they almost invariably claim to be showing the uniform validity of a series when in fact they are only showing that it is uniformly ordered. However, if a perturbation series is constructed on the basis of good intuitive insight into a problem, and if it is uniformly ordered, then it frequently turns out to be uniformly valid as well. (A very elementary example in which this is *not* the case will be given in Sec. 2.1.)

With regard to uniform ordering, Poincaré series occupy a special place. Recall that a series is a Poincaré series if its coefficients do not depend on ε; see Eq. (6). In this case, if the domain D is compact and the coefficients $u_k(x)$ are continuous, then the coefficients are automatically bounded and the series is uniformly ordered (but still not automatically uniformly valid). However, even a Poincaré series may fail to be uniformly ordered if D is not compact or if D is allowed to depend on ε. We have not considered this latter possibility until now, but, in fact, in many problems one must consider domains $D(\varepsilon)$ that depend upon ε. An important example is a boundary layer, a thin domain near the boundary of some larger region, whose thickness depends on ε. For such domains, the definitions (15) and (16) of uniform validity and uniform ordering are the same, except that they are to hold for all x in $D(\varepsilon)$; that is, for each value of ε, Eqs. (15) or (16) hold for a different range of x.

It is now possible to explain one of the principal divisions in the subject of perturbation theory, the division of perturbation problems into *regular* and *singular*. A perturbation problem is regular if there exists a Poincaré series that is uniformly valid on the intended domain; it is singular if it is necessary to use a generalized series in order to obtain a uniformly valid solution on the intended domain. This is the only correct definition. Many textbooks give partial definitions such as "a perturbation problem for a differential equation is singular if ε multiplies the highest derivative." Such a definition, which refers only to the differential equation without stating an intended domain, cannot be correct. The presence of an ε multiplying the highest derivative does, of course, affect the domain on which a problem can be regular; we will see below that a problem such as $\ddot{u} + u + \varepsilon u^3 = 0$ is regular on any fixed interval $[0, T]$ but singular on an expanding interval $[0, 1/\varepsilon]$, while a problem such as $\varepsilon\ddot{u} + (t^2 + 1)\dot{u} + u = 0$ is regular only on a shrinking interval $[0, \varepsilon]$ and singular on a fixed interval.

2 Nonlinear Oscillations and Dynamical Systems

In this section, we discuss the major perturbation methods used in the study of nonlinear ordinary differential equations (dynamical systems). Typical problems include the location of rest (or equilibrium) points and periodic or quasi-periodic solutions; the approximation of solutions close to these; the solution of initial value problems for systems that become solvable (usually either linear or integrable Hamiltonian) when a small parameter vanishes; and the splitting of a homoclinic orbit under perturbation. In all of these problems, there is an interplay between qualitative and quantitative behavior. Advance knowledge of the qualitative behavior may assist the choice of a successful perturbation method. Alternatively, if the qualitative behavior is unknown, perturbation methods may be used in an exploratory fashion, as long as it is kept in mind that very frequently, especially in nonlinear problems, a perturbation solution may appear correct but may predict qualitative features (such as periodicity, stability, or presence of chaos) erroneously. Many such faulty results appear in the literature, and more are published every year. As a general rule, qualitative results obtained from perturbation solutions (or any other approximate solutions) should be taken as conjectural,

until supported by some combination of numerical or experimental evidence and rigorous mathematical proof.

2.1 Rest Points and Regular Perturbations

Given a system of differential equations of the form

$$\dot{x} = f(x, \varepsilon) = f_0(x) + \varepsilon f_1(x) + \mathcal{O}(\varepsilon^2), \tag{17}$$

with $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$, the *rest points* are the solutions of the algebraic system $f(x, \varepsilon) = 0$. If a rest point a_0 is known when $\varepsilon = 0$, it may continue to exist as a function $a(\varepsilon)$ with

$$f(a(\varepsilon), \varepsilon) = 0 \tag{18}$$

for ε near zero, or it may bifurcate into two or more rest points, or disappear altogether; the results may differ for $\varepsilon > 0$ and $\varepsilon < 0$. A crucial role is played by the matrix $A_0 = f_x(x_0, 0)$ of partial derivatives of f at the unperturbed rest point and its eigenvalues. If A_0 is invertible (zero is not an eigenvalue), then a unique continuation $a(\varepsilon)$ exists and is computable as a series

$$a(\varepsilon) = a_0 + \varepsilon a_1 + \mathcal{O}(\varepsilon^2). \tag{19}$$

This is the simplest example of a perturbation series. Putting Eq. (19) into Eq. (18) and expanding gives $f_0(a_0) + \varepsilon(A_0 a_1 + f_1(a_0)) + \cdots = 0$, or (since $f_0(a_0) = 0$)

$$a_1 = -A_0^{-1} f_1(a_0). \tag{20}$$

The solution may be continued to higher order, and the matrix function

$$A(\varepsilon) = f_x(a(\varepsilon), \varepsilon) \tag{21}$$

may also be studied, since (in most cases) it determines the stability of the rest point. If all of the eigenvalues of A_0 are off the imaginary axis (A_0 is hyperbolic), the same will be true for $A(\varepsilon)$ for small ε, and the stability type (dimensions of the stable and unstable manifolds) of the rest point will not change. When this is not the case, the methods described in Sec. 4 determine the spectrum of $A(\varepsilon)$, and thus usually suffice to decide how the stability changes.

When A_0 is not invertible, bifurcation (change in the number of rest points) usually occurs. Even when A_0 has only one zero eigenvalue, various possibilities (such as saddle-node and pitchfork bifurcations) exist, and it is not possible to give details here. A reference treating the subject from the standpoint of perturbation theory is Iooss and Joseph (1980); a quite different modern treatment is Golubitsky and Schaeffer (1985).

Now suppose that a solution

$$x(t, \varepsilon) = x_0(t) + \varepsilon x_1(t) + \mathcal{O}(\varepsilon^2) \tag{22}$$

of the system in Eq. (17) is to be found, with initial condition $x(0, \varepsilon) = a(\varepsilon)$ (no longer a rest point) given by Eq. (19). Substituting Eq. (22) into Eq. (17), expanding, and equating terms of the same degree in ε yields

$$\begin{aligned} \dot{x}_0 &= f_0(x_0), \\ \dot{x}_1 &= f_{0x}(x_0(t))x_1 + f_1(x_0(t)). \end{aligned} \tag{23}$$

If the first equation of (23), which is the same as the unperturbed equation ((17) with $\varepsilon = 0$), can be solved with initial condition $x_0(0) = a_0$, its solution $x_0(t)$ can be placed in the second equation of (23), which then becomes an inhomogeneous *linear* equation for x_1; it is to be solved with initial condition $x_1(0) = a_1$. Equations of

this type are not necessarily easy to solve, but are certainly easier than the nonlinear Eq. (17). If this is successful, the procedure may be continued to higher order.

This is usually called the *regular perturbation method* or the method of *straightforward expansion*. According to our earlier definition, a perturbation method is regular if it provides a Poincaré series that is uniformly valid on the intended domain. Here Eq. (22) is a Poincaré series, and it can be shown to be uniformly valid on any *finite* interval $[0, T]$; that is, the error bound is of the order of the first omitted term, and once T is chosen, the coefficient in the error bound is fixed. So the term "regular" is justified if this is the intended domain. In many problems (below) one seeks a solution valid on an "expanding" interval such as $[0, 1/\varepsilon]$; the straightforward expansion is usually not valid for this purpose.

There are situations in which straightforward expansion is valid for much longer than finite intervals. For instance, if a solution is approaching a sink (a rest point with all eigenvalues in the left half plane), the straightforward expansion is valid for all future time (t in $[0, \infty)$). More generally, if the first equation of (23) has a solution that connects two hyperbolic rest points, then a straightforward expansion (to any order) beginning with that solution will be *shadowed* by an exact solution of Eq. (17) connecting two hyperbolic rest points of that system; that is, the approximate and exact solutions will remain close (to the order of the first omitted term) for all time, both past and future, although the two solutions may not have any point in common. (In particular, the approximate and shadowing solutions will not satisfy the same initial conditions.) For a precise statement, see Murdock (1996).

2.2 Simple Nonlinear Oscillators and Lindstedt's Method

The "hard" nonlinear spring or unforced Duffing equation is given by

$$\ddot{u} + u + \varepsilon u^3 = 0 \tag{24}$$

for $\varepsilon > 0$. It can be expressed as a first-order system in the $(u, \dot{u})$ phase plane in the form $\dot{u} = v$, $\dot{v} = -u - \varepsilon u^3$. In the phase plane, the orbits are closed curves surrounding the rest point at the origin; this may be seen from the conservation of energy. (The "soft" spring with $\varepsilon < 0$ behaves differently.) Since the solutions of Eq. (24) with any initial conditions $u(0) = a$, $\dot{u}(0) = b$ are smooth functions of ε, they may be expanded as Taylor series having the form (if we retain two terms)

$$u(t, \varepsilon) \cong u_0(t) + \varepsilon u_1(t). \tag{25}$$

The coefficients may be determined by substituting Eq. (25) into Eq. (24), expanding the u^3 term, dropping powers of ε higher than the first, and setting each order in ε separately equal to zero. This gives two *linear* equations that can be solved (recursively) for u_0 and u_1. The result, for $b = 0$ (and it is enough to consider this case because every solution is at rest momentarily when its amplitude is at its maximum), is

$$u(t, \varepsilon) \cong a \cos t - \varepsilon \frac{a^3}{32} \times (\cos t + 12t \sin t - \cos 3t). \tag{26}$$

Upon examining Eq. (26) for uniform ordering, we discover that all functions of t appearing there are bounded for all t except for $12t \sin t$, which becomes unbounded as t approaches infinity. This

is an example of a so-called "secular" term, one which grows over the "ages" (*saecula* in Latin). We conclude from this that Eq. (26) is uniformly ordered for t in any finite interval $D = [0, T]$ but not for $D = [0, \infty)$. According to the general principles discussed in Sec. 1.4, this shows that Eq. (26) is not uniformly valid on $[0, \infty)$, and it *leads us to suspect, but does not prove* that Eq. (16) is uniformly valid on $[0, T]$. In the present case, this conjecture is correct. If the intended domain D for the solution of Eq. (24) is a finite interval, then we have obtained a uniform approximation in the form of a Poincaré series, and the problem is a regular one. If a solution valid for a longer time is desired, the problem will prove to be singular.

In an effort to extend the validity of the solution, we recall that the actual solutions of Eq. (24) are periodic, whereas Eq. (26) is not. The problem is that the period of the exact solution depends upon ε, and there is no way that a Poincaré series can have such a period since the coefficients are not allowed to depend on ε. To remedy this, we seek to approximate the (unknown) frequency of the solution in the form

$$\nu(\varepsilon) \cong \tilde{\nu}(\varepsilon) = \nu_0 + \varepsilon\nu_1 + \varepsilon^2\nu_2 \qquad (27)$$

with the solution itself being represented as

$$u(t, \varepsilon) \cong u_0(\tilde{\nu}(\varepsilon)t) + \varepsilon u_1(\tilde{\nu}(\varepsilon)t), \qquad (28)$$

which is now a generalized series. Notice that we have carried Eq. (27) to one more order than Eq. (28). Now we substitute Eqs. (27) and (28) into Eq. (24) and attempt to determine $\nu_0, u_0, \nu_1, u_1, \nu_2$ recursively, in that order, in such a way that each u_i is periodic. The mechanics of doing this will be explained in the next paragraph; the result is

$$u(t, \varepsilon) \cong a\cos t^+ + \varepsilon\left(-\frac{1}{32}a^3\cos t^+ + \frac{1}{32}a^3\cos 3t^+\right) \qquad (29)$$

where

$$t^+ = \tilde{\nu}(\varepsilon)t = \left(1 + \varepsilon\frac{3}{8}a^2 - \varepsilon^2\frac{21}{256}a^4\right)t. \qquad (30)$$

Examining Eq. (29) we see that it is uniformly ordered for all time, since the coefficients are bounded (there are no secular terms). One might therefore conjecture that the solution is uniformly valid for all time, *but this would be incorrect!* (This example is an excellent warning as to the need for proofs of validity in perturbation theory.) The difficulty is that t^+ uses the approximate frequency $\tilde{\nu}(\varepsilon)$ in place of the exact frequency $\nu(\varepsilon)$; there is no escape from this, as the exact frequency remains unknown. Therefore, Eq. (29) gradually gets out of phase with the exact solution. The reason for taking Eq. (27) to one higher order than Eq. (28) is to minimize this effect. It can be shown that Eq. (29) is uniformly valid on the *expanding interval* $D(\varepsilon) = [0, 1/\varepsilon]$. (This is our first example of a domain that depends on ε, as discussed in Sec. 1.4.) If the intended domain is such an expanding interval, then Eq. (29) provides a uniformly valid generalized series, and the problem is seen to be singular. (If the intended domain is all t, the problem is simply impossible to approximate asymptotically.)

In order to complete the example in the last paragraph, we must indicate how to obtain Eqs. (29) and (30). The easiest way is to substitute $\tau = \nu(\varepsilon)t$ into Eq. (24) to obtain

$$\nu(\varepsilon)^2\frac{d^2u}{d\tau^2} + u + \varepsilon u^3 = 0. \qquad (31)$$

Then substitute Eqs. (27) and (28) into Eq. (31), expand, and set the coefficient of each power of ε equal to zero as usual. It is easy to find that $\nu_0 = 1$ and $u_0 = a\cos\tau$ (which in the end becomes $a\cos t^+$ because we do not know the exact frequency ν). The crucial step arises when examining the equation for u_1, which is (writing $' = d/d\tau$)

$$\begin{aligned} u_1'' + u_1 &= -a^3\cos^3\tau + 2\nu_1 a\cos\tau \\ &= \left(-\frac{3}{4}a^3 + 2\nu_1 a\right)\cos\tau \\ &\quad - \frac{1}{4}a^3\cos 3\tau. \end{aligned} \tag{32}$$

From the Fourier series expansion (the second line of Eq. (32)), we see that the term in $\cos\tau$ will be resonant with the free frequency, and hence produce unbounded (secular) terms in u_1, unless the coefficient of $\cos\tau$ vanishes. In this way, we conclude that

$$\nu_1 = \frac{3}{8}a^2 \tag{33}$$

and, after deleting the $\cos\tau$ term from Eq. (32), we solve it for u_1. This procedure is repeated at each subsequent stage.

The previous example is typical of *unforced conservative* oscillators, where every solution (at least in a certain region) is periodic. There are two additional classes of oscillators that must be mentioned, although we cannot give them as much space as they deserve: self-excited oscillators and forced oscillators.

The standard example of a self-excited oscillator is the *Van der Pol equation*

$$\ddot{u} + \varepsilon(u^2 - 1)\dot{u} + u = 0. \tag{34}$$

Instead of a region in the phase plane filled with periodic solutions, this equation has a single periodic solution (limit cycle) for $\varepsilon > 0$. The Lindstedt method, described above, can be used to approximate the periodic solution, but must be modified slightly: the initial condition can no longer be assigned arbitrarily, because to do so will in general yield a nonperiodic solution for which the Lindstedt method fails. (These solutions can be found by averaging or multiple scales; see Secs. 2.3 and 2.5.) Suppose that the limit cycle crosses the positive x axis (in the phase plane) at $(a(\varepsilon), 0)$ and has frequency $\nu(\varepsilon)$. Then the solution is sought in the form of Eqs. (27) and (28) together with an additional expansion $a(\varepsilon) = a_0 + \varepsilon a_1 + \cdots$; the coefficients u_i, ν_i, and a_i are determined recursively, choosing ν_i and a_i so that no secular terms arise in u_{i+1}. This example shows the effect of the dynamics of a system on the correct formulation of a perturbation problem.

The general nearly linear, periodically forced oscillator can be written as

$$\ddot{u} + u = \varepsilon f(u, \dot{u}, \omega t), \tag{35}$$

where $f(u, v, \theta)$ is 2π-periodic in θ; thus, the period of the forcing is $2\pi/\omega$. The dynamics of Eq. (35) can be complicated, and we will limit ourselves to one type of periodic solution, the *harmonic* oscillations, which are *entrained* by the forcing so that they have the same frequency ω; these harmonic solutions occur for ε small and ω close to 1 (the frequency of the solutions when $\varepsilon = 0$). Since the problem contains two parameters (ε and ω) and we are limited to one-parameter methods, it is necessary to express the statement "ω is close to 1" in terms of the perturbation parameter ε. (A study using two independent parameters would uncover the phenomenon of "resonance horns" or "resonance tongues." See Murdock,

(1999), Sec. 4.5.) It turns out that an efficient way to do so is to write

$$\omega^2 = 1 + \varepsilon\beta, \tag{36}$$

where β is a new parameter that is considered fixed (not small). With Eq. (36), the Eq. (35) can have one or more isolated periodic solutions with unknown initial conditions $(a(\varepsilon), b(\varepsilon))$. (We can no longer assume $b = 0$.) On the other hand, the frequency of the solution this time is not unknown but equals ω. So the Lindstedt method undergoes another modification dictated by the dynamics: u, a, and b are expanded in ε and solved recursively, choosing the coefficients of a and b to eliminate secular terms from u. In contrast with the previous cases, there is no accumulating phase error since the frequency is known, and the perturbation approximations are uniformly valid for all time.

2.3 Averaging Method for Single-Frequency Systems

All of the systems discussed in the previous section, and a great many others, can be expressed in *periodic standard form*,

$$\dot{x} = \varepsilon f(x, t, \varepsilon), \tag{37}$$

where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and where f is 2π-periodic in t. The solutions of Eq. (37) may be approximated by the *method of averaging*, which not only locates the periodic solutions (and proves their existence) but also determines their stability or instability and approximates the transient (nonperiodic) solutions. The method of averaging has been rediscovered many times and exists (with slight variations) under a variety of names, including: method of Van der Pol; method of Krylov–Bogoliubov–Mitropolski (KBM method); method of slowly varying amplitude and phase; stroboscopic method; Struble's method; Von Zeipel's method (in the Hamiltonian case); method of Lie series or Lie transforms. Some of the differences in these "methods" pertain to how the original system is put into periodic standard form, and others to details about how the near-identity transformations (described below) are handled.

To illustrate how a system may be put into periodic standard form, consider the Van der Pol equation (34), or, written as a system,

$$\begin{aligned} \dot{u} &= v, \\ \dot{v} &= -u + \varepsilon(1 - u^2)v. \end{aligned} \tag{38}$$

Rotating polar coordinates (r, φ) may be introduced by $u = r\cos(\varphi - t)$, $v = r\sin(\varphi - t)$, giving

$$\begin{aligned} \dot{r} &= \varepsilon(1 - r^2\cos^2(\varphi - t))r\sin^2(\varphi - t), \\ \dot{\varphi} &= \varepsilon(1 - r^2\cos^2(\varphi - t)) \\ &\quad \times \sin(\varphi - t)\cos(\varphi - t), \end{aligned} \tag{39}$$

which is in periodic standard form with $x = (r, \varphi)$. The same result may be achieved by seeking a solution for Eq. (34) by variation of parameters, in the form $u = r\cos(\varphi - t)$ where r and φ are variables, and imposing the requirement that $\dot{u} = r\sin(\varphi - t)$; the motivation for these choices is that with r and φ constant, these solve Eq. (34) for $\varepsilon = 0$. The transformation to periodic standard form is merely a change of coordinates, not an assumption about the nature of the solutions.

In its crudest form, the method of averaging simply consists in replacing Eq. (37) by

$$\dot{z} = \varepsilon\overline{f}(z), \tag{40}$$

where

$$\bar{f}(z) = \frac{1}{2\pi}\int_0^{2\pi} f(z, t, 0)\, dt. \qquad (41)$$

System (40) is easier to solve than Eq. (37) because it is autonomous. The form of Eq. (40) can be motivated by the fact that for small ε, x in Eq. (37) is slowly varying and therefore nearly constant over one period; therefore, to a first approximation we might hold x constant while integrating over one period in Eq. (37) to find the "average" influence due to f. But this sort of motivation gives no idea how to estimate the error or to extend the method to higher-order approximations. A much better procedure is to return to Eq. (37) and perform a near-identity change of variables of the form

$$x = y + \varepsilon u_1(y, t, \varepsilon), \qquad (42)$$

where u is a periodic function of t, which is to be determined so that the transformed equation has the form

$$\dot{y} = \varepsilon g_1(y) + \varepsilon^2 \hat{g}(y, t, \varepsilon) \qquad (43)$$

where g_1 is independent of t. It turns out that such a transformation is possible only if we take $g_1 = \bar{f}$; by doing so, Eq. (40) can be obtained from Eq. (43) simply by deleting the ε^2 term. When it is formulated in this way, the entire method of average is seen to consist of nothing but coordinate changes (first into periodic standard form, then into form (43)), followed by truncation; it is only the truncation that introduces error, and this error can be estimated using Gronwall's inequality. It is also clear how to proceed to higher orders; simply replace Eq. (42) by

$$x = y + \varepsilon u_1(y, t, \varepsilon) + \cdots + \varepsilon^k u_k(y, t, \varepsilon) \qquad (44)$$

and Eq. (43) by

$$\dot{y} = \varepsilon g_1(y) + \cdots + \varepsilon^k g_k(y) + \varepsilon^{k+1}\hat{g}(y, t, \varepsilon); \qquad (45)$$

the averaged equations are obtained by deleting $\hat{g}$. It is of course necessary to determine the u_i and g_i recursively in such a way that the u_i are periodic and the g_i are independent of t; this is where the technical details of various versions of averaging come into play.

The final conclusion of the method of averaging is that if Eq. (45) is truncated to order ε^k and solved, and the solutions are put back into the transformation (44), the resulting approximate solutions of Eq. (37) will differ from the exact solutions (with the same initial condition) by $\mathcal{O}(\varepsilon^k)$ during a time interval of length $\mathcal{O}(1/\varepsilon)$. Since the error is of the same order as the last term in Eq. (44), this term can be omitted when the solution of Eq. (45) is inserted. A recent variation in the method omits the last term of Eq. (44) altogether, with a change in the form of Eq. (45); the arguments are harder, and the result is the same in the case of smooth ordinary differential equations, but it is much more widely applicable, for instance, to partial differential equations. See Ellison et al. (1990) and the work of A. Ben Lemlih cited there. For additional information about the use of averaging in partial differential equations, see Verhulst (1999).

As for regular perturbation theory, under special conditions it is possible to obtain results on half-infinite or infinite intervals of time. See Sanders and Verhulst (1985) for the Sanchez-Palencia theorem (for solutions approaching a sink) and Murdock (1996) for shadowing.

2.4 Multifrequency Systems and Hamiltonian Systems

Oscillatory problems that cannot be put into periodic standard form can often be put into the following *angular standard form*:

$$\dot{r} = \varepsilon f(r, \theta, \varepsilon),$$
$$\dot{\theta} = \Omega(r) + \varepsilon g(r, \theta, \varepsilon), \qquad (46)$$

where $r = (r_1, \ldots, r_n)$ is a vector of amplitudes and $\theta = (\theta_1, \ldots, \theta_m)$ a vector of angles (so that f and g are periodic in each θ_i with period 2π). This form includes the periodic standard form, by taking $m = 1$ and $\dot{\theta} = 1$. The "naive" method of averaging for Eq. (46) would be to replace f and g by their averages over θ, for instance

$$\bar{f}(r) = \frac{1}{(2\pi)^m} \int_0^{2\pi} \cdots \int_0^{2\pi} f(r, \theta, 0)\, d\theta_1 \cdots d\theta_m. \qquad (47)$$

To justify this process, and to extend the method to higher order, one tries (as in the method of averaging) to make a near-identity change of variables $(r, \theta) \to (\rho, \varphi)$ that will render the system independent of θ up through a given order k in ε. However, one encounters at once the famous difficulty of "small divisors," which make the existence of such a transformation doubtful. If f is expanded in a convergent multiple Fourier series

$$f(r, \theta, 0) = \sum_{\nu} a_\nu(r) e^{i(\nu_1\theta_1 + \cdots + \nu_m\theta_m)}, \qquad (48)$$

then the transformation to averaged form necessarily involves the series

$$\sum_{\nu \neq 0} \frac{a_\nu(r)}{i(\nu_1\Omega_1(r) + \cdots + \nu_m\Omega_m(r))} \times e^{i(\nu_1\theta_1 + \cdots + \nu_m\theta_m)}, \qquad (49)$$

which may not converge because the denominators $i(\nu_1\theta_1 + \cdots + \nu_m\theta_m)$ may be small (or even zero), causing the coefficients to become large. It is of no use at this point to say that "perturbation theory is not concerned with the convergence of series", since the series in question are not being used for approximation, but to prove the existence of a transformation that is needed in order to justify a method.

Some preliminary progress can be made by considering the case in which the series (48), and hence Eq. (49), are finite. In this case, convergence difficulties cannot arise, but there is still the difficulty that for some r one or more of the denominators of Eq. (49) may become zero. Since $\Omega_i(r)$ are the frequencies of the free oscillations ($\varepsilon = 0$) of Eq. (46), we see that averaging must fail when there exists a *resonance relationship* among these free frequencies (more precisely, when there exists a resonance relationship defined by an integer vector ν for which $a_\nu \neq 0$). In general, for each ν there will be a manifold of r, called a *resonance surface*, for which the resonance relation defined by ν holds. On (or near) any such surface it is not permissible to average overall angles θ, although it may be possible to average over a subset of these angles or over certain integral linear combinations of them.

Results beyond these have been obtained in the important special case of Hamiltonian systems; the *Kolmogorov–Arnol'd–Moser (or KAM) theorem* and the *Nekhoroshev theorem* are the high point of modern perturbation theory and together give the definitive answer to the problem of the stability (in the sense of Laplace) of the (idealized Newtonian) solar system, with which this article began. Consider a system defined by a Hamiltonian function of

the form

$$H(r,\theta,\varepsilon) = H_0(r) + \varepsilon H_1(r,\theta) + \varepsilon^2 H_2(r,\theta) + \cdots, \quad (50)$$

where r and θ are as before except that $m = n$. Written in the form (46), this system is

$$\dot{r} = \varepsilon \frac{\partial H_1}{\partial \theta} + \cdots,$$
$$\dot{\theta} = -\frac{\partial H_0}{\partial r} - \varepsilon \frac{\partial H_1}{\partial r} + \cdots. \quad (51)$$

Since H_1 is assumed to be periodic in the components of θ, it may be expanded in a multiple Fourier series like Eq. (48); differentiating with respect to any component of θ then eliminates the constant term ($a_0(r)$, which is the average). It follows that $\partial H_1/\partial\theta$ has zero average, so that the (naive) first-order averaged equation for r becomes

$$\dot{r} \cong 0. \quad (52)$$

This suggests that the motion is oscillatory with nearly constant amplitudes; if this is true, the solar system (and all other systems having the same general form) will be stable (in the sense of Laplace). Of course, the argument we have given does not prove the result, unless the small-divisor problem can be overcome in this situation. This is exactly what the KAM and Nekhoroshev theorems accomplish. The KAM theorem states that (if a certain determinant does not vanish) the great majority of initial conditions will lead to motion on an invariant torus close to a torus $r =$ constant. The Nekhoroshev theorem states that even for those initial conditions that do not lie on invariant tori, the drift in r (called *Arnol'd diffusion*) takes place exponentially slowly (as $\varepsilon \to 0$). (Notice that n-dimensional tori in $2n$-dimensional space do not have an inside, so the presence of many such invariant tori does not prevent other solutions from slowly drifting off to infinity.) For details see Lochak and Meunier (1988).

In all applications of averaging and related methods to Hamiltonian systems, it is necessary to have a means of handling near-identity transformations that preserve the Hamiltonian form of the equations; that is, one needs to construct near-identity transformations that are *canonical* (or *symplectic*). Classically, such transformations can be constructed from their *generating functions* (in the sense of Hamilton-Jacobi theory); averaging procedures carried out in this way are called *von Zeipel's method*. Currently, this approach can be regarded as obsolete. It has been replaced by the method of *Lie transforms*, in which near-identity canonical transformations are generated as the flows of Hamiltonian systems in which ε takes the place of time. (The Lie method is not limited to Hamiltonian systems, but is particularly useful in this context.) Algorithmic procedures for handling near-identity transformations in this way have been developed, and they are considerably simpler than using generating functions. See Nayfeh (1973), Sec. 5.7.

2.5 Multiple-Scale Method

The earliest perturbation problem, that of planetary motion, illustrates the appeal of the idea of multiple scales. A single planet under the influence of Newtonian gravitation would travel around the sun in an elliptic orbit characterized by certain quantities called the *Keplerian elements* (the eccentricity, major axis, and certain angles giving the position of the ellipse in space). Since the actual (perturbed) motion of the

planets fits this same pattern for long periods of time, it is natural to describe the perturbed motion as "elliptical motion with slowly varying Keplerian elements." A simpler example would be a decaying oscillation of the form $e^{-\varepsilon t}\sin t$, which could be described as a periodic motion with slowly decaying amplitude. Solutions of nonlinear oscillations obtained by the method of averaging frequently have this form, where time appears both as t and as εt, the latter representing slow variation; sometimes other combinations such as $\varepsilon^2 t$ appear.

This leads to the question whether it is possible to arrive at such solutions more directly, by postulating the necessary timescales in advance. The "method of multiple scales" is the result of such an approach, and is sometimes regarded as the most flexible general method in perturbation theory, since it is applicable both to oscillatory problems (such as those covered by averaging) and to boundary layer problems (discussed below). However, this very flexibility is also its drawback, because the "method" exists in an immense variety of *ad hoc* formulations adapted to particular problems. (See Nayfeh, (1973) for examples of many of these variations.) There are two-scale methods using fast time t and slow time εt; two-scale methods using strained fast time $(\nu_0 + \varepsilon\nu_1 + \varepsilon^2\nu_2 + \cdots)t$ (similar to the strained time in the Lindstedt method) and slow time εt; multiple-scale methods using $t, \varepsilon t, \varepsilon^2 t, \ldots, \varepsilon^n t$; and methods using scales that are nonlinear functions of t. The scales to be used must be selected in advance by intuition or experience, while in other methods (averaging and matching) the required scales are generated automatically. Sometimes the length of validity of a solution can be increased by increasing the number of scales, but (contrary to popular impression) this is by no means always the case. Some problems come with more than one timescale from the beginning, for instance, problems that contain a "slowly varying parameter" depending on εt. It may seem natural to treat such a system by the method of multiple scales, but another possibility is to introduce $\tau = \varepsilon t$ as an additional independent variable subject to $\dot{\tau} = \varepsilon$. In summary, the popularity of multiple scales results from its shorter calculations, but this aside, other methods have greater power.

The general outlines of the method are as follows. Suppose the chosen timescales are t, τ, σ with $\tau = \varepsilon t, \sigma = \varepsilon^2 t$. An approximate solution is sought as a series taken to a certain number of terms, such as

$$x_0(t,\tau,\sigma) + \varepsilon x_1(t,\tau,\sigma) + \varepsilon^2 x_2(t,\tau,\sigma). \tag{53}$$

In substituting Eq. (53) into the differential equations to be solved, the definitions of the scales (such as $\tau = \varepsilon t$) are used, so that ordinary derivatives with respect to t are replaced by combinations of partial derivatives with respect to the different scales; thus

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \varepsilon\frac{\partial}{\partial \tau} + \varepsilon^2\frac{\partial}{\partial \sigma}. \tag{54}$$

From this point on (until the very end, when τ and σ are again replaced by their definitions), the separate timescales are treated as independent variables. This has the effect of changing ordinary differential equations into partial differential equations that are highly underdetermined, so that various free choices are possible in expressing the solution. The point of the method is now to make these choices

skillfully so that the final series (53) is uniformly ordered (and, it is hoped, uniformly valid) on the desired domain.

As an illustration, we return to the Van der Pol equation (34) with initial conditions $u(0) = a$, $\dot{u}(0) = 0$. Choosing timescales t and $\tau = \varepsilon t$, and writing the solution as $u \cong u_0(t, \tau) + \varepsilon u_1(t, \tau)$, one finds recursively that (with subscripts denoting partial derivatives) $u_{0tt} + u_0 = 0$ and $u_{1tt} + u_1 = -2u_{0t\tau} - u_0^2 u_{0t}$. The first equation gives $u_0(t, \tau) = A(\tau)\cos t + B(\tau)\sin t$, a modulated oscillation with slowly varying coefficients. The solution remains underdetermined, since there is nothing here to fix $A(\tau)$ and $B(\tau)$. The solution for u_0 is now substituted into the right-hand side of the differential equation for u_1, and $A(\tau)$ and $B(\tau)$ are chosen to eliminate resonant terms so that the solution for u_1 will remain bounded. (This is similar to the way the undetermined quantities are fixed in the Lindstedt method.) The result is

$$u(t) \cong u_0(t, \varepsilon t) = \frac{2a}{\sqrt{a^2 + (4 - a^2)e^{-\varepsilon t}}} \cos t. \quad (55)$$

This is the same result (to first order) as would be found by applying averaging to Eq. (39), but it has been found without any preliminary coordinate transformations. On the other hand, the possibility of constructing the solution depended on the correct initial guess as to the timescales to be used; the method of averaging generates the needed timescales automatically. The solution (55) exhibits oscillations tending toward a limit cycle that is a simple harmonic motion of amplitude 2. This is qualitatively correct, but the motion is not simple harmonic; carrying the solution to higher orders will introduce corrections.

2.6 Normal Forms

Suppose that the origin is a rest point for a system $\dot{x} = f(x)$, $x \in \mathbb{R}^n$, and it is desired to study solutions of the system near this point. (Any rest point can be moved to the origin by a shift of coordinates.) The system can be expanded in a (not necessarily convergent) series

$$\dot{x} = Ax + f_2(x) + f_3(x) + \cdots, \quad (56)$$

where A is a matrix, f_2 consists of homogeneous quadratic terms, and so forth. The matrix A can be brought into real canonical form by a change of coordinates (or into Jordan canonical form, if one is willing to allow complex variables and keep track of the conditions guaranteeing reality in the original variables). The object of normal form theory is to continue this simplification process into the higher-order terms. This is usually done recursively, one degree at a time, by applying changes of coordinates that differ from the identity by terms having the same degree as the term to be simplified. This is an example of a *coordinate perturbation* (Sec. 1.2), since it is $\|x\|$ that is small, not a perturbation parameter. However, writing $x = \varepsilon\xi$ turns Eq. (56) into

$$\dot{\xi} = A\xi + \varepsilon f_2(\xi) + \varepsilon^2 f_3(\xi) + \cdots, \quad (57)$$

which is an ordinary perturbation of a linear problem.

When A is *semisimple* (diagonalizable using complex numbers), it is possible to bring all of the terms $f_2, f_3, \ldots$ (up to any desired order) into a form that exhibits symmetries determined by A. For instance,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \alpha_1 \left(x^2 + y^2\right) \begin{bmatrix} x \\ y \end{bmatrix}$$

$$+ \beta_1 \left(x^2 + y^2\right) \begin{bmatrix} -y \\ x \end{bmatrix}$$
$$+ \alpha_2 \left(x^2 + y^2\right)^2 \begin{bmatrix} x \\ y \end{bmatrix}$$
$$+ \beta_2 \left(x^2 + y^2\right)^2 \begin{bmatrix} -y \\ x \end{bmatrix} + \cdots \quad (58)$$

is the normal form for any system having this 2×2 matrix for its linear part; all terms of even degree have been removed, and all remaining terms of odd degree are symmetrical under rotation. Because of this symmetry, the system is quite simple in polar coordinates:

$$\dot{r} = \alpha_1 r^3 + \alpha_2 r^5 + \cdots + \alpha_k r^{2k+1},$$
$$\dot{\theta} = 1 + \beta_1 r^2 + \beta_2 r^4 + \cdots + \beta_k r^{2k}. \quad (59)$$

This is solvable by quadrature, and (even without integration) the first nonzero α_j determines the stability of the origin. In general, when A is semisimple the system in normal form always gains enough symmetry to reveal certain geometrical structures called *preserved foliations*, and frequently is solvable by quadrature. These solutions have error estimates (due to truncation) similar to those of the method of averaging, to which the method of normal forms is closely related.

When A is not semisimple (its Jordan form has off-diagonal ones), the results of normalization are not so easy to explain or to use, because the nonlinear terms acquire a symmetry different from that of the linear term. Nevertheless, the normal form in such cases has proven essential to the study of such problems as the Takens–Bogdanov and Hamiltonian Hopf bifurcations. A full exposition of normal form theory is given in Murdock (2003). Popular expositions covering only the semisimple case are Nayfeh (1993) and Kahn and Zarmi (1998).

2.7 Perturbation of Stable Manifolds; Melnikov Functions

With the steadily increasing importance of nonlinear phenomena such as chaos and strange attractors, finding solutions of specific initial value problems often becomes less important than finding families (manifolds) of solutions characterized by their qualitative behavior. Many of these problems are accessible by means of perturbation theory. We will briefly describe one example. If a dynamical system has a rest point of saddle type, there will exist a *stable manifold* and an *unstable manifold* of the saddle point; the former consists of all points that approach the saddle point as $t \to \infty$, the latter of points approaching the saddle as $t \to -\infty$. In some cases, the stable and unstable manifold will coincide; that is, points that approach the saddle in the distant future also emerged from it in the distant past. (The simplest case occurs in the plane when the stable and unstable manifolds form a figure-eight pattern with the saddle at the crossing point.) If such a system is perturbed, it is important to decide whether the stable and unstable manifolds separate, or continue to intersect; and if they intersect, whether they are transverse. (The latter case leads to chaotic motion.) The criterion that in many cases decides between these alternatives is based on the *Melnikov function*; if this function has simple zeroes, the manifolds will intersect transversely and there will be a chaotic region. The Melnikov function is an integral over the homoclinic orbit of the normal component of the perturbation; the form of the integral is determined by applying regular perturbation methods to the solutions in the stable and unstable manifolds and measuring the distance

between the approximate solutions. For details see Wiggins (2003).

3 Initial and Boundary Layers

The problems considered in Secs. 2.2–2.5 are regular perturbation problems when considered on a fixed interval of time, but become singular when considered on an expanding interval such as $0 \le t \le 1/\varepsilon$. We now turn to problems that are singular even on a fixed interval. It is not easy to solve these problems even numerically, because, for sufficiently small ε, they are what numerical analysts call "stiff". Each of these problems has (in some coordinate system) a small parameter multiplying a (highest order) derivative.

3.1 Multiple-Scale Method for Initial Layer Problems

As a first example, we consider initial value problems of the form

$$\begin{aligned} &\varepsilon\ddot{u} + b(t)\dot{u} + c(t)u = 0, \\ &u(0) = \alpha, \\ &\dot{u}(0) = \frac{\beta}{\varepsilon} + \gamma. \end{aligned} \tag{60}$$

One may think, for instance, of an object of small mass ε subjected to a time-dependent restoring force and friction; at time zero, the position and velocity have just reached α and γ when the object is subjected to an impulse imparting momentum β, increasing the velocity by an amount β/ε, which is large since ε is small. We will use this example to explain two methods that are applicable to many problems in which a small parameter multiplies the highest derivative.

In approaching any perturbation problem, one first tries to understand the case $\varepsilon = 0$, but here it does not make sense to set $\varepsilon = 0$. On one hand, the differential equation drops from second order to first, and can no longer accept two initial conditions; on the other hand, the second initial condition becomes infinite. Progress can be made, however, by introducing the "stretched" time variable

$$\tau = \frac{t}{\varepsilon}. \tag{61}$$

Upon substituting Eq. (61) into Eq. (60) and writing $' = d/d\tau$, we obtain

$$\begin{aligned} &u'' + b(\varepsilon\tau)u' + \varepsilon c(\varepsilon\tau)u = 0, \\ &u(0) = \alpha, \\ &u'(0) = \beta + \varepsilon\gamma. \end{aligned} \tag{62}$$

This problem is regular (for a fixed interval of τ) and can be solved readily. For a first approximation, it suffices to set $\varepsilon = 0$ in Eq. (62) to obtain $u'' + b_0 u' = 0$ with $b_0 = b(0)$; the solution is

$$u_0^i = -\frac{\beta}{b_0}e^{-b_0\tau} + \alpha + \frac{\beta}{b_0}, \tag{63}$$

called the *first-order inner solution*. (Higher-order approximations can be found by substituting a perturbation series $u_0^i + \varepsilon u_1^i + \cdots$ into Eq. (62).) The name "inner solution" comes from the fact that Eq. (63) is only uniformly valid on an interval such as $0 \le \tau \le 1$, which translates into $0 \le t \le \varepsilon$ in the original time variable; this is a narrow "inner region" close to the initial conditions. It is necessary somehow to extend this to a solution valid for a fixed interval of t. This is of course equivalent to an expanding interval of τ, and one might attempt to solve Eq. (62) on such an expanding interval by previously discussed methods. The equation cannot be put in

a form suitable for averaging. However, the method of multiple scales is flexible enough to be adapted to this situation. One takes as timescales τ and t, and seeks a solution in the form

$$u \cong \{u_0^i(\tau) + u_0^{cor}(t)\} + \varepsilon\{u_1^i(\tau) + u_1^{cor}(t)\} + \cdots. \tag{64}$$

(We could have taken $u_0(\tau, t) + \varepsilon u_1(\tau, t) + \cdots$, but the solution turns out to be the sum of the previously calculated u^i and a "correction" u^{cor}, so it is convenient to postulate this form initially.) One can differentiate Eq. (64) with respect to τ using Eq. (61) and substitute it into Eq. (62), or equivalently differentiate with respect to t and substitute into Eq. (60). Assuming that $u^{cor}(0) = 0$ (since the inner part u^i should suffice initially), one finds that u^i must satisfy Eq. (62) as expected, and that u^{cor} satisfies a first-order differential equation together with the assumed initial condition $u^{cor}(0) = 0$; thus, u^{cor} is fully determined. At the first order, u_0^{cor} in fact satisfies the differential equation obtained from Eq. (60) by setting $\varepsilon = 0$; this is the very equation that we initially discarded as unlikely to be meaningful. Upon solving this equation (with zero initial conditions) and adding the result to u_0^i we obtain the *composite solution*

$$u_0^c = u_0^i + u_0^{cor} = -\frac{\beta}{b_0}e^{-b_0 t/\varepsilon} + \left(\alpha + \frac{\beta}{b_0}\right)\exp\left[-\int_0^t \frac{c(s)}{b(s)}\,ds\right]. \tag{65}$$

This solution is uniformly valid on any fixed interval of t.

3.2 Matching for Initial Layer Problems

Although the method of multiple scales is successful for problems of this type, it is not used as widely as the *method of matched asymptotic expansions*, probably because multiple scales require that the choices of gauges and scales be made in advance, whereas matching allows for the discovery of the correct gauges and scales as one proceeds. (Recall that gauges are the functions $\delta_i(\varepsilon)$, usually just powers ε^i, that multiply successive terms of a perturbation series; scales are the stretched time or space variables used.) To apply the matching method to Eq. (60), begin with the first-order inner solution (63) that is valid near the origin. Assume that at some distance from the origin, a good first approximation should be given by setting $\varepsilon = 0$ in Eq. (60) and discarding the initial conditions (which we have already seen do not make sense with $\varepsilon = 0$); the result is

$$u_0^o = A\exp\left[-\int_0^t \frac{c(s)}{b(s)}\,ds\right], \tag{66}$$

called the *first-order outer solution*. Since we have discarded the initial conditions, the quantity A in Eq. (66) remains undetermined at this point. At this point, one compares the inner solution (63) with the outer solution (66) in an effort to determine the correct value of A so that these solutions "match." In the present instance, the inner solution decays rapidly (assuming $b_0 > 0$) to $\alpha + \beta/b_0$, while the outer solution has A as its initial value (at $t = 0$). One might try to determine where the "initial layer" ends, and choose A so that u_0^i and u_0^o agree at that point; but in fact it is sufficient to set $A = \alpha + \beta/b_0$ on the assumption that the inner solution reaches this value at a point close enough to $t = 0$ to allow taking it as the initial condition for the outer solution. Finally, we note that adding the inner and outer solutions would duplicate the quantity $\alpha + \beta/b_0$ with which one ends and the other begins, so we subtract this "common part" u^{io} of the inner and outer solutions to obtain the

composite solution

$$u^c = u^i + u^o - u^{io}, \tag{67}$$

which is equal to the result (65) obtained by multiple scales.

In the last paragraph, we have cobbled together the inner and outer solution in a very *ad hoc* manner. In fact, several systematic procedures exist for carrying out the matching of u^i and u^o to any order and extracting the common part u^{io}. The most common procedure consists of what are sometimes called the *Van Dyke matching rules*, details of which will be given below. Although this procedure is simple to use, it does not always lead to correct results, in particular, when it is necessary to use logarithmic gauges. The other methods, *matching in an intermediate variable* and *matching in an overlap domain* are too lengthy to explain here (see Lagerstrom (1988)), but give better results in difficult cases. None of these methods has a rigorous justification *as a method*, although it is often possible to justify the results for a particular problem or class of problems. Occasionally, one encounters problems in which the inner and outer solutions cannot be matched. These cases sometimes require a "triple deck," that is, a third (or even fourth) layer timescale. In other cases, there does not exist a computable asymptotic approximation to the exact solution.

To explain the Van Dyke matching rules, we will first assume that the inner and outer solutions $u^i(\tau, \varepsilon)$ and $u^o(t, \varepsilon)$ have been computed to some order ε^k. In the problem we have been studying, the outer solution contains undetermined constants whose value must be determined, and the inner solution contains none, but in more general problems to be considered below there may be undetermined constants in both. It is important to understand that the inner and outer solutions are naturally computed in such a way that u^i is "expanded in powers of ε with τ fixed" while u^o is "expanded in powers of ε with t fixed." We are about to reexpand each solution with the opposite variable fixed. *The first step* is to express u^i in the outer variable t by setting $\tau = \varepsilon t$. The resulting function of t and ε is then expanded in powers of ε to order ε^k, holding t fixed. This new expansion is called u^{io}, the *outer expansion of the inner solution*. Notice that in computing u^{io} we retain only the terms of degree $\leq k$, so that in effect part of u^i is discarded because it moves up to order higher than k; the meaning of this is that the discarded terms of u^i are insignificant, at the desired order of approximation, in the outer region. *The second step* is to express u^o in the inner variable τ by setting $t = \tau/\varepsilon$, and expand the resulting function of τ and ε in powers of ε to order k holding τ constant. The result, called u^{oi} or the *inner expansion of the outer solution*, contains those parts of u^o that are significant in the inner region (to order k), arranged according to their significance in the inner region. *The third step* is to set $u^{io} = u^{oi}$ and use this equation to determine the unknown constants. The rationale for this is that if the domains of validity of the inner and outer regions overlap, then, since the inner solution is valid in the overlap, but the overlap belongs to the outer region, u^{io}, which is the inner solution stripped of the part that is insignificant in the outer region, should be valid there; similarly, since the outer solution is valid in the overlap, but the overlap belongs to the inner region, u^{oi} should be valid there. Now in setting $u^{io} = u^{oi}$ it is not possible to carry out the necessary computations unless both are expressed in the same variable, so it is necessary to choose either t or τ and

express both sides in that variable before attempting to determine the unknown constants. *The fourth step* is to compute the composite solution $u^c = u^i + u^o - u^{io}$. At this stage, u^{io} (which is equal to u^{oi}) is known as the *common part* of u^i and u^o; it is subtracted because otherwise it would be represented twice in the solution.

3.3 Slow–Fast Systems

The systems considered above, and many others, can be put into the form

$$\dot{x} = f(x, y, \varepsilon),$$
$$\varepsilon\dot{y} = g(x, y, \varepsilon), \qquad (68)$$

with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, which is called a *slow–fast system*. When $\varepsilon = 0$, the second equation changes drastically, from differential to algebraic; the motion is confined to the set $g(x, y) = 0$, called the *slow manifold*, or (when $n = m = 1$, which we now assume for simplicity) the *slow curve*. For $\varepsilon \neq 0$ the entire (x, y) plane is available, but (assuming $\partial g/\partial y < 0$, in which case we say the slow curve is *stable*) any point moves rapidly toward the slow curve and then slowly along it. These two stages of the motion can be approximated separately as inner and outer solutions and then matched. To obtain the inner solution (the rapid part), one rescales time by setting $t = \varepsilon\tau$ and obtains (with $' = d/d\tau$)

$$x' = \varepsilon f(x, y, \varepsilon),$$
$$y' = g(x, y, \varepsilon), \qquad (69)$$

in which ε no longer multiplies a derivative. This problem is regular (Sec. 2.1) on finite intervals of τ, which are short intervals of t. For details of the matching see Murdock (1999) Sec. 7.7 and Smith (1985) Chs. 6, 7.

An interesting case arises when the slow curve is S shaped, with the upper and lower branches stable and the middle section (the doubled-over part) unstable. A point can move along a stable branch until it reaches a vertical tangent point, then "fall off" and move rapidly to the other stable branch, then move along that branch to the other vertical tangent point and "fall off" the other way, leading to a cyclic motion called a *relaxation oscillation*. In a further, very unusual scenario, the point may actually turn the corner at the vertical tangent point and follow the unstable branch for some distance before "falling." This rather recently discovered phenomenon is called a *canard*. The explanation of canards is that several timescales become involved; the solution is actually "falling" away from the unstable branch all the time, but doing so at a rate that is slow even compared to the already slow motion along the branch. For relaxation oscillations and canards, see Grasman (1987).

Recently, an approach to slow–fast systems (in any number of dimensions) called *geometric singular perturbation theory* has come into prominence. Initiated by Fenichel, the idea is to prove that Eq. (68) for ε near zero has an actual invariant manifold close to the slow manifold defined above, and that solutions off this manifold are (in the stable case) asymptotic to solutions in the manifold, with asymptotic phase. The emphasis is on a clear geometric description of the motion rather than on computation, but computational aspects are included. A good introduction is Jones (1994).

3.4 Boundary Layer Problems

Problems in which a small parameter multiplies the highest derivative are encountered among boundary value problems at least as frequently as among initial

value problems. Since the basic ideas have been covered in the previous sections, it is only necessary to point out the differences that arise in the boundary value case. Either the multiple-scale or matching methods may be used; we will use matching. The method will be illustrated here with an ordinary differential equation; a partial differential equation will be treated in Sec. 3.5.

Consider the problem

$$\begin{aligned} &\varepsilon y'' + b(x)y' + c(x)y = 0, \\ &y(0) = \alpha, \\ &y(1) = \beta \end{aligned} \tag{70}$$

on the interval $0 \le x \le 1$. The differential equation here is the same as Eq. (60), only the independent variable is a space variable rather than time in view of the usual applications. If $b(x)$ is positive throughout $0 < x < 1$, there will be a boundary layer at the left endpoint $x = 0$; if negative, the boundary layer will be at the right endpoint; and if $b(x)$ changes sign, there may be internal transition layers as well. We will consider the former case. To the first order, the outer solution y^o will satisfy the first-order equation $b(x)y' + c(x)y = 0$ obtained by setting $\varepsilon = 0$ in Eq. (70); it will also satisfy the right-hand boundary condition $y(1) = \beta$. Therefore, the outer solution is completely determined. The first-order inner solution y^i will satisfy the equation $d^2y/d\xi^2 + b_0 y = 0$ with $b_0 = b(0)$, obtained by substituting the stretched variable $\xi = x/\varepsilon$ into Eq. (70) and setting $\varepsilon = 0$; it will also satisfy the left-hand boundary condition $y(0) = \alpha$. Since this is a second-order equation with only one boundary condition, it will contain an undetermined constant that must be identified by matching the inner and outer solutions. The differential equations satisfied by the inner and outer solutions are the same as in the case of Eq. (60), the only difference being that this time the constant that must be fixed by matching belongs to the inner solution rather than the outer.

3.5 WKB Method

There are a great variety of problems that are more degenerate than the one we have just discussed, which can exhibit a wide range of exotic behaviors. These include internal layers, in which a stretched variable such as $\xi = (x - x_0)/\varepsilon$ must be introduced around a point x_0 in the interior of the domain; triple decks, in which two stretched variables such as x/ε and x/ε^2 must be introduced at one end; and problems in which the order of the differential equation drops by more than one. The simplest example of the latter type is

$$\varepsilon^2 y'' + a(x)y = 0. \tag{71}$$

This problem is usually addressed by a technique called the *WKB* or *WKBJ method*. This method is rather different in spirit than the others we have discussed, because it depends heavily on the linearity of the perturbed problem. Rather than pose initial or boundary value problems, one finds approximations for two linearly independent solutions of the linear equation (71) on the whole real line. The general solution then consists of the linear combinations of these two. If $a(x) = k^2(x) > 0$, these approximate solutions are

$$y^{(1)} \cong \frac{1}{\sqrt{k(x)}} \cos \frac{1}{\varepsilon} \int k(x)\, dx \tag{72}$$

and

$$y^{(2)} \cong \frac{1}{\sqrt{k(x)}} \sin \frac{1}{\varepsilon} \int k(x)\, dx. \tag{73}$$

If $a(x) = -k^2(x) < 0$ the two solutions are

$$y^{(1),(2)} \cong \frac{1}{\sqrt{k(x)}} \exp \frac{1}{\varepsilon} \int \pm k(x)\, dx. \quad (74)$$

If $a(x)$ changes sign, one has a difficult situation called a *turning point problem*. This can be addressed in various ways by matching solutions of these two types or by using Airy functions. The latter are solutions of the differential equation $y'' + xy = 0$, which is the simplest problem with a turning point at the origin. These Airy functions can be considered as known (they can be expressed using Bessel functions of order 1/3), and solutions to more general turning point problems can be expressed in terms of them. For an introduction to turning point problems see Lakin and Sanchez (1970), Chapter 2, and for theory see Wasow (1976), Chapter 8.

3.6 Fluid Flow

We will conclude this section with a brief discussion of the problem of fluid flow past a flat plate because of its historical importance (see the Introduction) and because it illustrates several aspects of perturbation theory that we have avoided so far: the use of perturbation theory for partial differential equations, and the need to combine undetermined scales with undetermined gauges. The classic reference for this material is Van Dyke (1975). Consider a plane fluid flow in the upper half plane, with a "flat plate" occupying the interval $0 \le x \le 1$ on the x-axis; that is, the fluid will adhere to this interval, but not to the rest of the x-axis. The stream function $\psi(x, y)$ of such a fluid will satisfy

$$\begin{aligned} &\varepsilon(\psi_{xxxx} + 2\psi_{xxyy} + \psi_{yyyy}) \\ &\quad - \psi_y(\psi_{xxx} + \psi_{xyy}) \\ &\quad + \psi_x(\psi_{xxy} + \psi_{yyy}) = 0 \end{aligned} \quad (75)$$

with $\psi(x, 0) = 0$ for $-\infty < x < \infty$, $\psi_y(x, 0) = 0$ for $0 \le x \le 1$, and $\psi(x, y) \to y$ as $x^2 + y^2 \to \infty$. The latter condition describes the flow away from the plate, and this in fact gives the leading order outer solution as

$$\psi^o(x, y) = y. \quad (76)$$

To find an inner solution, we stretch y by an undetermined scale factor,

$$\eta = \frac{y}{\mu(\varepsilon)}, \quad (77)$$

and expand the inner solution using undetermined gauges, giving (to first order)

$$\psi^i = \delta(\varepsilon)\Psi(x, \eta). \quad (78)$$

Substituting this into Eq. (75) and discarding terms that are clearly of higher-order yields

$$\frac{\varepsilon}{\mu}\Psi_{\eta\eta\eta\eta} + \delta\left[\Psi_x\Psi_{\eta\eta\eta} - \Psi_\eta\Psi_{x\eta\eta}\right] = 0. \quad (79)$$

The relative significance of ε/μ and δ has not yet been determined, but if either of them were dominant, the other term would drop out of Eq. (79) to first order, and the resulting solution would be too simple to capture the behavior of the problem. So we must set

$$\frac{\varepsilon}{\mu} = \delta \quad (80)$$

and conclude that the first-order inner solution satisfies

$$\Psi_{\eta\eta\eta\eta} + \Psi_x\Psi_{\eta\eta\eta} - \Psi_\eta\Psi_{x\eta\eta} = 0. \quad (81)$$

It is not possible to solve Eq. (81) in closed form, but it is possible to express the solution as

$$\Psi(x, \eta) = \sqrt{2x} f\left(\frac{\eta}{\sqrt{2x}}\right), \qquad (82)$$

where f is the solution of the ordinary differential equation $f''' + ff'' = 0$ with $f(0) = 0$, $f'(0) = 0$, and $f'(\infty) = 1$. In attempting to match the inner and outer solutions, it is discovered that this is only possible if $\delta = \mu$. Together with Eq. (8) this finally fixes the undetermined scales and gauges as

$$\delta = \mu = \sqrt{\varepsilon}. \qquad (83)$$

Upon attempting to continue the solution to higher orders, obstacles are encountered that can only be overcome by introducing triple decks and other innovations. See Sychev (1998) and Rothmayer and Smith (1998).

4
Perturbations of Matrices and Spectra

In this section, we address the question: if

$$A(\varepsilon) = A_0 + \varepsilon A_1 + \cdots \qquad (84)$$

is a matrix or linear operator depending on a small parameter, and the spectrum of A_0 is known, can we determine the spectrum of $A(\varepsilon)$ for small ε? For the case of a matrix, the spectrum is simply the set of eigenvalues (values of λ for which $Av = \lambda v$ for some nonzero column vector v called an *eigenvector*). More generally, the spectrum is defined as the set of λ for which $A - \lambda I$ is not invertible; for linear transformations on infinite-dimensional spaces (such as Hilbert or Banach spaces), this need not imply the existence of an eigenvector. Our attention here will be focused on the matrix case, but many of the procedures (excluding those that involve the determinant or the Jordan normal form) are applicable as well to any operators whose spectrum consists of eigenvalues. The classical reference for the general infinite dimensional case is Kato (1966). For matrices that are not diagonalizable, one can (and should) ask not only for the eigenvalues and eigenvectors but also for generalized eigenvectors v for which $(A - \lambda I)^k v = 0$ for some integer $k > 1$.

The most direct approach (which we do not recommend) to finding the eigenvalues of Eq. (84) in the matrix case would be to examine the *characteristic equation*

$$P(\lambda, \varepsilon) = \det(A(\varepsilon) - \lambda I) = 0, \qquad (85)$$

having the eigenvalues as roots. There are standard perturbation methods for finding the roots of polynomials (see Murdock (1999) chapter 1), the simplest of which is to substitute

$$\lambda(\varepsilon) = \lambda_0 + \varepsilon\lambda_1 + \cdots \qquad (86)$$

into Eq. (85) and solve recursively for λ_i. This method works if λ_0 is a simple root of $P(\lambda, 0) = 0$; that is, it will work if A_0 has distinct eigenvalues. If there are repeated eigenvalues, then, in general, it is necessary to replace Eq. (86) with a *fractional power series* involving gauges $\delta_i(\varepsilon) = \varepsilon^{i/q}$ for some integer q that is most readily determined by using Newton's diagram. Although these are the best available perturbation methods for finding roots of general polynomials, they have two drawbacks in the case of eigenvalues: if the matrices are large, it is difficult to compute the characteristic polynomial; and, more importantly, these methods do not take into account the special features of eigenvalue problems. For instance, if $A(\varepsilon)$ is a symmetric matrix, then its eigenvalues will be real, and fractional

powers will not be required (even if A_0 has repeated eigenvalues). But the fact that A is symmetric is lost in passing to the characteristic polynomial, and one cannot take advantage of these facts.

For these reasons, it is best to seek not only the eigenvalues, but also at the same time the eigenvectors that go with them. The general procedure (which must be refined in particular situations) is to seek solutions of

$$A(\varepsilon)v(\varepsilon) = \lambda(\varepsilon)v(\varepsilon) \tag{87}$$

in the form

$$\begin{aligned} \lambda(\varepsilon) &= \lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2 + \cdots, \\ v(\varepsilon) &= v_0 + \varepsilon v_1 + \varepsilon^2 v_2 + \cdots. \end{aligned} \tag{88}$$

In the first 2 orders, the resulting equations are

$$\begin{aligned} (A_0 - \lambda_0 I)v_0 &= 0, \\ (A_0 - \lambda_0 I)v_1 &= (\lambda_1 I - A_1)v_0. \end{aligned} \tag{89}$$

We will now discuss how to solve Eq. (89) under various circumstances.

The simplest case occurs if A_0 is real and symmetric (or complex and Hermitian), and also has distinct eigenvalues. In this case, the first equation of (89) can be solved simply by choosing an eigenvector v_0 for each eigenvalue λ_0 of A_0. It is convenient to normalize v_0 to have length one, that is, $(v_0, v_0) = 1$ where $(\cdot, \cdot)$ is the inner (or "dot") product. Now we fix a choice of λ_0 and v_0 and insert these into the second equation of Eq. (89). The next step is to choose λ_1 so that the right-hand side lies in the image of $A_0 - \lambda_0 I$; once this is accomplished, it is possible to solve for v_1. To determine λ_1, we rely upon special properties of the eigenvectors of a symmetric matrix; that is, they are orthogonal (with respect to the inner product). Thus, there exists an orthogonal basis of eigenvectors in which A_0 is diagonal; examining $A_0 - \lambda_0 I$ in this basis, we see that its kernel (or null space) is spanned by v_0 and its image (or range) is spanned by the rest of the eigenvectors. Therefore, the image is perpendicular to the kernel. It follows that $(\lambda_1 I - A_1)v_0$ will lie in the image of $A_0 - \lambda_0 I$ if and only if its orthogonal projection onto v_0 is zero, that is, if and only if $(\lambda_1 v_0 - A_1 v_0, v_0) = 0$, or, using $(v_0, v_0) = 1$,

$$\lambda_1 = (A_1 v_0, v_0). \tag{90}$$

It is not necessary to find v_1 unless it is desired to go on to the next stage and find λ_2. (There is a close similarity between these steps and those of the Lindstedt method, Sec. 2.2, in which each term in the frequency expansion is determined to make the next equation solvable.)

If A_0 has distinct eigenvalues but is not symmetric, most of the last paragraph still applies, but the eigenvectors of A_0 need not be orthogonal. The vector space still decomposes as a direct sum of the image and kernel of $A_0 - \lambda_0 I$, but the inner product can no longer be used to effect the decomposition; λ_1 can still be determined but cannot be written in the form (90).

If A_0 does not have distinct eigenvalues, the situation can become quite complicated. First, suppose $A(\varepsilon)$ is symmetric for all ε, so that all A_i are symmetric. In this case, every eigenvalue has a "full set" of eigenvectors (as many linearly independent eigenvectors as its algebraic multiplicity). However, suppose that A_0 has an eigenvalue λ_0 of multiplicity two, with eigenvectors w and z. It is likely that for $\varepsilon \neq 0$ the eigenvalue λ_0 splits into two distinct eigenvalues having separate eigenvectors. In this case, it is not possible to choose an arbitrary eigenvector v_0 from the plane of z and w to use in the second equation of (89); only the limiting positions (as

$\varepsilon \to 0$) of the two eigenvectors for $\varepsilon \neq 0$ are suitable candidates for v_0. Since these are unknown in advance, one must put $v_0 = az + bw$ (for unknown real a and b) into Eq. (89), then find two choices of a, b, and λ_1 that make the second equation solvable. It also may happen that the degeneracy cannot be resolved at this stage but must be carried forward to higher stages before the eigenvalues split; or, of course, they may never split.

If $A(\varepsilon)$ is not symmetric, and hence not necessarily diagonalizable, the possibilities become even worse. The example

$$A(\varepsilon) = \begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix} \tag{91}$$

shows that a full set of eigenvectors may exist when $\varepsilon = 0$ but not for $\varepsilon \neq 0$; the contrary case (diagonalizable for $\varepsilon \neq 0$ but not for $\varepsilon = 0$) is exhibited by

$$A(\varepsilon) = \begin{bmatrix} 1 & 1 \\ 0 & 1+\varepsilon \end{bmatrix}. \tag{92}$$

These examples show that the Jordan normal form of $A(\varepsilon)$ is not in general a continuous function of ε.

There is a normal form method, closely related to that of Sec. 2.6, that is successful in all cases. It consists in simplifying the terms of Eq. (84) by applying successive coordinate transformations of the form $I + \varepsilon^k S_k$ for $k = 1, 2, \ldots$ or a single coordinate transformation of the form $T(\varepsilon) = I + \varepsilon T_1 + \varepsilon^2 T_2 + \cdots$; the matrices S_k or T_k are determined recursively. It is usually assumed that A_0 is in Jordan canonical form, hence is diagonal if possible. If A_0 is diagonal and $A(\varepsilon)$ is diagonalizable, the normalized A_k will be diagonal for $k \geq 1$, so that Eq. (84) will give the asymptotic expansion of the eigenvalues and $T(\varepsilon)$ the asymptotic expansion of all the eigenvalues (as its columns). In more complicated cases, the normalized series Eq. (84) will belong to a class of matrices called the *Arnol'd unfolding* of A_0, and although it will not always be in Jordan form, it will be in the simplest form compatible with smooth dependence on ε. Still further simplifications (the *metanormal form*) can be obtained using fractional powers of ε. This theory is described in Murdock (2003), Chapter 3.

Glossary

Asymptotic approximation: An approximate solution to a perturbation problem that increases in accuracy at a known rate as the perturbation parameter approaches zero.

Asymptotic Series: A series, the partial sums of which are asymptotic approximations of some function to successively higher order.

Averaging: A method of constructing asymptotic approximations to oscillatory problems. In the simplest case, it involves replacing periodic functions by their averages to simplify the equations to be solved.

Bifurcation: Any change in the number or qualitative character (such as stability) of the solutions to an equation as a parameter is varied.

Boundary Layer: A transition layer located near the boundary of a region where boundary values are imposed.

Composite Solution: A solution uniformly valid on a certain domain, created by matching an inner solution and an outer solution each valid on part of the domain.

Gauge: A monotonic function of a perturbation parameter used to express the order of a term in an asymptotic series.

Generalized Series: An asymptotic series of the form $\sum \delta_i(\varepsilon)u_i(x,\varepsilon)$ in which the perturbation parameter ε appears both in the gauges and in the coefficients. See Poincaré Series.

Initial Layer: A transition layer located near the point at which an initial value is prescribed.

Inner Solution: An approximate solution uniformly valid within a transition layer.

Lie Series: A means of representing a near-identity transformation by a function called a *generator*. There are several forms; in Deprit's form, if $W(x,\varepsilon)$ is the generator, then the solution of $dx/d\varepsilon = W(x,\varepsilon)$ with $x(0) = y$ for small ε is a near-identity transformation of the form $x = y + \varepsilon u_1(y) + \cdots$.

Lindstedt Method: A method of approximating periodic solutions whose frequency varies with the perturbation parameter by using a scaled time variable.

Matching: Any of several methods for choosing the arbitrary constants in an inner and an outer solution so that they both approximate the same exact solution.

Multiple Scales: The simultaneous use of two or more variables having the same physical significance (for instance, time or distance) but proceeding at different "rates" (in terms of the small parameter), for instance, "normal time" t and "slow time" $\tau = \varepsilon t$. The variables are treated as if they were independent during part of the discussion, but at the end are reduced to a single variable again.

Outer Solution: An approximate solution uniformly valid in a region away from a transition layer.

Overlap Domain: A region in which both an inner and an outer approximation are valid, and where they can be compared for purposes of matching.

Perturbation Parameter: A parameter, usually denoted ε, occurring in a mathematical problem, such that the problem has a known solution when $\varepsilon = 0$ and an approximate solution is sought when ε is small but nonzero.

Perturbation Series: A finite or infinite series obtained as a formal approximate solution to a perturbation problem, in the hope that it will be uniformly asymptotically valid on some domain.

Poincaré Series: An asymptotic series of the form $\sum \delta_i(\varepsilon)u_i(x)$ in which the perturbation parameter ε appears only in the gauges. See Generalized Series.

Regular Perturbation Problem: A perturbation problem having an approximate solution in the form of a Poincaré series that is uniformly valid on the entire intended domain.

Relaxation Oscillation: A self-sustained oscillation characterized by a slow buildup of tension (in a spring, for instance) followed by a rapid release or relaxation. The rapid phase is an example of a transition layer.

Rescaled Coordinate: A coordinate that has been obtained from an original variable by a transformation depending on the perturbation parameter, usually by multiplying by a scaling factor. For instance, time t may be rescaled to give a "slow time" εt (see multiple scales) or a "strained time" $(\omega_0 + \varepsilon\omega_1 + \cdots)t$ (see Lindstedt method).

Resonance: In linear problems, an equality of two frequencies. In nonlinear problems, any integer relationship holding between two or more frequencies, of

the form $\nu_1\omega_1 + \cdots + \nu_k\omega_k = 0$, especially one involving small integers or one that produces zero denominators in a Fourier series.

Self-Excited Oscillation: An oscillation about an unstable equilibrium, that occurs because of the instability and has its own natural frequency, rather than an oscillation in response to an external periodic forcing.

Singular Perturbation Problem: A perturbation problem that cannot be uniformly approximated by a Poincaré series on the entire intended domain, although this may be possible over part of the domain. For singular problems one seeks a solution in the form of a generalized series.

Transition Layer: A small region in which the solution of a differential equation changes rapidly and in which some approximate solution (outer solution) that is valid elsewhere fails.

Triple Deck: A problem that exhibits a transition layer within a transition layer and that therefore requires the matching of three approximate solutions rather than only two.

Unfolding: A family of perturbations of a given system obtained by adding several small parameters. An unfolding is universal if (roughly) it exhibits all possible qualitative behaviors for perturbations of the given system using the least possible number of parameters.

List of Works Cited

Ellison, J.A., Sáenz, A.W., Dumas, A.S. (1990), *J. Differ. Equations* **842**, 383–403.

Golubitsky, M., Schaeffer, D. G. (1985), *Singularities and Groups in Bifurcation Theory*, Vol. 1. New York: Springer-Verlag.

Grasman, J. (1987), *Asymptotic Methods for Relaxation Oscillations and Applications*. New York: Springer-Verlag.

Iooss, G., Joseph, D. D. (1980), *Elementary Stability and Bifurcation Theory*. New York: Springer-Verlag.

Jones, C. K. R. T. (1994), *Geometric Singular Perturbation Theory, in Dynamical Systems*, (Montecatini Terme, 1994), Lecture Notes in Mathematics 1609. New York: Springer-Verlag, pp. 44–118.

Kahn, P. B., Zarmi, Y. (1998), *Nonlinear Dynamics: Exploration Through Normal Forms*. New York: Wiley.

Kato, T. (1966), *Perturbation Theory for Linear Operators*. New York: Springer-Verlag.

Lagerstrom, P. A. (1988), *Matched Asymptotic Expansions*. New York: Springer-Verlag.

Lakin, W. D., Sanchez, D. A. (1970), *Topics in Ordinary Differential Equations*. New York: Dover.

Lochak, P., Meunier, C. (1988), *Multiphase Averaging for Classical Systems*. New York: Springer-Verlag.

Murdock, J. A. (1996), *Appl. Anal.* **62**, 161–179.

Murdock, J. A. (1999), *Perturbations*. Philadelphia: Society for Industrial and Applied Mathematics.

Murdock, J. A. (2003), *Normal Forms and Unfoldings for Local Dynamical Systems*. New York: Springer-Verlag.

Nayfeh, A. (1973), *Perturbation Methods*. New York: Wiley.

Nayfeh, A. (1993), *Method of Normal Forms*. New York: Wiley.

Rothmayer, A. P., Smith, F. T. (1998), Incompressible Triple-Deck theory, in: R. W. Johnson (Ed.), *The Handbook of Fluid Dynamics*. Boca Raton: CRC Press.

Sanders, J. A., Verhulst, F. (1985), *Averaging Methods in Nonlinear Dynamical Systems*. New York: Springer-Verlag.

Smith, D. R. (1985), *Singular-Perturbation Theory*. Cambridge: Cambridge University Press.

Sychev, V. V., Ruban, A. I., Sychev, V. V., Korolev, G. L. (1998), *Asymptotic Theory of Separated Flows*. Cambridge: Cambridge University Press.

Van Dyke, M. (1975), *Perturbation Methods in Fluid Mechanics*, Annotated Edition. Stanford: Parabolic Press.

Verhulst, F. (1999), in: A. Degasperis and G. Gaeta (Ed.), *International Workshop on Symmetry and Perturbation Theory (SPT 98)*, Rome,

December 16, 1998; Singapore: World Scientific, pp. 79–95.
Wasow, W. (1976), *Asymptotic Expansions for Ordinary Differential Equations*. Huntington: Robert E. Krieger Publishing Company.
Wiggins, S. (2003), *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Second edition. New York: Springer-Verlag.

Further Reading

Andrianov, I. V., Manevitch, L. I. (2002), *Asymptotology*. Dordrecht: Kluwer.
Bush, A. W. (1992), *Perturbation Methods for Engineers and Scientists*. Boca Raton: CRC Press.
Hinch, E. J. (1991), *Perturbation Methods*. Cambridge: Cambridge University Press.
Kevorkian, J., Cole, J. D. (1981), *Perturbation Methods in Applied Mathematics*. New York: Springer-Verlag, corrected second printing 1985.
Nayfeh, A. (1981), *Introduction to Perturbation Techniques*. New York: Wiley.
O'Malley, R. E. (1991), *Singular Perturbation Methods for Ordinary Differential Equations*. New York: Springer-Verlag.

Quantum Computation

Samuel L. Braunstein
SEECS, University of Wales, Bangor, UK

Introduction *417*
1 Computing at the Atomic Scale *419*
2 Reversible Computation *419*
3 Classical Universal Machines and Logic Gates *420*
3.1 FANOUT and ERASE *420*
3.2 Computation without ERASE *421*
4 Elementary Quantum Notation *422*
5 Logic Gates for Quantum Bits *423*
6 Logic Gates in the Laboratory *425*
7 Model Quantum Computer and Quantum Code *426*
8 Quantum Parallelism: Period of a Sequence *427*
9 The Complexity of Factoring *429*
10 Security and RSA *430*
11 Shor's Result: Factoring Numbers *431*
12 Quantum Error Correction *432*
13 Prospects *434*
Glossary *435*
Appendix *436*
List of Works Cited *436*

Introduction

A quantum computer is a device that can arbitrarily manipulate the quantum state of a part of itself. The field of quantum computation is largely a body of theoretical promises for some impressively fast algorithms that could be executed on quantum computers. However, since the first significant algorithm was proposed in 1994

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

(Shor, 1994) experimental progress has been rapid with several schemes yielding two–(Turchette et al., 1995; Monroe et al., 1995) and three–quantum-bit manipulations (Gershenfeld and Chuang, 1997). At the writing of this article it does not seem unreasonable to expect that small quantum computers capable of manipulating the quantum states of five or six two-level systems will be available within around five years. In addition, with the discovery of quantum error-correction schemes (Shor, 1995), such machines have the promise of providing long-term storage of quantum information and possibly allowing the ability to manipulate many more bits. It is still too early to tell whether the promises of rapid computation are achievable, nor is it yet well understood how broad a class of problems could be significantly speeded up by quantum computers.

Quantum computers were first discussed by Benioff (1980, 1981, 1982) in the context of simulating classical Turing machines (very elementary conventional computers) with quantum unitary evolution. Feynman (1982, 1986) considered the converse question of how well classical computers can simulate quantum systems. It was concluded that classical computers invariably suffer from an exponential slowdown in trying to simulate quantum systems, but that quantum systems could, in principle, simulate each other without this slowdown. It was Deutsch (1985a, 1985b), however, who first suggested that quantum superposition might allow quantum evolution to perform many classical computations in parallel.

To demonstrate where such capabilities may lie hidden, we review an elementary quantum mechanical experiment. The two-slit experiment is prototypic for observing one key feature of quantum mechanics: a source emits photons, electrons, or other particles that arrive at a pair of slits. These particles undergo unitary evolution and finally measurement. We see an interference pattern, with both slits open, which wholly vanishes if either slit is covered. In some sense, each particle passes through both slits in parallel. If such unitary evolution were to represent a calculation (or an operation within a calculation) then the quantum system would be performing computations in parallel. In some sense this quantum parallelism comes for free without our having to construct many copies of the "processing unit." The output of this system would be given by the constructive interference among the parallel computations.

In this article we give a tutorial on how quantum mechanics can be used to improve computation. We concentrate on the only known algorithm that demonstrates an exponential speedup relative to the best known classical algorithms. Our challenge: solving a problem that is exponentially difficult for a conventional computer – that of factoring a large number. As a prelude, we review the standard tools of computation, universal gates and machines. These ideas are then applied first to classical, dissipationless computers and then to quantum computers. A schematic model of a quantum computer is described as well as some of the subtleties in its programming. The Shor algorithm (Shor, 1994; Ekert and Jozsa, 1996) for efficiently factoring numbers on a quantum computer is presented in two parts: the quantum procedure within the algorithm and the classical algorithm that calls the quantum procedure. The mathematical structure within the factoring problem is discussed, making it clear what contribution the quantum computer makes to the calculation. The complexity of

the Shor algorithm is compared with that of factoring on conventional machines, and its relevance to public-key cryptography is noted. In addition, we discuss the experimental status of the field and also quantum error correction, which may in the long run help solve some of the most pressing difficulties. We conclude with an outlook as to the feasibility and prospects for quantum computation in the coming years.

1
Computing at the Atomic Scale

Quantum computers will perform computations at the atomic scale (DiVincenzo, 1995b). We might ask at this point how close conventional computations are to this scale already. Figure 1 shows a survey made by Keyes (1988): the number of dopant impurities in the bases of bipolar transistors used for digital logic against the year. This plot may be thought of as showing the number of electrons required to store a single bit of information. An extrapolation of the plot suggests that we might be within reach of the atomic-scale computations within the next two decades.

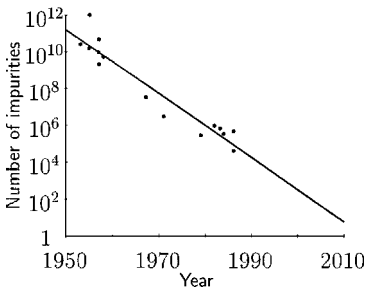


Fig. 1 Number of dopant impurities involved in logic in bipolar transistors versus year. (From Keyes, 1998; copyright 1988 by International Business Machines Corporation, reprinted with permission.)

Another way of viewing this plot is perhaps even more relevant for the development of quantum computation: conventional computers have been improving in speed and miniaturization at an exponential rate since their earliest days. Clearly there is a bound to our ability to miniaturize conventional electronics, and we will likely be touching that limit within the next 20 years. The question is raised: can we continue to expect to see an exponential improvement in performance 20 and more years from now? As we approach some of the physical limits to conventional computational construction we may begin to see a slowdown of this exponential rate. A detailed study of quantum computation may help us understand the fundamental physical limitations upon computation-conventional or otherwise.

2
Reversible Computation

What are the difficulties in trying to build a classical computing machine on such a small scale? One of the biggest problems with the program of miniaturizing conventional computers is the difficulty of dissipated heat. As early as 1961 Landauer studied the physical limitations placed on computation from dissipation (Landauer, 1961). Surprisingly, he was able to show that all but one operation required in computation could be performed in a reversible manner, thus dissipating no heat! The first condition for any deterministic device to be reversible is that its input and output be uniquely retrievable from each other. This is called logical reversibility. If, in addition to being logically reversible, a device can actually run backwards then it is called physically reversible and the second law

of thermodynamics guarantees that it dissipates no heat.

The work on classical, reversible computation laid the foundation for the development of quantum mechanical computers. On a quantum computer, programs are executed by unitary evolution of an input that is given by the state of the system. Since all unitary operators U are invertible with $U^{-1} = U^{\dagger}$, we can always "uncompute" (reverse) a computation on a quantum computer.

3 Classical Universal Machines and Logic Gates

We now review the basic logic elements used in computation and explain how conventional computers may be used for any "reasonable" computation. A reasonable computation is one that may be written in terms of some (possibly large) Boolean expression, and any Boolean expression may be constructed out of a fixed set of logic gates. Such a set (e.g., AND, OR, and NOT) is called universal. In fact we can get by with only two gates, such as AND and NOT or OR and NOT. Alternatively, we may replace some of these primitive gates by others, such as the exclusive-OR (called XOR or often controlled-NOT); then AND and XOR form a universal set. The truth tables for these gates are displayed in Table 1. Any machine that can build up arbitrary combinations of logic gates from a universal set is then a universal computer. Further, which universal set of gates is chosen makes little difference: a theorem by Muller (1956) states that the complexity of the simplest circuits needed to compute any reasonable Boolean function is affected by at most a constant multiplicative factor.

Tab. 1 Truth table defining the operation of some simple logic gates. Each row shows two input values *A* and *B* and the corresponding output values for gates AND, OR, and XOR. The output for the NOT gate is shown only for input *B*

A	*B*	AND	OR	XOR	NOT*B*
0	0	0	0	0	1
0	1	0	1	1	0
1	0	0	1	1	1
1	1	1	1	0	0

Which of the above gates is reversible? Since AND, OR, and XOR are many-to-one operations, information is lost and they are not, as they stand, logically reversible. Before we discuss how these logic gates may be made reversible we consider some nonstandard gates that we shall require.

3.1 FANOUT and ERASE

Although the above gates are sufficient for the mathematics of logic, they are not sufficient to build a machine. A useful computer will also require the FANOUT and ERASE gates (Fig. 2).

First consider the FANOUT gate: is it reversible? Certainly no information has been destroyed, and so it is at least logically reversible. Landauer (1961) showed that

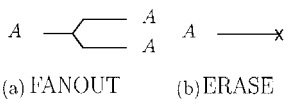


Fig. 2 Two nonstandard gates that are required to build a computer, in addition to a universal set of logic gates: (a) the FANOUT gate, which duplicates an input *A*; and (b) the ERASE gate, which deletes its input

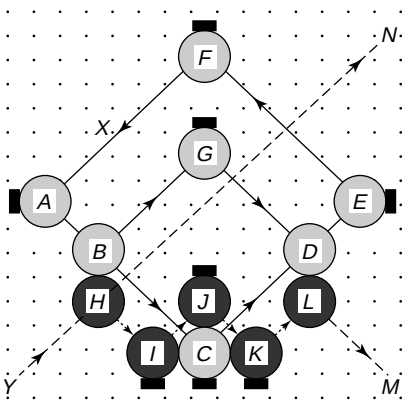


Fig. 3 A reversible measurement of the existence of a (light) ball in a trap of mirrors (dark rectangles). A (dark) ball enters the trap from *Y*. In the absence of a light ball in the trap the dark ball will follow the path *HN*. In presence of a light ball (timed to start at *X*) the dark ball will deflect the light one from its unhindered trajectory *ABCDEF* to *ABGDEF* and will follow the path *HIJKLM* itself. (From Bennett, 1998; copyright 1988 by International Business Machines Corporation, reprinted with permission.)

it could also be physically reversible. Let us describe a simple model for FANOUT based on Bennett's (1988) scheme for a reversible measurement (Fig. 3). [We note, however, that the concepts involved in this scheme come from Toffoli (1980) and Fredkin and Toffoli (1982).] Here a dark ball is used to determine the presence or absence of a second (light) ball inside a trap. The trap consists of a set of mirrors and may be thought of as a one-bit memory register. If the trap is occupied then the dark ball is reflected and leaves along direction *M* (with the light ball continuing along its original trajectory); otherwise it passes unhindered towards *N*. Upon leaving the trap, the dark ball's direction is used to populate, or not, another trap.

Let us now consider the ERASE operation, required to "clean out" the computer's memory periodically. One type of erasure can be performed reversibly: if we have a backup copy of some information, we can erase further copies by uncomputing the FANOUT gate. The difficulty arises when we wish to erase our last copy, referred to here as the primitive ERASE.

Consider a single bit represented by a pair of equally probable classical states of some particle. To erase the information about the particle's state we must irreversibly compress phase-space by a factor of two. If we allowed this compressed phase space to expand, at temperature T, to its original size, we could obtain an amount of work equal to $k_B T \ln 2$ (where k_B is Boltzmann's constant). Landauer (1961) concluded, on the basis of simple models and more general arguments about the compression of phase space, that the erasure of a bit of information at temperature T requires the dissipation of at least $k_B T \ln 2$ heat (a result known as Landauer's principle).

3.2 Computation without ERASE

Fortunately, the primitive ERASE is not absolutely essential in computation. To see why, consider what is required to compute arbitrary functions using reversible logic (where the primitive ERASE is forbidden). Landauer showed how any function $f(a)$ could be made one-to-one by keeping a copy of the input:

$$f: a \longrightarrow (a, f(a)). \qquad (1)$$

Here the bold parentheses represent an ordered set of values, in this case, two. Extra "slots" will be added (or removed) as required in our discussion below.

How can this trick be used to perform reversible logic? One solution, known as

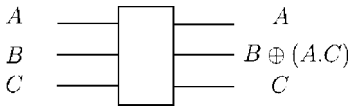


Fig. 4 Three-input, three-output universal reversible Toffoli gate. This gate is clearly reversible since a second application of it retrieves the original input

the Toffoli gate (Toffoli, 1980; Fredkin and Toffoli 1982; Landauer 1961), is shown in Fig. 4. The output of this gate may be decomposed into various gates:

$$B \oplus (A \cdot C) = \begin{cases} A \cdot C, & \text{for } B = 0 \text{ (AND)}, \\ A \oplus B, & \text{for } C = 1 \text{ (XOR)}, \\ \bar{A}, & \text{for } B = C = 1 \text{ (NOT)}, \\ A, & \text{for } B \neq C = 1 \text{ (FANOUT)}, \end{cases} \quad (2)$$

where $A \cdot B$ represents an AND gate, $A \oplus B$ represents an XOR gate, and $\bar{A}$ represents a NOT gate. We see that this gate is universal, because it performs AND, XOR, NOT, or FANOUT depending on its inputs. A combination of many such gates could then be used for any computation and would still be reversible.

As noted by Landauer, this procedure leads to an immediate problem because of the absence of the primitive ERASE. The more gates we employ, the more "junk" bits we accumulate: at each gate we must save input bits in order to preserve reversibility. In other words a computer built out of reversible logic instead of conventional, irreversible logic gates would behave like

$$f : a \longrightarrow (a, j(a), f(a)), \quad (3)$$

with many extra junk bits $j(a)$.

Bennett (1973 and 1989; Li et al., 1997) solved this problem by showing that the junk bits could be reversibly erased at intermediate steps with minimal run-time and memory costs. The spirit of Bennett's solution may be understood in terms of the following procedure:

$$f : a \longrightarrow (a, j(a), f(a)), \quad (4a)$$

$$\text{FANOUT} : (a, j(a), f(a)) \longrightarrow (a, j(a), f(a), f(a)), \quad (4b)$$

$$f^{\dagger} : (a, j(a), f(a), f(a)) \longrightarrow (a, f(a)), \quad (4c)$$

where $f^{\dagger}$ denotes uncomputing f, as opposed to computing f^{-1}. First, f is computed, producing both junk bits and the desired output. Then the FANOUT gate is applied to duplicate the output. Finally, we uncompute the original function f by running its computation backwards. This procedure removes the junk bits and the original output. The duplicate, however, remains!

This completes our discussion of the construction of classical, reversible computers. We have found that reversibility does not bar the logical design of computing machines. Before mapping these ideas to quantum systems, however, we introduce some elementary quantum mechanical notation.

4 Elementary Quantum Notation

A simple quantum system is the two-level spin-1/2 particle. Its basis states, spin down $|\downarrow\rangle$ and spin up $|\uparrow\rangle$, may be relabeled to represent binary zero and one, i.e., $|0\rangle$ and $|1\rangle$, respectively. The state of a single such particle is described by the wave function $\psi = \alpha|0\rangle + \beta|1\rangle$, i.e., a linear superposition among any of the possible "classical" states of the system. The squares of the complex coefficients $|\alpha|^2$ and $|\beta|^2$ represent the probabilities for finding the particle in the corresponding

states. Generalizing this to a set of k spin-1/2 particles we find that there are now 2^k basis states (quantum mechanical vectors that span a Hilbert space) corresponding, say, to the 2^k possible bit strings of length k. Freely moving between decimal, binary, and spin labels then we might write, for example for $k = 5$, a state $|25\rangle = |11001\rangle = |\uparrow\uparrow\downarrow\downarrow\uparrow\rangle$.

The dimensionality of the Hilbert space grows exponentially with k. In some very real sense quantum computations make use of this enormous size latent in even the smallest systems.

5 Logic Gates for Quantum Bits

In this section we describe how arbitrary logic gates may be constructed for quantum bits. We start by considering various one-bit unitary operations and a single two-bit one – the XOR operation. Combinations of these are sufficient to construct a Toffoli gate for quantum bits or, indeed, *any* unitary operation on a finite number of bits.

Start with a single quantum bit. If we represent the states $|\downarrow\rangle$ and $|\uparrow\rangle$ (i.e., $|0\rangle$ and $|1\rangle$) as the vectors $\begin{pmatrix}1\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$, respectively, then the most general unitary transformation corresponds to a 2×2 matrix of the form

$$U_\theta \equiv \begin{pmatrix} e^{i(\delta+\sigma/2+\tau/2)}\cos(\theta/2) & e^{i(\delta+\sigma/2-\tau/2)}\sin(\theta/2) \\ -e^{i(\delta-\sigma/2+\tau/2)}\sin(\theta/2) & e^{i(\delta-\sigma/2-\tau/2)}\cos(\theta/2) \end{pmatrix}, \tag{5}$$

where we typically take $\delta = \sigma = \tau = 0$ (Barenco et al., 1995a). Using this operator we can flip bits via

$$U_\pi|0\rangle = -|1\rangle \quad \text{and} \quad U_\pi|1\rangle = |0\rangle. \tag{6}$$

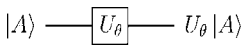


Fig. 5 Schematic of the quantum circuit diagram for a one-bit gate. The line represents a single quantum bit (such as a spin-1/2 particle). Initially, this bit has a state described by $|A\rangle$; after it has "passed" through this circuit it comes out in the state $U_\theta|A\rangle$.

The extraneous sign represents a phase factor that does not affect the logical operation of the gates and may be removed if we wish, now or at a later stage. Such one-bit computations are illustrated schematically as a quantum circuit in Fig. 5 (Barenco et al., 1995a; DiVincenzo 1995a).

Another important one-bit gate is $U_{-\pi/2}$, which maps a spin-down particle to an equal superposition of down and up:

$$U_{-\pi/2}|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle). \tag{7}$$

Consider a string of k spin-1/2 particles initially spin-down. If we apply this gate independently to each particle we obtain a superposition of every possible bit string of length k:

$$|0\rangle \longrightarrow \frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a\rangle, \tag{8}$$

where $q = 2^k$. Our computer is now in a superposition of an exponentially large number of integers a from 0 to $2^k - 1$. Suppose further that we could construct a unitary operation that maps a pair of bit strings $|a; 0\rangle$ into the pair $|a; f(a)\rangle$ for some function $f(a)$. Then such a unitary operator acting on the superposition of states

$$\frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a; 0\rangle \longrightarrow \frac{1}{\sqrt{q}}\sum_{a=0}^{q-1}|a; f(a)\rangle, \tag{9}$$

would compute *f(a)* in parallel an exponentially large number of times for the various inputs *a*. This follows from the linearity of quantum mechanics.

To see how such unitary operators may be constructed from a few elementary ones we must also consider the XOR gate (Barenco et al., 1995a; DiVincenzo 1995a). Writing the two-particle basis states as the vectors

$$|00\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |01\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \tag{10a}$$

$$|10\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad |11\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{10b}$$

we may represent the XOR gate as a unitary operator.

$$U_{\mathrm{XOR}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{11}$$

Here the first particle acts as a conditional gate to flip the state of the second particle. It is easy to check that the state of the second particle corresponds to the action of the XOR gate given in Table 1. The quantum circuit for an XOR gate is illustrated in Fig. 6. This circuit is equivalent to the elementary instruction

$$\text{if } (|A\rangle = 1)|B\rangle = \mathrm{NOT}|B\rangle \tag{12}$$

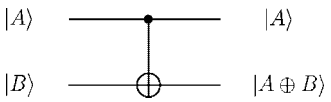


Fig. 6 Quantum circuit diagram for an XOR gate. The lower bit $|B\rangle$ is flipped whenever the upper bit $|A\rangle$ is set

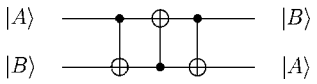


Fig. 7 Circuit for swapping a pair of bits

which may be thought of as example of quantum computer code (suggest by DiVincenzo at the 1995 Quantum Computation, Villa Gualino Workshop). The ket brackets $|\rangle$ are reminders that we are dealing with quantum rather than classical bits. The XOR gate allows us to move information around, as is illustrated in Fig. 7.

How do we construct the Toffoli gate? One major problem with this gate is that it requires three bits in and three out. Quantum mechanically, this seems to correspond to a scattering process involving three-particle collisions (Deutsch, 1989) calling for a (possibly) unreasonable control of the particles. Fortunately, the Toffoli gate may be constructed by two-particle scattering processes alone (DiVincenzo, 1995a; Barenco et al., 1995b; Sleator and Weinfurter, 1995; Deutsch et al., 1995; Lloyd, 1995). In particular, we show a construction here involving the XOR gate and some one-bit gates U_θ (Fig. 8) (Barenco et al., 1995a). Not only is the XOR sufficient for all logic operations on a quantum computer, but it can be used to construct arbitrary unitary transformations on any finite set of bits. Numerous proposals for producing such gates have

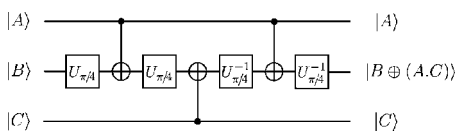


Fig. 8 Toffoli gate built from two-bit XOR gates plus some one-bit gates (Barenco et al., 1995a). This circuit introduces some extra signs in the unitary matrix U_{XOR}, which may be removed at a later stage

been considered (Ekert and Jozsa 1996; DiVincenzo 1995b), and we discuss some promising experimental results in the next section.

6
Logic Gates in the Laboratory

In this section we briefly review two recent experiments that demonstrate conditional dynamics of a type that is promising for constructing quantum logic gates (Turchette et al., 1995; Monroe et al., 1995). These two results appeared back to back in the same issue of *Physical Review Letters*.

The first experiment, by Turchette et al. (1995), demonstrated that the phase of a weak coherent optical field could be controlled by the intensity of a second coherent field at a slightly different frequency. The chief result is that such a high nonlinear susceptibility was achieved that a large phase shift (up to 16°) was produced by a change in intensity corresponding to a single photon in the second field. The coupling between the optical fields was obtained using the hyperfine level structure of cesium. A stream of Cs atoms was dropped through an optical cavity, which effectively restrained the atomic decay modes to the cavity modes. This allowed the atoms to couple strongly to the optical fields passing through the cavity with minimal incoherent emission into free space. Since coherent instead of single-photon states were used in this experiment, however, there could be no direct demonstration of the coherence retained in the final state of the optical fields. In this scheme a qubit would need to be represented by single-photon states rather than weak coherent states, but that does not appear to be a great difficulty.

The second experiment, by Monroe et al. (1995), involved a direct demonstration of an XOR gate in a radio-frequency ion trap (also known as a Paul trap). Here the lowest vibrational excitation of a single $^9Be^+$ ion in the trap and its hyperfine state represented the two qubits. A pair of off-resonant laser beams were used to drive stimulated Raman transitions between the basis states of these two qubits. The XOR gate was executed using three laser pulses, as had been suggested by Cirac and Zoller (1995). The coherence of the qubits in this system was reported to have survived for around to 10–20 XOR operations (Thompson, 1996).

Cirac and Zoller had suggested using a linear ion trap to hold a set of ions in a well localized manner by mutual electrostatic repulsion. Each ion would be "addressed" separately by its own laser. By tuning the lasers shining on individual ions to the appropriate levels, either single-qubit operations could be performed or the internal ionic state could be transferred to that of the lowest two vibrational modes of the trap. In this way two-qubit gates could be simulated between even non-contiguous ions via their interactions with the trap's vibrational modes. This ability significantly reduces the complexity of elementary operations over other proposals (Lloyd, 1993). The fact that this theoretical proposal was implemented within a few months, though in a slightly modified form, suggests that few-qubit processors could become a reality within a relatively short period of time. In Sec. 13 we discuss the short-term prospects for such machines.

In quantum computation we normally aim at having the qubits couple maximally to each other and minimally to the outside world. This lessens the actions of

environment-induced decoherence. From this perspective the latter scheme, involving ion traps, appears more nearly ideal. However, there is another class of quantum logic processor, which aims at communicating quantum information between various, possibly distantly, separated subsystems. This might allow for the combining of smaller quantum computers into larger ones operating on more qubits through combination. It is likely also to be important for the area of quantum communication and potential technologies that few-bit quantum logic processors might enable. For such tasks the former scheme involving "flying" qubits appears a more likely direction. It is possible that a mature quantum computation technology would have components incorporating the positive aspects of both the above schemes.

7 Model Quantum Computer and Quantum Code

In this section we describe a simple abstract model for a quantum computer based on a classical computer instructing a machine to manipulate a set of spins. This model has some intrinsic limitations that make designing algorithms in a high-level language somewhat tricky. We discuss some of the rules for writing such quantum computer code as a high-level language and give an example.

Consider the following model for the operation of a quantum computer. Several thousand spin-1/2 particles (or two-level systems) are initially in some well defined state, such as spin down. A classical machine takes single spins or pairs of spins and entangles them (performing an elementary one-bit operation U_θ or the two-bit XOR gate); see Figs. 9(a), 9(b), and 9(c). These stages are repeated on different pairs of spins according to the instructions of a conventional computer program. Since the spins are entangled, we must not look at the spins at intermediate stages. We must keep the quantum superposition intact. Furthermore, nothing else may interfere with the spins that could destroy their orientation or interrupt their unitary evolution. Once this well-defined cycle of manipulation is complete the orientations of the spins are measured [Fig. 9(d)]. This set of measured orientations is the output of the computation.

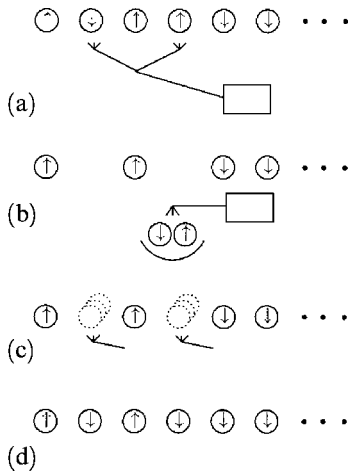


Fig. 9 Model quantum computer as pictured by Shor (presented at the 1994 Quantum Computation, Villa Gualino Workshop). Initially all particles are spin-down. In stage (a) a classical machine takes a single or pair of spirts and in stage (b) it performs a selected one-bit or two-bit operation; in stage (c) the "entangled" particles are returned to their original locations. These three stages are repeated many times in accord with the instructions given by an ordinary classical computer. When this cycle is complete stage (d) consists of measuring the state of the particles (leaving them is some particular bit string); this bit string is the result of the computation

Given this paradigm for a quantum computer, what might its high-level language (its computer code) look like? The most

serious difficulty that must be dealt with is that the quantum information is manipulated by a conventional computer in a completely blind manner – without any access to the values of this quantum information. This means that the program cannot utilize "shortcuts" conditional on the value of a quantum variable (or register or bit). For example, loops must be iterated through exactly the same number of times independent of the values of the quantum variables. Similarly, conditional branches around large pieces of code must be broken down into repeated conditions for each step. In addition, each instruction performed upon the quantum bits must be logically reversible. Thus, ordinary assignments of a value to a variable, such as $|a\rangle = n$, are not legal and must instead be performed as increments on an initially zeroed variable, such as $|a\rangle = |a\rangle + n$.

An example of such code that could run on this machine might look like this (suggested by DiVincenzo at the *1995 Quantum Computation, Villa Gualino Workshop*):

```
   do 10 k = 1, worstdiv
   |a⟩ = |a⟩ − n
   if (|a⟩ > = 0) |q⟩ = |q⟩ + 1
10 continue
   do 20 k = 1, worstdiv
   if (k > |q⟩) |a⟩ = |a⟩ + n
20 continue
```

This code fragment could be used to calculate the quotient and the remainder, placed in $|q\rangle$ and $|a\rangle$, respectively, for the division of $|a\rangle$ by n; the constant worstdiv is the worst-case number of times the loop must be traversed. Here $|q\rangle$ is initially zero. Each instruction here is either a conventional computer instruction or one involving some quantum variables. The former are direct instructions for the external computer, while the latter must be interpreted as a sequence of manipulations to be performed upon the quantum bits. As it stands, this code is *not* reversible (neither is it very efficient); e.g., the label 10 gives no specification of which routes might be used to get to it. It can, however, be easily rewritten.

8
Quantum Parallelism: Period of a Sequence

We now have sufficient ingredients to understand how a quantum computer can perform logical operations and compute just like an ordinary computer. In this section we describe an algorithm that makes use of the quantum parallelism that we have hinted at already: finding the period of a long sequence.

Consider the sequence

$$f(0), f(1), \ldots, f(q-1), \tag{13}$$

where $q \equiv 2^k$; we shall use quantum parallelism to find its period. We start with a set of initially spin-down particles, which we group into two sets (two quantum registers, or quantum variables):

$$|0; 0\rangle = |\downarrow, \downarrow, \ldots; \downarrow, \downarrow, \ldots\rangle, \tag{14}$$

the first set having k bits, the next having sufficient for our needs. (In fact other registers are required, but by applying Bennett's solution to space management they may be suppressed in our discussion here.) On each bit of the first register we perform the $U_{-\pi/2}$ one-bit operation, yielding a superposition of every possible bit-string of length k in this register:

$$\longrightarrow \frac{1}{\sqrt{q}} \sum_{a=0}^{q-1} |a; 0\rangle. \tag{15}$$

The next stage is to break down the computation corresponding to the function $f(a)$

into a set of one-bit and two-bit unitary operations. The sequence of operations is designed to map the state $|a;0\rangle$ to the state $|a;f(a)\rangle$ for any input a. Now we see that the number of bits required for this second register must be at least sufficient to store the longest result $f(a)$ for any of these computations. When, however, this sequence of operations is applied to our exponentially large superposition, instead of the single input, we obtain

$$\longrightarrow \frac{1}{\sqrt{q}} \sum_{a=0}^{q-1} |a;f(a)\rangle. \qquad (16)$$

An exponentially large amount of computation has been performed essentially for free.

The final computational step, like the first, is again a purely quantum mechanical one. Consider a discrete "quantum" Fourier transform on the first register

$$|a\rangle \longrightarrow \frac{1}{\sqrt{q}} \sum_{c=0}^{q-1} e^{2\pi iac/q}|c\rangle. \qquad (17)$$

It is easy to see that this is reversible via the inverse transform, and indeed it is readily verified to be unitary. Further, an efficient way to compute this transform with one-bit and two-bit gates has been described by Coppersmith (1994; Cleve 1994; DiVincenzo 1995b) (Fig. 10).

When this quantum Fourier transform is applied to our superposition, we obtain

$$\longrightarrow \frac{1}{q} \sum_{a=0}^{q-1} \sum_{c=0}^{q=1} e^{2\pi iac/q}|c;f(a)\rangle. \qquad (18)$$

The computation is now complete and we retrieve the output from the quantum computer by measuring the state of all spins in the first register (the first k bits). Indeed, once the Fourier transform has been performed the second register may even be discarded (Chuang et al., 1996).

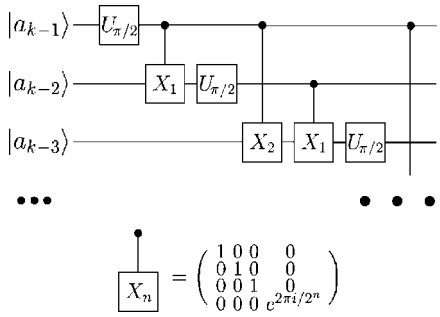


Fig. 10 Circuit for the quantum Fourier transform of the variable $|a_{k-1}\ldots a_1 a_0\rangle$ using Coppersmith's (1994; Cleve 1994; DiVincenzo 1995b) fast Fourier-transform approach. The two-bit "X_n" gate may itself be decomposed into various one-bit and XOR gates (Barenco et al., 1995a)

What will the output look like? Suppose $f(a)$ has period r so that $f(a+r) = f(a)$. The sum over a will yield constructive interference from the coefficients $e^{2\pi iac/q}$ only when c/q is a multiple of the reciprocal period $1/r$. (In fact, we must be careful that the discrete Fourier transform yields sufficient resolution to extract the multiple of the inverse period from c/q. This is always possible provided the number of bits k in the first quantum register satisfies $r^2 \leq q = 2^k$.) All other values of c/q will produce destructive interference to a greater or lesser extent. Thus, the probability distribution for finding the first register with various values is shown schematically by Fig. 11.

One complete run of the quantum computer yields a random value of c/q underneath one of the peaks in the probability of each result prob(c). That is, we obtain a random multiple of the inverse period. To extract the period itself we need only repeat this quantum computation roughly $\log(r/k)$ times in order to have a high probability for at least one of the multiples to be relatively prime to the period r – uniquely determining it (Shor

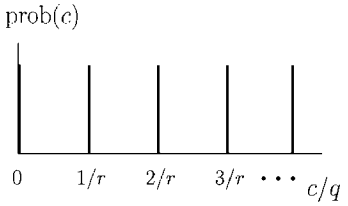


Fig. 11 Idealized plot of the probability of each result prob(c) *versus* c/q. Constructive interference produces narrow peaks at multiples of the inverse period of the sequence $1/r$. (The discrete approximation means that the peaks will actually have a nonzero width.)

1994). Thus, this algorithm yields only a probabilistic result. Fortunately, we can make this probability as high as we like.

All the above work may appear a little anticlimactic. We have gone to a lot of trouble to design a quantum computer to find the period of a sequence. The point is, however, that the sequence is calculated in parallel and is exponentially long – even for a small value of, say, $k = 270$ bits in the first register, the quantum computer has calculated and stored more results than there are particles in the universe.

This algorithm for finding the period of an exponentially long sequence on a quantum computer lies at the heart of efficiently factoring numbers. We first proceed to review the computational difficulty of factoring for conventional computers. Then we discuss the implications this computational difficulty has had for the secure transmission of private information via public key cryptosystems. We then follow these discussions with Shor's (1994) new result for efficient factoring.

9
The Complexity of Factoring

How can we quantify the difficulty of solving a problem with a conventional computer? Surely once the computer program is written and debugged we may simply let it run and wait for the answer. But this brings us to the crux of the difficulty. For a given problem how long must we expect to wait for the solution? When more carefully phrased this becomes the simplest measure of computational difficulty of an algorithm, yielding the "algorithmic complexity" of the problem.

To be more specific, without getting into technicalities, let us consider the problem of factoring a number N into its prime factors (e.g., the number 51 688 may be decomposed as $2^3 \times 7 \times 13 \times 71$). A convenient way to quantify how quickly a particular algorithm may solve this, or any, problem is to ask how the number of steps to complete the algorithm scales with the size of the "input" the algorithm is fed. For the factoring problem, this input is just the number N we wish to factor; hence the length of the input is $\log N$. (The base of the logarithm is determined by our numbering system. Thus a base of 2 gives the length in binary; a base of 10 in decimal. For example, the number 51 688 requires 16 binary digits, but only five decimal digits, to specify it.) "Reasonable" algorithms are ones that scale as some small-degree polynomial in the input size (with a degree of perhaps 2 or 3). One famous example of a fast algorithm is the fast Fourier transform, which requires roughly $O(M \log_2 M)$ steps to perform the discrete Fourier transform of M points (so for a fixed precision the input scales as M); by contrast, a conceptually simpler algorithm equivalent to matrix multiplication would require $O(M^2)$ computational steps (Press et al., 1988). This modest improvement from a quadratic to a roughly linear complexity has made many image-processing applications possible with even quite modest computers.

On conventional computers the best known factoring algorithm runs in $O(\exp[(64/9)^{1/3} \times (\ln N)^{1/3}(\ln \ln N)^{2/3}])$ steps (Odlyzko, 1995). This algorithm, therefore, scales exponentially with the input size $\log N$. For instance, in 1994 a 129-digit number (known as RSA129; Rivest et al., 1978) was successfully factored using this algorithm on approximately 1600 work-stations scattered around the world; the entire factorization took eight months (Atkins et al., 1995). Using this to estimate the prefactor of the above exponential scaling, we find that it would take roughly 800 000 years to factor a 250-digit number with the same computer power; similarly, a 1000-digit number would require 10^{25} years (significantly longer than the age of the universe). This difficulty is, however, almost certainly exaggerated since it takes no account of the constant improvement in factoring algorithms and the constant speedup of computer hardware. In fact, both of these components have shown a more or less exponential improvement over the last few decades (Odlyzko, 1995), with each contributing roughly equally to the increased computational power. The difficulty of factoring large numbers is crucial for public-key cryptosystems, such as ones used by banks. There, such codes rely on the difficulty of factoring numbers with around 250 digits.

10
Security and RSA

Cryptography as a discipline aims at minimizing the affect of the dishonest. One such situation is the need for secure communication between two parties across an insecure channel. The sender encrypts a plain-text message with an encryption key yielding the cypher text. The message is sent across an insecure medium where, we must assume, an eavesdropper may have access. The receiver takes the cypher text and uses a decryption algorithm to restore the plaintext. If we assume that the eavesdropper has access to the decryption algorithm, then the receiver must have something in addition to ensure the security of the transmitted message; this is the decryption key. According to Shannon's information theory the cypher text must contain some information about the plain-text message unless the decryption key is at least as long as the message itself (Goldreich, 1995). Such a perfect cipher was invented in 1917, and is known as the Vernam cipher or one-time pad; it requires a key equal in size to the plain-text message. Thus, for perfect security we have the problem of distributing the key itself, which must be done over a secure channel such as by trusted courier. In many situations, such as banking transactions where the volume of information is very large, this is unreasonable.

An understanding of computational complexity allows us to "circumvent" the restriction of Shannon's theory. A pseudorandom-number generator can be used to generate a long almost random key from a much smaller secret key. If the lack of randomness cannot be discovered except through an unreasonable amount of computational effort by the eavesdropper then we have a secure encryption system. An excellent example of such a scheme is the U.S. Data Encryption Standard (DES), which uses a 56-bit key (Press et al., 1988). One estimate of the security of DES suggests that for around one million dollars a special-purpose machine could be built to try all 2^{56} keys in a few hours, though security can be easily enhanced by multiple application and a larger effective key (Odlyzko, 1994). Clearly

short keys reduce the burden of key distribution amongst single pairs of users, but for n users $n(n-1)/2$ keys would be required to allow any pair to communicate securely. This becomes unwieldy for commercial applications where millions of users may be involved.

Another approach, also based on computational complexity, is known as public-key encryption. The most popular scheme, and one used in may commercial applications, is RSA encryption (Rivest et al., 1978). A person wishing to receive secret communications simply publishes a pair of numbers (N, e) that form the public key. Encryption involves converting the message to numerical data and dividing it into blocks of numbers m_j each smaller than N. Each block m_j of the message is then encrypted by its modular exponentiation

$$c_j \equiv m_j^e (\text{mod } N), \tag{19}$$

where mod N represents modulo arithmetic (the expression is computed and only the remainder after division by N is retained). The encrypted blocks c_j are then transmitted to the receiver via a public channel. Thus, RSA (and any other public key scheme) efficiently solves the key distribution problem.

Decryption by the receiver requires knowing the inverse operation, i.e., knowing the d such that

$$m_j \equiv c_j^d (\text{mod } N), \tag{20}$$

reconstructs the original message m_j from the encrypted data c_j. The size of N makes the direct determination of d too difficult. Instead, d is constructed along with the public key pair (N, e) in an efficient manner. This construction involves choosing $N = pq$ as the product of a pair of comparably sized primes, with e relatively prime to both $p-1$ and $q-1$, and solving the much simpler problem

$$ed \equiv 1 (\text{mod } p-1) \tag{21a}$$

$$ed \equiv 1 (\text{mod } q-1). \tag{21b}$$

It is important to note that the best algorithms for finding d proceed by first factoring N; thus the security of RSA relies on the assumed difficulty of factoring (Schneier, 1994).

11 Shor's Result: Factoring Numbers

Recently, an algorithm was developed by Shor (1994; Beckman et al., 1996) of AT&T for factoring numbers on a quantum computer that runs in $O((\log N)^3)$ steps. This is cubic in the input size, so that factoring a 250-digit number with such an algorithm would require only a few billion steps. The implication is that public-key cryptosystems based on factoring may be breakable. In this section we give the classical portion of Shor's algorithm, which relates factoring to finding the period of an exponentially long sequence and hence makes the problem tractable for a quantum computer.

We wish to factor the number N. It will be sufficient to find even a single factor, since then we can reduce the problem to a simpler one. First, select a number x. Euclid's algorithm (see Appendix) could be used to compute efficiently the common factors between N and x, hence reducing our problem. We therefore assume that these numbers are co-prime. Next, consider the sequence formed by the function $f(a) = x^a (\text{mod } N)$. This sequence has the form

$$1, x, \ldots, x^{r-1}, x^r, x^{r+1}, \ldots$$
$$\underbrace{1, x, \ldots, x^{r-1}}_{r \text{ terms}}, \underbrace{1, x, \ldots,}_{r \text{ terms}} \underbrace{1, x, \ldots}_{r \text{ terms}} \tag{22}$$

Here the top row is just the sequence of powers $\{x^a\}$; the bottom row is the same sequence written in modulo arithmetic, namely $\{x^a(\text{mod } N)\}$. The number r is just the first nontrivial power where $x^r \equiv 1(\text{mod } N)$. A close look at this sequence shows that it has a periodic structure with period r. Using standard algorithms this period would not be readily accessible for a long sequence. However, with the quantum computer algorithm described in Sec. 8 it could be calculated efficiently. This possibility opens up a novel way to find the factors of N, as we shall now describe.

Let us suppose that we have obtained the period r by the above quantum computer algorithm. [Note that since the period r is not known beforehand, we require $N^2 \leq q = 2^k$ for the Fourier transform step to yield sufficient resolution (Shor 1994; Ekert and Jozsa, 1996)]. If now this period is even we may proceed with our factoring algorithm. If not, we must select another x and start again. A randomly chosen x will yield a suitably even period r 50% of the time, and so not too many trials will be needed (Shor, 1994; Ekert and Jozsa, 1996).

Having chosen an x so that the sequence $\{x^a(\text{mod } N)\}$ has an even period r, we rewrite the expression $x^r \equiv 1(\text{mod } N)$ as the difference of two squares:

$$(x^{r/2})^2 - 1 \equiv 0 \text{ (mod } N). \tag{23}$$

Expressing the left-hand-side as a product between a sum and difference we obtain

$$(x^{r/2} + 1)(x^{r/2} - 1) \equiv 0 \text{ (mod } N). \tag{24}$$

This says that the product of the two terms on the left is a multiple of the number N we wish to factor. Thus, either one or the other of these terms must have a factor in common with N. The final step in the algorithm then is to calculate the greatest common divisor of these terms individually with N (see the Appendix for an efficient classical algorithm); any nontrivial common divisor will be a factor we have sought. This completes our search.

As an example, consider the number $N = 91$. Choosing $x = 3$ we find that the sequence $3^a(\text{mod } 91)$ has the form:

$$\begin{aligned} &a : 0, 1, 2, 3, 4, 5, 6, 7, \ldots \\ &3^a : 1, 3, 9, 27, 81, 243, 729, 2187, \ldots \\ &3^a(\text{mod } 91) : 1, 3, 9, 27, 81, 61, 1, 3, \ldots . \end{aligned} \tag{25}$$

A quantum computer could calculate the period in parallel; however, it is sufficient here to see by eye that this sequence has a period of $r = 6$ (since it is even we may proceed with the algorithm). Rearranging the expression $3^6 \equiv 1(\text{mod } 91)$ as discussed above we conclude that $28 \times 26 \equiv 0(\text{mod } 91)$. This implies that either gcd(28, 91) or gcd(26, 91) will be a nontrivial factor of 91 (where gcd is the greatest common divisor function). In fact, in this case, the two terms yield different factors, 7 and 13, respectively. This completes the prime factorization of 91 yielding $91 = 7 \times 13$.

12 Quantum Error Correction

Building a quantum computer is a daunting task. Even within apparently small atomic-scale systems, quantum computation runs on the enormous size of Hilbert space. Quantum computation involves building a trajectory from a standard initial state to a complex final state. The main difficulty is keeping to this trajectory. To fail is to be lost in Hilbert space. The largest problem is hypersensitivity to

perturbations, shifting the computational trajectory randomly from its path. Such perturbations come from an unintentional coupling to external noise (Unruh, 1995). In this section we briefly touch on quantum error correction and fault-tolerant computing, both introduced by Shor (1995, 1996), which promise to alleviate greatly the problem associated with unwanted perturbations.

In straight quantum error correction the state of a fragile quantum system is encoded into a quantum system having more degrees of freedom. By choosing the mapping to a suitable subspace of the larger system, a limited class of errors that occur on this larger space may be corrected. In Fig. 12 we see a circuit for Shor's original scheme. [Since then much work has been done in the last year (Steane, 1996, 1997a, 1997b; Calderbank and Shor, 1996; Calderbank et al., 1997; Laflamme et al., 1996; Knill and Laflamme, 1997; Gottesman, 1996, to cite just a few).] The unprotected single qubit $|\psi\rangle$ is processed by the left half of the circuit shown in Fig. 12 until just before the shaded region. The resulting combined nine-qubit state represents the now error-protected encoded state. Any single-qubit error (represented by the shaded region) on this encoded state may now be "undone" by sending the state through the decoding and correcting circuit shown to the right of the shaded region.

We have introduced three new circuit elements in Fig. 12, which we now explain. The first three-qubit gate involves a single control qubit (at the heavy dot) from qubit 1 to qubits 4 and 7. This gate is a shorthand notation for a pair of XOR gates (or controlled-NOT gates) between gates 1 and 4 and gates 1 and 7. The second new three-qubit gate involves two control qubits (see, for example, the last gate in the decoding circuit). This is a Toffoli gate with the condition to flip qubit 1 given by the logical AND of qubits 4 and 7 (recall that these logical operations occur separately for each branch of the wave function). (In this

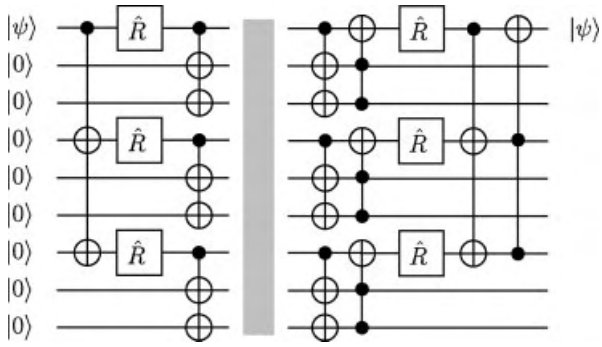


Fig. 12 Schematic of Shor's (1995) original error-correction scheme to protect one qubit in a nine-qubit code against arbitrary one-qubit decoherence. The unprotected qubit enters as qubit 1 from the left. Its error protection encoded state is just to the left of the shaded region. This region represents an arbitrary one-qubit error that may be introduced through coupling to the environment. To recover the original qubit one performs the remainder of the circuit to the right of the decoherence. (The text describes the new circuit elements.)

circuit a Toffoli gate without extraneous phases is required.) Finally, the single-qubit gate labeled $\hat{R}$ is defined by

$$\hat{R}|0\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \tag{26a}$$

$$\hat{R}|1\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}, \tag{26b}$$

which is a variation of the operation $\hat{U}_{-\pi/2}$ discussed earlier.

The above scheme and its variations allow for the long-term storage of quantum information. Unfortunately, they require that the encoding and decoding circuitry operate without error. Further, the above circuitry only slows the rate of decoherence, but does not eliminate its effect. By contrast, fault-tolerant quantum computation can get around these problems and, in principle, allows for unlimited quantum computation provided a threshold in accuracy for each elementary gate's operation in the presence of decoherence can be surpassed. Fault-tolerant computation involves a redesign of the error-correction and computation circuitry so as to allow for the imperfect operation of individual gates or for a decoherence event during computation. The future success of quantum computation will almost certainly rely on these techniques and upon our ability to determine theoretically, and experimentally surpass, the threshold (see Shor, 1996, DiVincenzo and Shor, 1996; Plenio et al., 1997; Aharonov and Ben-Or, 1996; Steane, 1997d; Zalka, 1996; Gottesman, 1997).

13 Prospects

We now describe the likely prospects for quantum computation. In this article we have discussed a single algorithm yielding an exponential speedup over conventional methods: effectively the calculation of the period of a long sequence. To date this is the only algorithm displaying such a speedup. This algorithm was applied to a traditional computer-science problem, factoring, only by recognizing a deeper structure within that problem. This requirement appears to be a general one: quantum parallelism will only yield an exponential speedup in problems whose structure avoids the need to try exponentially many solutions (Jozsa, 1991; Deutsch and Jozsa, 1992; Yao, 1993; Bennett et al., 1997). Thus, a brute-force approach to some of the hardest computational questions, known as NP-complete problems, will not succeed with the aid of quantum parallelism. Any progress for such problems will require finding a deeper structure within them. Thus, it is possible that the direction the field of quantum computation will take will be along the line of simulating (or manipulating) small quantum systems (Feynman, 1982, 1986; DiVincenzo, 1995b; Lloyd, 1996). Notwithstanding the above comments, another class of algorithms for quantum computers gives more modest gains, though still sizable in comparison with the best classical algorithms. This class is based on Grover's (1996a, 1996b, 1997; Boyer et al., 1996) database-searching algorithm which searches a "virtual" database for a specific entry. Unlike a classical database, a virtual database is a function that must be computed each time it is used. For searching a unique item from such a virtual database of N items, only $\pi\sqrt{N/4}$ iterations calling the virtual database are required for a close to 100% efficiency compared with only a 50% probability of success requiring $N/2$ iterations for a conventional computer. Certainly if quantum

computers ever get to rival the complexity of conventional computers, the Grover algorithm will play a central role in their surpassing their rival's performance for the traditionally hard problems of computer science.

How difficult will it be to build a quantum computer? Currently, several implementations are being considered by theoreticians and experimentalists worldwide (Barenco et al., 1995b; Sleator and Weinfurter, 1995; Turchette et al., 1995; Monroe et al., 1995; Cirac and Zoller, 1995; Pellizzari et al., 1995; Lloyd, 1993; Gershenfeld and Chuang, 1997; Loss and DiVincenzo, 1997). The most promising scheme to date appears to involve linear ion traps (Monroe et al., 1995; Cirac and Zoller, 1995; Pellizzari et al., 1995). There are already several theoretical studies investigating the limitations to these systems, and for numerous reasons it seems that ion-trap quantum computers will be limited to computations involving no more than around 10–20 ions (Plenio and Knight, 1996a, 1996b; Hughes et al., 1996; Steane, 1997c). It seems likely, then, that the first generation of quantum computers will not be performing traditional computations but will be used for the manipulation of small amounts of quantum information: a quantum information processor employed possibly for quantum cryptography, quantum teleportation, quantum storage, and quantum communication of quantum information. Indeed, these nontraditional tasks will probably lead to new types of technology, even with the relatively modest quantum computers we will be capable of building within a few years.

Over the next couple of decades we will approach computing at the atomic scale. Heat dissipation will become an ever increasing problem. The lessons from reversible classical computation and quantum computation may help us overcome this engineering hurdle and may even open doors to new faster algorithms.

Glossary

Algorithmic Complexity: See Computational Complexity.

Computational Complexity: A measure of the complexity of a question as measured by the number of computational steps that must be performed to find the answer. The scaling relation of the number of steps required to the size of the input information described in the question is typically quoted as the associated complexity of that question (or class of questions).

Environment: Any uncontrolled external degrees of freedom.

Fault-Tolerant Computation: Computation that is robust even when a modest amount of noise is present in the logic gates performing operations and the wires connecting them.

Modular Exponentiation: The operation of taking the remainder after division by N of some number m to a power e, is called the modular exponential; it is written m^e (mod N).

Public-Key Cryptography: A method of encoding secret information to a specific recipient who has made his/her encoding key publically known. The security of this scheme relies on the computational complexity of any attempt to invert the known encoding procedure.

Quantum Bit: See Qubit.

Quantum Error Correction: A scheme to "reverse" the affects of dissipation, decoherence, or dephasing due to inadvertent coupling to an environment.

Quantum Fourier Transform: A quantum mechanical analog of the discrete Fourier transform acting on a superposition of quantum states.

Qubit: Quantum bit of information; the amount of information that can be stored by a two-level quantum system.

Reversible Computer: A hypothetical computer that can actually "run" backwards, yielding its input from its output.

RSA: An acronym based on the names, Rivest, Shamir, and Adleman, of the inventors of the prototypically public-key cryptographic system.

Toffoli Gate: A specific three-bit gate that is universal for classical reversible computations.

Turing Machine: Prototypical classical universal computer using a read/write head and one or more paper tapes.

Appendix

Here we describe Euclid's algorithm for finding the greatest common divisor (gcd) between a pair of numbers $n_0 \geq n_1$ (Hardy and Wright 1979). The algorithm proceeds by calculating the sequence of divisions with remainder for these numbers:

$$\begin{aligned} n_0 &= d_1 \times n_1 + n_2 \\ n_1 &= d_2 \times n_2 + n_3 \\ &\vdots \\ n_{m-2} &= d_{m-1} \times n_{m-1} + n_m \\ n_{m-1} &= d_m \times n_m + 0, \end{aligned} \tag{A1}$$

where the d_m are the quotients and $n_{m-1} \geq n_m$ at each stage. The last nonzero remainder n_m yields the answer, i.e., $\gcd(n_0, n_1) = n_m$. For example, the sequence

$$91 = 3 \times 28 + 7 \tag{A2a}$$

$$28 = 4 \times 7 + 0 \tag{A2b}$$

shows that $\gcd(28, 91) = 7$ in just two steps. The worst-case number of steps required to complete Euclid's algorithm is $O(\log \log n_1)$.

List of Works Cited

Aharonov, D., Ben-Or, M. (1996), LANL pre-print No. quant-ph/9611025.

Atkins, D., Graff, M., Lenstra, A. K., Leyland, P. C. (1995), in: J. Pieprzyk, R. Safavi-Naini (Eds.), *Advances in Cryptology – ASIACRYPT '94*, Lecture Notes in Computer Science No. 917, Berlin: Springer Verlag, pp. 263–277.

Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D. P., Margolus, N., Shor, P., Sleator, T., Smolin, J. A. (1995a), *Phys. Rev. A* **52**, 3457–3467.

Barenco, A., Deutsch, D., Ekert, A. (1995b), *Phys. Rev. Lett.* **74**, 4083–4086.

Beckman, D., Chari, A. N., Devabhaktuni, S., Preskill, J. (1996), *Phys. Rev. A* **54**, 1034–1063.

Benioff, P. (1980), *J. Stat. Phys.* **22**, 563–591.

Benioff, P. (1981), *J. Math. Phys.* **22**, 495–507.

Benioff, P. (1982), *Phys. Rev. Lett.* **48**, 1581–1585.

Bennett, C. H. (1973), *IBM J. Res. Develop.* **17**, 525–532.

Bennett, C. H. (1988), *IBM J. Res. Develop.* **32**, 16–23.

Bennett, C. H. (1989), *SIAM J. Comput.* **18**, 766–776.

Bennett, C. H., Bernstein, E., Brassard, G., Vazirani, U. V. (1997), *SIAM J. Comput.* **26**, 1510–1523.

Boyer, M., Brassard, G., Hoeyer, P., Tapp, A. (1996), *Fortschr. Phys.* **46**, 493–505.

Calderbank, A. R., Shor, P. W. (1996), *Phys. Rev. A* **54**, 1098–1105.

Calderbank, A. R., Rains, E. M., Shor, P. W., Sloane, N. J. A. (1997), *Phys. Rev. Lett.* **78**, 405–408.

Chuang, I. L., Laflamme, R., Shor, P. W., Zurek, W. H. (1996), *Science* **270**, 1633–1635.

Cirac, J. I., Zoller, P. (1995), *Phys. Rev. Lett.* **74**, 4091–4094.

Cleve, R. (1994), University of British Columbia preprint.

Coppersmith, D. (1994), IBM Research Report No. RC19642.

Deutsch, D. (1985a), *Int. J. Theor. Phys.* **24**, 1–41.

Deutsch, D. (1985b), *Proc. Roy. Soc. London, A* **400**, 97–117.
Deutsch, D. (1989), *Proc. Roy. Soc. London, A* **425**, 73–90.
Deutsch, D., Jozsa, R. (1992), *Proc. R. Soc. London, A* **439**, 553–558.
Deutsch, D., Barenco, A., Ekert, A. (1995), *Proc. R. Soc. London, A* **449**, 669–677.
DiVincenzo, D. P. (1995a), *Phys. Rev. A* **51**, 1015–1022.
DiVincenzo, D. P. (1995b), *Science* **270**, 255–261.
DiVincenzo, D. P., Shor, P. W. (1996), *Phys. Rev. Lett.* **77**, 3260–3263.
Ekert, A., Jozsa, R. (1996), *Rev. Mod. Phys.* **68**, 733–753.
Feynman, R. P. (1982), *Int. J. Theor. Phys.* **21**, 467–488.
Feynman, R. P. (1986), *Found. Phys.* **16**, 507–531.
Fredkin, E., Toffoli, T. (1982), *Int. J. Theor. Phys.* **21**, 219–253.
Gershenfeld, N., Chuang, I. (1997), *Science* **275**, 350–356.
Goldreich, O. (1995), *Foundations of Cryptography (Fragments of a book)*, http://theory.lcs.mit.edu/~oded.
Gottesman, D. (1996), *Phys. Rev. A* **54**, 1862–1868.
Gottesman, D. (1997), LANL preprint No. quant-ph/9702029.
Grover, L. K. (1996a), in: *Proceedings, 28th Annual ACM Symposium on the Theory of Computing* New York: ACM, pp. 212–219.
Grover, L. K. (1996b), LANL preprint No. quant-ph/9607024.
Grover, L. K. (1997), *Phys. Rev. Lett.* **79**, 325–328.
Hardy, G. H., Wright, E. M. (1979), *An Introduction to the Theory of Numbers*, Oxford: Clarendon Press.
Hughes, R. J., James, D. F. V., Knill, E. H., Laflamme, R., Petschek, A. G. (1996), *Phys. Rev. Lett.* **77**, 3240–3243.
Jozsa, R. (1991), *Proc. R. Soc. London, A* **435**, 563–574.
Keyes, R. W. (1988), *IBM J. Res. Develop.* **32**, 24–28.
Knill, E., Laflamme, R. (1997), *Phys. Rev. A* **55**, 900–911.
Laflamme, R., Miquel, C., Paz, J. P., Zurek, W. H. (1996), *Phys. Rev. Lett.* **77**, 198–201.
Landauer, R. (1961), *IBM J. Res. Develop.* **3**, 183–191.
Li, M., Tromp, J., Vitanyi, P. (1997), *Physica D* **120**, 168–176.
Lloyd, S. (1993), *Science* **261**, 1569–1571.
Lloyd, S. (1995), *Phys. Rev. Lett.* **75**, 346–349.
Lloyd, S. (1996), *Science* **273**, 1073–1078.
Loss, D., DiVincenzo, D. P. (1997), *Phys. Rev. A* **57**, 120–126.
Monroe, C., Meekhof, D. M., King, B. E., Itano, W. M., Wineland, D. J. (1995), *Phys. Rev. Lett.* **75**, 4714–4717.
Muller, D. E. (1956), *IRE Trans. Elec. Comput.* **5**, 15–19.
Odlyzko, A. M. (1994), *AT&T Tech. J.* **73**, 17–23.
Odlyzko, A. M. (1995), *Cryptobytes: The Technical Newsletter of RSA Laboratories*, Summer.
Pellizzari, T., Gardiner, S. A., Cirac, J. I., Zoller, P. (1995), *Phys. Rev. Lett.* **75**, 3788–3791.
Plenio, M. B., Knight, P. L. (1996a), *Phys. Rev. A* **53**, 2986–2990.
Plenio, M. B., Knight, P. L. (1996b), *Proc. R. Soc. Lond. A* **453**, 2017–2041.
Plenio, M. B., Vedral, V., Knight, P. L. (1997), *Phys. Rev. A* **55**, 4593–4596.
Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1988), *Numerical Recipes: The Art of Scientific Computing*, Cambridge, U.K.: Cambridge Univ. Press, p. 228 and p. 390.
Rivest, R., Shamir, A., Adleman, L. (1978), *Communications ACM* **21**, 120–126.
Schneier, B. (1994), *Applied Cryptography*, New York: Wiley.
Shor, P. W. (1994), in: S. Goldwasser (Ed.), *Proceedings of the 35th Annual Symposium on the Foundations of Computer Science*, Los Alamitos, CA: IEEE Computer Society Press, pp. 124–134.
Shor, P. W. (1995), *Phys. Rev. A* **52**, R2493–R2496.
Shor, P. W. (1996), in: *Proceedings of the 37th Annual Symposium on the Foundations of Computer Science*, Los Alamitos, CA: IEEE Computer Society Press, pp. 56–65.
Sleator, T., Weinfurter, H. (1995), *Phys. Rev. Lett.* **74**, 4087–4090.
Steane, A. M. (1996), *Phys. Rev. Lett.* **77**, 793–797.
Steane, A. M. (1997a), *Proc. Roy. Soc. London*, **452**, 2551–2576.
Steane, A. M. (1997b), *Phys. Rev. A* **54**, 4741–4751.
Steane, A. M. (1997c), *Appl. Phys. B* **B64**, 623–643.
Steane, A. M. (1997d), *Phys. Rev. Lett.* **78**, 2252–2255.
Thompson, T. (1996), *Byte*, **21**(4), 45–54.
Toffoli, T. (1980), in: J. W. de Bakker, J. van Leeuwen (Eds.), *Automata, Languages and*

Programming, New York: Springer-Verlag, pp. 632–644.

Turchette, Q. A., Hood, C. J., Lange, W., Mabuchi, H., Kimble, H. J. (1995), *Phys. Rev. Lett.* **75**, 4710–4713.

Unruh, W. G. (1995), *Phys. Rev. A* **51**, 992–997.

Yao, A. C. C. (1993), in: *Proceedings of the 34th Annual Symposium on Foundations of Computer Science*, New York: IEEE, pp. 352–361.

Zalka, C. (1996), LANL preprint, quant-ph/9612028.

Quantum Logic

David J. Foulis
University of Massachusetts, Amherst, Massachusetts, USA (Secs. 1–3)

Richard J. Greechie
Louisiana Tech University, Ruston, Louisiana, USA (Secs. 1–3)

Maria Louisa Dalla Chiara and Roberto Giuntini
Università di Firenze, Florence, Italy (Secs. 4–10)

Introduction ***440***
1 Brief History of Quantum Logic ***441***
1.1 The Origin of Quantum Logic *441*
1.2 The Work of Birkhoff and von Neumann *441*
1.3 The Orthomodular Law *442*
1.4 The Interpretation of Meet and Join *443*
2 Standard Quantum Logic ***444***
2.1 The Orthomodular Lattice of Projections on a Hilbert Space *444*
2.2 Observables *445*
2.3 States *446*
2.4 Superposition of States *447*
2.5 Dynamics *447*
2.6 Combinations of Standard Quantum Logics *448*
3 Orthoalgebras as Models for a General Quantum Logic ***448***
3.1 Orthoalgebras *448*
3.2 Compatibility, Conjunction, and Disjunction in an Orthoalgebra *449*
3.3 Probability Measures on and Supports in an Orthoalgebra *450*
3.4 Cartesian and Tensor Products of Orthoalgebras *451*
3.5 The Logic of a Physical System *451*
3.6 The Canonical Mapping *453*
3.7 Critique of Quantum Logic *454*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

4 The Logician's Approach ***454***
5 Algebraic and Possible-World Semantics ***455***
6 Orthodox Quantum Logic ***457***
6.1 Semantic Characterizations of QL *457*
6.2 An Axiomatization of QL *459*
7 Orthologic and Unsharp Quantum Logics ***460***
8 Hilbert-Space Models of the Brouwer–Zadeh Logics ***463***
9 Partial Quantum Logics ***464***
9.1 Algebraic Semantics for WPaQL *464*
9.2 An Axiomatization of Partial Quantum Logics *465*
10 Critique of Abstract Quantum Logics ***466***
Glossary ***466***
List of Works Cited ***471***

Introduction

In formulating and studying principles of valid reasoning, logicians have been guided not only by introspection and philosophical reflection, but also by an analysis of various rational procedures commonly employed by mathematicians and scientists. Because these principles have a multitude of disparate sources, efforts to consolidate them in a single coherent system have been unsuccessful. Instead, philosophers, logicians, and mathematicians have created a panoply of competing logical formalisms, each with its own domain of applicability. Among these formalisms are Boolean-based propositional and predicate calculi, modal and multivalued logics, intuitionistic logic, and quantum logic.

Our purpose in this article is to outline the history and present some of the main ideas of quantum logic. In what follows, it will be helpful to keep in mind that there are four levels involved in any exposition of logic and its relation to the experimental sciences.

1. *Philosophical*: Addresses the epistemology of the experimental sciences. Guides and motivates the activities at the remaining levels while assimilating and coordinating the insights gained from these activities.
2. *Syntactic*: Emphasizes the formal structure of a general calculus of experimental propositions.
3. *Semantic*: Focuses on the construction of classes of mathematical models for a logical calculus.
4. *Pragmatic*: Concentrates on a specific mathematical model pertinent to a particular branch of experimental science.

For instance, studies regarding the logics associated with classical physics could be categorized as follows:

1. Philosophical writings extending back at least to Aristotle.
2. Propositional and predicate calculi.
3. The class of Boolean algebras.
4. The Boolean σ algebra of all Borel subsets of the phase space of a mechanical system.

Likewise, for quantum logic, we have

1. Philosophical writings beginning with Schrödinger, von Neumann, Bohr, Einstein, et al.
2. Quantum-logical calculi.
3. The class of orthoalgebras.
4. The lattice of projection operators on a Hilbert space.

For expository reasons, our survey proceeds roughly in the order **1, 4, 3, 2**. Thus, we give a brief history of quantum logic in Sec. 1, outline the standard quantum logic of projections on a Hilbert space in Sec. 2, introduce orthoalgebras as models for quantum logic in Sec. 3, and discuss a general quantum-logical calculus of propositions in Sec. 4.

1 Brief History of Quantum Logic

1.1 The Origin of Quantum Logic

The publication of John von Neumann's *Mathematische Grundlagen der Quantenmechanik* (1932) was the genesis of a novel system of logical principles based on propositions affiliated with quantum-mechanical entities.

According to von Neumann, a quantum-mechanical system $\mathscr{S}$ is represented mathematically by a separable (i.e., countable dimensional) complex Hilbert space $\mathscr{H}$, observables for $\mathscr{S}$ correspond to self-adjoint operators on $\mathscr{H}$, and the spectrum of a self-adjoint operator is the set of all numerical values that could be obtained by measuring the corresponding observable. Hence, a self-adjoint operator with spectrum consisting at most of the numbers 0 and 1 can be regarded as a *quantum-mechanical proposition* by identifying 0 with "false" and 1 with "true." Since a self-adjoint operator has spectrum contained in $\{0, 1\}$ if and only if it is an (orthogonal) projection onto a closed linear subspace of $\mathscr{H}$, von Neumann (1955, p. 253) observed that

> ... the relation between the properties of a physical system on the one hand, and the projections on the other, makes possible a sort of logical calculus with these.

1.2 The Work of Birkhoff and von Neumann

In 1936, von Neumann, now in collaboration with Garrett Birkhoff, reconsidered the matter of a logical calculus for physical systems and proposed an axiomatic foundation for such a calculus. They argued that the experimental propositions regarding a physical system $\mathscr{S}$ should band together to form a lattice L (Birkhoff, 1967) in which the meet and join operations are formal analogs of the *and* and *or* connectives of classical logic (although they admitted that there could be a question of the experimental meaning of these operations). They also argued that L should be equipped with a mapping carrying each proposition $a \in L$ into its negation $a' \in L$. In present-day terminology, they proposed that L forms an *orthocomplemented lattice* with $\wedge, \vee$, and $a \to a'$ as meet, join, and orthocomplementation, respectively (Kalmbach, 1983; Pták and Pulmannová, 1991).

Birkhoff and von Neumann observed that the experimental propositions concerning a classical mechanical system $\mathscr{S}$ can be identified with members of a field of subsets of the phase space for $\mathscr{S}$, (or, more accurately, with elements of a quotient of such a field by an ideal). In any case, for a classical mechanical system $\mathscr{S}$, they concluded that L forms a *Boolean algebra*.

An orthocomplemented lattice L is a Boolean algebra if and only if it satisfies the *distributive law*:

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z). \qquad (1)$$

An example in which $a \in L$ denotes the observation of a wave packet on one side of a plane, $a' \in L$ its observation on the other side, and $b \in L$ its observation in a state symmetric about the plane shows that

$$b = b \wedge (a \vee a') \neq (b \wedge a) \vee (b \wedge a') = 0,$$

so that the distributive law of classical logic breaks down even for the simplest of quantum-mechanical systems. As Birkhoff and von Neumann observed,

> ... whereas logicians have usually assumed that properties of negation were the ones least able to withstand a critical analysis, the study of mechanics points to the *distributive identities* as the weakest link in the algebra of logic.

Invoking the desirability of an "*a priori* thermo-dynamic weight of states," Birkhoff and von Neumann argued that L should satisfy a weakened version of Eq. (1), called the *modular law*, and having the following form:

$$\text{If } z \leq x, \quad \text{then } x \wedge (y \vee z) = (x \wedge y) \vee z. \qquad (2)$$

If Eq. (1) holds, then so does Eq. (2) in view of the fact that $z \leq x$ implies $x \wedge z = z$. However, Eq. (2) is weaker than Eq. (1) since the projection operators for a Hilbert space of finite dimension $n \geq 2$ form a modular, but nondistributive, lattice. Thus, Birkhoff and von Neumann proposed an *orthocomplemented modular lattice* as a model for a quantum-mechanical calculus of logic, although they admitted that it would be satisfying if one could interpret the modular law in Eq. (2) "by simpler phenomenological properties of quantum physics."

Birkhoff and von Neumann also gave an example to show that the projection operators on an infinite-dimensional Hilbert space fail to satisfy the modular law. Evidently, von Neumann considered this to be a possible serious flaw of the Hilbert-space formulation of quantum mechanics as proposed in his own *Grundlagen*. Much of von Neumann's work on continuous geometries (1960) and rings of operators (Murray and von Neumann, 1936) was motivated by his desire to construct concrete complemented modular lattices carrying an "*a priori* thermo-dynamic weight of states," that is, a continuous dimension or trace function.

1.3 The Orthomodular Law

Although the projection lattice of an infinite-dimensional Hilbert space fails to satisfy the modular law in Eq. (2), it was discovered by Husimi (1937) that it does satisfy the following weaker condition, now called the *orthomodular law*:

$$\text{If } z \leq x, \quad \text{then } x = (x \wedge z') \vee z. \qquad (3)$$

If Eq. (2) holds, then so does Eq. (3) in view of the fact that $x = x \wedge (z' \vee z)$. The same condition was rediscovered independently by Loomis (1955) and Maeda (1955) in connection with their work on extension of the Murray–von Neumann dimension theory of rings of operators to orthocomplemented lattices. An orthocomplemented lattice satisfying Eq. (3) is called an *orthomodular lattice*.

In 1957, Mackey published an expository article on quantum mechanics in Hilbert space based on notes for lectures that he was then giving at Harvard. These notes were later published in the form of a

monograph (Mackey, 1963) in which the basic principles of quantum mechanics were introduced in terms of a function

$$p = \mathrm{Prob}(A, \psi, E) \tag{4}$$

interpreted as the probability p that a measurement of the observable A in state ψ results in a value in a set of E of real numbers. The square A^2 of A is then defined by the condition

$$\mathrm{Prob}(A^2, \psi, E) = \mathrm{Prob}(A, \psi, F),$$

where F is the set of all real numbers x such that $x^2 \in E$. If $A = A^2$, then A is called a *question*. Under certain more or less reasonable hypotheses, it can be shown that the set of all questions forms an orthomodular lattice L.

The generality of Mackey's formulation and the natural way in which Mackey's questions give rise to an orthomodular lattice engendered the heady idea of a universal logical calculus for all of the experimental sciences. Such a calculus would be based on the class of all orthomodular lattices – including the Boolean algebras that would serve as models for the logics affiliated with classical mechanical systems. Would this be the realization of Leibniz's dream of a *calculus ratiocinator*? This captivating thought helped to motivate an ongoing study of the theory of orthomodular lattices by a relatively small but devoted group of researchers. An authoritative account of the resulting theory of orthomodular lattices as developed up to about 1983 can be found in Kalmbach (1983).

1.4
The Interpretation of Meet and Join

In spite of the appeal of a general scientific logic based on orthomodular lattices, a nagging question raised in the 1936 paper of Birkhoff and von Neumann was still unresolved. If, for a quantum-mechanical system, most pairs of observations are incompatible and cannot be made simultaneously, what experimental meaning can one attach to the meet $p \wedge q$ of two propositions? Two and a half decades after his initial paper with von Neumann, Birkhoff returned to this question (Birkhoff, 1961), calling for an autonomous quantum logic that draws its authority directly from experiments. (A similar question arises in connection with the logic of relativistic physics where the traditional notion of simultaneity is meaningless for spatially separated events.) After all, simultaneity is an indispensable constituent of classical propositional conjunction.

An obvious way to avoid the interpretation issue for $p \wedge q$ is to replace the assumption that $p \wedge q$ always exists with the weaker assumption that it exists if p and q are *compatible* in the sense that they can be *simultaneously tested* by means of a single experiment. In this connection, Birkhoff and von Neumann were careful to point out that "... one may regard a set of compatible measurements as a single composite 'measurement'."

Thus, for compatible propositions, experimental meaning can be bestowed upon the meet and join by regarding these connectives as the conjunction and disjunction in the usual sense of classical logic.

Although the mainstream effort to develop a viable quantum logic has concentrated on the use of orthomodular lattices as the basic models (Jauch, 1968; Piron, 1976; Mittelstaedt, 1978; Beltrametti and Cassinelli, 1981), alternative models have been introduced that avoid the interpretation issue for meet and join by invoking the notion of compatibility.

Among these are the *orthomodular posets* introduced in the early 1960s (Foulis, 1962) and the *orthoalgebras* proposed in the late 1970s (Randall and Foulis, 1978; Hardegree and Frazer, 1981; Lock and Hardegree, 1984a,b). Thus, the evolution of quantum logic from the 1930s to the present has been the story of a slow retreat from Boolean-algebra-based logic and the concurrent development of more and more general mathematical models.

2 Standard Quantum Logic

2.1 The Orthomodular Lattice of Projections on a Hilbert Space

As a mathematical model for a calculus of quantum logic, the orthomodular lattice L of projection operators on a Hilbert space $\mathscr{H}$ is called a *standard quantum logic*. Wilbur (1977) has given a purely lattice-theoretic characterization of the standard quantum logics.

In the present section, we sketch the theory of standard quantum logics, considering only the special case in which $\mathscr{H}$ is a separable Hilbert space of dimension at least three over the complex number field $\mathbb{C}$. Thus, we leave aside real or quaternionic Hilbert spaces as well as the generalized Hilbert spaces of Gross and Keller (Keller, 1980). We regard $\mathscr{H}$ as the Hilbert space corresponding to a quantum-mechanical system $\mathscr{S}$. (For the time being, we do not consider superselection rules.)

If A is a bounded operator on $\mathscr{H}$, we denote by A^* the adjoint of A. Thus, $\langle A\psi|\phi\rangle = \langle\psi|A^*\phi\rangle$ for all $\psi, \phi \in \mathscr{H}$. A bounded operator P on $\mathscr{H}$ is called a *projection* if $P = P^* = P^2$, and we define $L = L(\mathscr{H})$ to be the set of all such projection operators. If $P \in L$ and

$$\mathscr{M} = P(\mathscr{H}) = \{P(\psi)|\psi \in \mathscr{H}\} \qquad (5)$$

is the range of P, then $\mathscr{M}$ is a closed linear subspace of $\mathscr{H}$; conversely, every closed linear subspace $\mathscr{M}$ of $\mathscr{H}$ is of the form given in Eq. (5) for a uniquely determined $P \in L$. If P and $\mathscr{M}$ are related as in Eq. (5), we say that P is the *projection onto* $\mathscr{M}$. The zero operator $\mathbb{0}$ is the projection onto $\{0\}$ and the identity operator $\mathbb{1}$ is the projection onto $\mathscr{H}$.

If P is the projection onto $\mathscr{M}$ and Q is the projection onto $\mathscr{N}$, we write $P \leq Q$ if and only if $\mathscr{M}$ is a linear subspace of $\mathscr{N}$. Thus, L is a partially ordered set (poset) under $\leq$. If $\mathscr{M}$ is a closed linear subspace of $\mathscr{H}$, we write $\mathscr{M}^\perp$ for the set of all vectors in $\mathscr{H}$ that are orthogonal to every vector in $\mathscr{M}$. Then $\mathscr{M}^\perp$ is again a closed linear subspace of $\mathscr{H}$. If P is the projection onto $\mathscr{M}$, we write the projection onto $\mathscr{M}^\perp$ as P'. Note that

$$P' = \mathbb{1} - P, \quad (P')' = P, \quad \mathbb{0}' = \mathbb{1}, \quad \text{and } \mathbb{1}' = \mathbb{0}.$$

Furthermore, if $P, Q \in L$ with $P \leq Q$, then $Q' \leq P'$.

If $\mathscr{M}$ and $\mathscr{N}$ are closed linear subspaces of $\mathscr{H}$, then so is the set-theoretic intersection $\mathscr{M} \cap \mathscr{N}$. If P is the projection onto $\mathscr{M}$ and Q is the projection onto $\mathscr{N}$, we define $P \wedge Q$ to be the projection onto $\mathscr{M} \cap \mathscr{N}$, noting that $P \wedge Q$ is the meet of P and Q in the poset L. If we define $P \vee Q = (P' \wedge Q')'$, we find that $P \vee Q$ is the join of P and Q in L. Also, $P \wedge P' = \mathbb{0}$ and $P \vee P' = \mathbb{1}$, and so L forms a lattice that is orthocomplemented by $P \to P'$. Furthermore, L satisfies Eq. (3), and hence it is an orthomodular lattice.

If P is the projection onto $\mathscr{M}$ and Q is the projection onto $\mathscr{N}$, then $\mathscr{M}$ is a linear

subspace of $\mathscr{N}^{\perp}$ if and only if $P \leq Q'$. If $P \leq Q'$, we say that P and Q are *orthogonal* to each other and write $P \perp Q$. It can be shown that $P \perp Q$ if and only if $P + Q$ is again a projection operator, in which case, $P + Q = P \vee Q$.

If $(\mathscr{M}_\alpha)$ is a family of closed linear subspaces of $\mathscr{H}$, then the set-theoretic intersection $\cap_\alpha \mathscr{M}_\alpha$ is again a closed linear subspace of $\mathscr{H}$. If P_α is the projection onto $\mathscr{M}_\alpha$ for all α and P is the projection onto $\cap_\alpha \mathscr{M}_\alpha$, then P is the *greatest lower bound* in L of the family (P_α), and we write $\wedge_\alpha P_\alpha = P$. Likewise, $(\wedge_\alpha (P_\alpha)')'$ is the *least upper bound* on L of the family (P_α), and we write $\vee_\alpha P_\alpha = (\wedge_\alpha (P_\alpha)')'$. Consequently, the standard quantum logic L is actually a *complete* orthomodular lattice.

The interpretation of L as a model for a logic of quantum mechanics is based on the following premise:

> *The two-valued (true/false), experimentally testable propositions for the quantum-mechanical system $\mathscr{S}$ are represented by the projections in the standard quantum logic* L *for the Hilbert space $\mathscr{H}$ corresponding to $\mathscr{S}$.*

Furthermore, if each experimental proposition for $\mathscr{S}$ is identified with its corresponding projection $P \in L$, it is assumed that

> *if* P, Q $\in$ L, *then* P $\leq$ Q *holds if and only if* P *and* Q *are simultaneously testable and, whenever they are both tested and* P *is found to be true, then* Q *will also be true.*

2.2 Observables

As is customary, we assume that an *observable* or *dynamical variable* for the quantum-mechanical system $\mathscr{S}$ is represented by a (not necessarily bounded) self-adjoint operator A on the Hilbert space $\mathscr{H}$. In particular, then, each projection operator $P \in L$ represents an observable that, when measured, can only produce the values 1 (true) or 0 (false). As we shall see, the connection between general observables and projection observables is effected by the celebrated *spectral theorem*.

The smallest collection of subsets of the real numbers $\mathbb{R}$ that contains all open intervals and is closed under the formation of complements and countable unions is called the σ field of *real Borel sets*. A *spectral measure* is a mapping $E \to P_E$ from real Borel sets into projections such that $P_\phi = \mathbb{0}$, $P_{\mathbb{R}} = \mathbb{1}$, and, for every pairwise disjoint sequence $E_1, E_2, E_3, \ldots$ of real Borel sets,

$$\bigvee_{k=1}^{\infty} P_{E_k} = P_{E_1 \cup E_2 \cup E_3 \cup \cdots}.$$

If $E \to P_E$ is a spectral measure, $\lambda \in \mathbb{R}$, and $J = (-\infty, \lambda]$, define $P_\lambda = P_J$. Then, by the *spectral theorem* there is a one-to-one correspondence between observables A and spectral measures $E \to P_E$ such that

$$A = \int_{-\infty}^{\infty} \lambda dP_\lambda.$$

The projections in the family (P_E) are called the *spectral projections* for the observable A.

We can now be quite explicit about the connection between observables in general and projection observables in particular. Suppose that A is an observable and that (P_E) is the corresponding family of spectral projections. Then

> $\mathrm{P_E}$ *represents the experimental proposition asserting that a measurement of the observable* A *yields a result* r *that belongs to the real Borel set* $\mathrm{E} \subseteq \mathbb{R}$.

In quantum mechanics it is understood that a family of observables (A_α) is *compatible* (that is, *jointly* or *simultaneously observable*) if and only if $A_\alpha A_\beta = A_\beta A_\alpha$ for all α, β (that is, if and only if the observables commute with each other). On the basis of this understanding, we can state the following:

> *A family of projections* $(\mathrm{P}_\alpha) \subseteq \mathrm{L}$ *is compatible (that is, simultaneously testable) if and only if* $\mathrm{P}_\alpha \mathrm{P}_\beta = \mathrm{P}_\beta \mathrm{P}_\alpha$ *for all projections in the family.*

It can be shown that two observables commute with each other if and only if their spectral projections commute with each other.

It is interesting to note that the question of whether or not two projections commute can be settled in *purely lattice-theoretic terms*. Indeed, for $P, Q \in L$,

$$PQ = QP \quad \text{if and only if}$$
$$P = (P \wedge Q) \vee (P \wedge Q').$$

The equation stating the condition is a special case of the distributive law in Eq. (1); hence, in a standard quantum logic, the failure of the distributive law is a direct consequence of the fact that there are incompatible pairs of quantum-mechanical observables.

2.3 States

A bounded self-adjoint operator W on $\mathscr{H}$ is said to be *nonnegative* if $\langle W\psi|\psi\rangle \geq 0$ for all $\psi \in \mathscr{H}$. A nonnegative operator W belongs to the *trace class* if the series

$$\mathrm{tr}(W) = \sum_{\psi \in B} \langle W\psi|\psi\rangle$$

converges for an orthonormal basis $B \subseteq \mathscr{H}$. Convergence on any one orthonormal basis implies convergence on all orthonormal bases.

A *(von Neumann) density operator* on $\mathscr{H}$ is a bounded, self-adjoint, nonnegative, trace-class operator W on $\mathscr{H}$ such that $\mathrm{tr}(W) = 1$. Denote by $\Omega = \Omega(\mathscr{H})$ the set of all density operators on $\mathscr{H}$. One of the basic assumptions of statistical quantum mechanics is the following:

> *There is a one-to-one correspondence between the possible states of the system* $\mathscr{S}$ *and the density operators* $\mathrm{W} \in \Omega$ *such that, for every experimental proposition* $\mathrm{P} \in \mathrm{L}$, *tr(*$\mathrm{WP}$*) is the probability that* P *will be true when tested in the state corresponding to* W.

In accordance with this assumption, we shall identify each possible state of the system $\mathscr{S}$ with the corresponding density operator W. In particular, if A is an observable with spectral family (P_E), the probability function, Eq. (4) in Mackey's formulation, is realized as

$$\mathrm{Prob}(A, W, E) = \mathrm{tr}(WP_E). \qquad (6)$$

Equation (6), one of the fundamental equations of quantum mechanics, says that

> *the probability that a measurement of the observable* A *in the state* W *yields a result* r *in the Borel set* E *is given by tr*$(\mathrm{WP_E})$.

By a *countably additive probability measure* on the orthomodular lattice L is meant a function $\omega: L \to [0, 1] \subseteq \mathbb{R}$ such that, for every sequence $P_1, P_2, P_3, \ldots$ of pairwise orthogonal projections in L,

$$\omega(\vee_k P_k) = \sum_k \omega(P_k).$$

By a celebrated theorem of Gleason (1957), ω is a countably additive probability measure on L if and only if there is

a (uniquely determined) density operator $W \in \Omega$ such that

$$\omega(P) = \mathrm{tr}(WP) \quad \text{for all } P \in L.$$

2.4 Superposition of States

If $W_1, W_2, W_3, \ldots \in \Omega$ is a sequence of density operators and $t_1, t_2, t_3, \ldots$ is a corresponding sequence of nonnegative real numbers such that $\Sigma t_k = 1$, then $W = \Sigma_k t_k W_k$ is again a density operator, which is referred to as a *mixture* or an *incoherent superposition* of the states W_k. For instance, W could be regarded as the state of a statistical ensemble of systems for which t_k is the fraction of the systems that are in the state W_k.

A state W is called a *pure state* if it cannot be obtained as a mixture of other states. It is customary to assume that individual physical systems are always in a pure state and that mixed states apply only to statistical ensembles of systems each of which is in a pure state, or to physical systems that are interactively coupled with other physical systems.

It can be shown that $W \in \Omega$ is a pure state if and only if it is a projection onto a one-dimensional linear subspace $\mathscr{M}$ of $\mathscr{H}$. Thus, any normalized vector $\psi \in \mathscr{H}$ determines a unique pure state, namely the projection onto the linear subspace of complex multiples of ψ. Such a state is called a *vector state*, and two normalized vectors determine the same vector state if and only if each can be obtained from the other by multiplying by a complex number of modulus 1 (a *phase factor*). Every state W is a mixture of pure (that is, vector) states.

We define the *support* of $W \in \Omega$, in symbols $\mathrm{supp}(W)$, to be the set of all $P \in L$ such that $\mathrm{tr}(WP) \neq 0$. This is the same as the set of all $P \in L$ for which $WP \neq 0$. If $W = \Sigma_k t_k W_k$ is an incoherent superposition of the sequence (W_k), then it is clear that $\mathrm{supp}(W)$ is contained in the set-theoretic union $\cup_k \mathrm{supp}(W_k)$.

More generally, if (W_α) is a family of states, we say that the state W is a *superposition* of the states W_α if and only if

$$\mathrm{supp}(W) \subseteq \cup_\alpha \mathrm{supp}(W_\alpha)$$

(Bennett and Foulis, 1990). If W as well as every W_α is a pure state, then W is said to be a *coherent superposition* of the states W_α. For instance, if W is the vector state determined by $\psi \in \mathscr{H}$, each W_α is the vector state determined by $\psi_\alpha \in \mathscr{H}$, and ψ differs from a normalized linear combination of the ψ_α by a phase factor, then W is a coherent superposition of the W_α.

2.5 Dynamics

By *dynamics* is meant a study of the way in which the states (*Schrödinger picture*) or the observables (*Heisenberg picture*) of a system change or evolve in time. The Schrödinger and Heisenberg pictures are mathematically equivalent. For definiteness, we adopt the Schrödinger picture. Thus, if the space Ω of density operators represents the state space of the quantum-mechanical system $\mathscr{S}$, then the dynamical evolution of the system is represented by a function $f(t, W)$ of the time $t \in \mathbb{R}$ and the state $W \in \Omega$ such that

$$f(t, W) \in \Omega, \quad f(0, W) = W \quad \text{and}$$
$$f(t + s, W) = f(t, f(s, W)). \tag{7}$$

The understanding in Eq. (7) is that $f(t, W)$ represents the state of the system after a time interval t if it is in state W at time 0. The function f is called the *dynamical law* for the system $\mathscr{S}$.

If the dynamical law f in Eq. (7) preserves superpositions, and is continuous in a suitable sense, it can be shown

(Mackey, 1963) that there is a family (U_t) of unitary operators continuously indexed by real numbers such that

$$f(t, W) = U_t W U_t^{-1}$$

holds for all $t \in \mathbb{R}$. Hence, by a celebrated representation theorem of Stone (1932), it follows that there is a self-adjoint operator H on $\mathscr{H}$ such that

$$U_t = e^{-itH} \tag{8}$$

for all $t \in \mathbb{R}$. Equation (8) is the operator form of the *Schrödinger equation* and H is the *Hamiltonian operator* for the system.

2.6 Combinations of Standard Quantum Logics

Suppose that $\mathscr{H}_1$ and $\mathscr{H}_2$ are complex separable Hilbert spaces with corresponding standard quantum logics L_1 and L_2. There are two natural ways to combine $\mathscr{H}_1$ and $\mathscr{H}_2$ to form a composite Hilbert space $\mathscr{H}$ with its own standard quantum logic L: We can form either the direct sum $\mathscr{H} = \mathscr{H}_1 \oplus \mathscr{H}_2$ or the tensor product $\mathscr{H} = \mathscr{H}_1 \otimes \mathscr{H}_2$ (Foulis, 1989). In neither case is the structure of the resulting standard quantum logic L easy to describe in terms of the structures of L_1 and L_2.

If $\mathscr{S}_1$ and $\mathscr{S}_2$ are quantum-mechanical systems represented by corresponding Hilbert spaces $\mathscr{H}_1$ and $\mathscr{H}_2$, it is customary to regard the tensor product $\mathscr{H} = \mathscr{H}_2 \otimes \mathscr{H}_2$ as the Hilbert space corresponding to the "combined system" $\mathscr{S} = \mathscr{S}_1 + \mathscr{S}_2$ (Jauch, 1968). If this is so, then in the combination $\mathscr{S}_1 + \mathscr{S}_2$ the systems can be tightly correlated, but they cannot exert instantaneous influences on each other (Kläy et al., 1987).

If W is a state for the combined system $\mathscr{S} = \mathscr{S}_1 + \mathscr{S}_2$, there exist uniquely determined states W_1 for $\mathscr{S}_1$ and W_2 for $\mathscr{S}_2$ such that for all $P_1 \in L_1$ and all $P_2 \in L_2$

$$\mathrm{tr}(W_1 P_1) = \mathrm{tr}(W(P_1 \otimes \mathbb{1})) \quad \text{and}$$
$$\mathrm{tr}(W_2 P_2) = \mathrm{tr}(W(\mathbb{1} \otimes P_2)).$$

The states W_1 and W_2 are called *reduced states*. In general, W is not determined by W_1 and W_2, but depends on the details of the coupling between $\mathscr{S}_1$ and $\mathscr{S}_2$. However, if W is a pure state and either W_1 or W_2 is pure, then both W_1 and W_2 are pure and $W = W_1 \otimes W_2$. Therefore, if $\mathscr{S} = \mathscr{S}_1 + \mathscr{S}_2$ is in a pure state and if $\mathscr{S}_1$ and $\mathscr{S}_2$ are correlated in any way, then neither $\mathscr{S}_1$ nor $\mathscr{S}_2$ can be in a pure state.

If $\mathscr{H} = \mathscr{H}_1 \oplus \mathscr{H}_2$, a *superselection rule* (Wick et al., 1952) may be imposed, in which case the quantum logic L associated with $\mathscr{H}$ is understood to consist only of projections that commute with the projections P_1 and P_2 of $\mathscr{H}$ onto the subspaces $\mathscr{H}_1 \oplus \{0\}$ and $\{0\} \oplus \mathscr{H}_2$, respectively. In this case, L is isomorphic to the Cartesian product $L_1 \times L_2$ of the standard quantum logics L_1 and L_2, but L is no longer a standard quantum logic. If such a superselection rule is imposed, it is assumed that the superselected observables are those with spectral projections in L and the super-selected states are represented by density operators W that commute with both P_1 and P_2.

3 Orthoalgebras as Models for a General Quantum Logic

3.1 Orthoalgebras

In this section, we present an axiomatic mathematical structure called an *orthoalgebra* (Foulis et al., 1992), which generalizes the standard quantum logics. The idea is to

endow a generic orthoalgebra with an *absolute minimum of mathematical structure* so that it becomes possible to investigate the meaning and consequences of the special features that distinguish particular orthoalgebras – for instance, Boolean algebras, orthomodular lattices, or standard quantum logics – as models for a calculus of experimental propositions.

By definition, an *orthoalgebra* is a set L containing two special elements 0 and 1 and equipped with a relation $\perp$ called *orthogonality* such that, for each pair $p, q \in L$ with $p \perp q$, an *orthogonal sum* $p \oplus q$ is defined in L and subject to the following four axioms:

(*Commutativity*) If $p \perp q$, then $q \perp p$ and $p \oplus q = q \oplus p$.

(*Associativity*) If $p \perp q$ and $(p \oplus q) \perp r$, then $q \perp r, p \perp (q \oplus r)$, and $p \oplus (q \oplus r) = (p \oplus q) \oplus r$.

(*Orthocomplementation*) For each $p \in L$ there is a unique $p' \in L$ such that $p \perp p'$ and $p \oplus p' = 1$.

(*Consistency*) If $p \perp p$, then $p = 0$.

We note that every orthomodular lattice L becomes an orthoalgebra if we define $p \oplus q = p \vee q$ whenever $p \leq q'$. In particular, every Boolean algebra and every standard quantum logic is an orthoalgebra.

If L is an orthoalgebra and $p, q \in L$, we define $p \leq q$ to mean that there exists $r \in L$ with $p \perp r$ such that $p \oplus r = q$. It can be shown that L is partially ordered by $\leq$; $0 \leq p \leq 1$ and $p = p''$ hold for all $p \in L$; and, if $p \leq q$, then $q' \leq p'$. Also, if $p \perp q$, then with respect to $\leq$, $p \oplus q$ is a *minimal* upper bound for p and q; that is,

$$p, q \leq p \oplus q \text{ and there exists no } r \in L \text{ with } p, q \leq r < p \oplus q.$$

However, $p \oplus q$ may not be the *least* upper bound for p and q; that is, the conditions $r \in L$ and $p, q \leq r$ do not necessarily imply that $p \oplus q \leq r$.

If $x, y \in L$ have a least upper bound (respectively, a greatest lower bound), we write it as $x \vee y$ (respectively, as $x \wedge y$). By definition, an *orthomodular poset* is an orthoalgebra L satisfying the condition that $p \oplus q = p \vee q$ whenever $p \perp q$. An orthomodular lattice is the same thing as an orthoalgebra in which every pair of elements x, y has a meet $x \wedge y$ and a join $x \vee y$. A Boolean algebra is the same thing as an orthomodular lattice satisfying the condition that $x \wedge y = 0$ only if $x \perp y$.

By a *subalgebra* of the orthoalgebra L, we mean a subset $S \subseteq L$ such that $0, 1 \in S$ and, if $p, q \in S$ with $p \perp q$, then $p \oplus q \in S$ and $p' \in S$. Evidently, a subalgebra of an orthoalgebra is an orthoalgebra in its own right under the operations inherited from the parent orthoalgebra. If, as an orthoalgebra in its own right, a subalgebra B of L is a Boolean algebra, we refer to B as a *Boolean subalgebra* of L. If $p \perp q$ in L, then

$$B = \{0, 1, p, q, p', q', p \oplus q, (p \oplus q)'\}$$

is a Boolean subalgebra of L, so L is a set-theoretic union of Boolean subalgebras.

3.2 Compatibility, Conjunction, and Disjunction in an Orthoalgebra

We say that a subset of an orthoalgebra L is a *compatible* set if it is contained in a Boolean subalgebra of L. A compatible set of pairwise orthogonal elements is called an *orthogonal* subset of L. If L is a standard quantum logic, then a subset of L is compatible if and only if the projections in the subset commute with one another.

Let $A = \{a_1, a_2, a_3, \ldots, a_n\}$ be a finite orthogonal subset of L. Then, it can be

shown that the least upper bound

$$\vee_B A = a_1 \vee_B a_2 \vee_B a_3 \vee_B \cdots \vee_B a_n$$

as calculated in any Boolean subalgebra B of L that contains A is independent of the choice of B. Thus, we define the *orthogonal sum*

$$\oplus A = \vee_B A$$

as calculated in any such B. If C and D are finite orthogonal subsets of L, then $\oplus C \perp \oplus D$ if and only if $C \cap D \subseteq \{0\}$ and $C \cup D$ is an orthogonal set, in which case $\oplus C \oplus \oplus D = \oplus(C \cup D)$. If $A = \{a_1, a_2\}$, then $\oplus A = a_1 \oplus a_2$. Thus, if $A = \{a_1, a_2, a_3, \ldots, a_n\}$, is an orthogonal set, we can define

$$a_1 \oplus a_2 \oplus a_3 \oplus \cdots \oplus a_n = \oplus A$$

without notational conflict.

If $p, q \in L$ and both p and q belong to a Boolean subalgebra B of L, then the greatest lower bound $p \wedge_B q$ and the least upper bound $p \vee_B q$ of p and q as calculated in B may well depend on the choice of B. If $p \wedge_B q$ is independent of the choice of B, we define the *conjunction* $p \,\&\, q$ of p and q by

$$p \,\&\, q = p \wedge_B q.$$

Likewise, if $p \vee_B q$ is independent of the choice of B, we define the *disjunction* $p + q$ of p and q by

$$p + q = p \vee_B q.$$

It can be shown that the compatible elements p and q have a conjunction if and only if they have a disjunction. Furthermore, if $p \,\&\, q$ and $p + q$ exist, then $p \,\&\, q$ is a maximal lower bound and $p + q$ is a minimal upper bound for p and q in L. If p and q are compatible and at least one of $p \wedge q$ or $p \vee q$ exists in L, then $p \,\&\, q$ and $p + q$ exist, $p \,\&\, q = p \wedge q$, and $p + q = p \vee q$. If $p \,\&\, q$ exists, then so do $p' \,\&\, q'$ and $p' + q'$, and we have $p + q = (p' \,\&\, q')'$ and $p \,\&\, q = (p' + q')'$. If $p \perp q$, then $p + q = p \oplus q$ and $p \,\&\, q = 0$.

If L is an orthomodular poset, then any two compatible elements $p, q \in L$ have a conjunction $p \,\&\, q = p \wedge q$ and a disjunction $p + q = p \vee q$; however, there are orthoalgebras containing compatible pairs of elements that do not admit conjunctions or disjunctions. There are non-Boolean orthoalgebras in which every pair of elements forms a compatible set. There exist orthomodular posets containing three elements that are pairwise compatible, but that do not form a compatible set; however, in an orthomodular poset, every pairwise orthogonal subset is an orthogonal subset. There exist orthoalgebras containing three elements that are pairwise orthogonal, but do not form an orthogonal set.

3.3 Probability Measures on and Supports in an Orthoalgebra

By a *probability measure* on an orthoalgebra L, we mean a mapping $\omega : L \to [0, 1] \subseteq \mathbb{R}$ such that, for $p, q \in L$, with $p \perp q$,

$$\omega(p \oplus q) = \omega(p) + \omega(q). \tag{9}$$

It is possible to define *σ-complete* orthoalgebras and *countably additive* probability measures thereon, and thus extend Eq. (9) to sequences in L, but we do not do so here. The set of all probability measures on L is denoted by $\Omega = \Omega(L)$. Evidently, Ω is a convex subset of the vector space of all real-valued functions on L.

If the elements of L are regarded as representing two-valued experimental propositions concerning a physical system

$\mathscr{S}$, then a probability measure $\omega \in \Omega$ may be interpreted in any of the following ways:

(*Frequency*) ω is a *complete stochastic model* for $\mathscr{S}$ in the sense that $\omega(p)$ is the "longrun relative frequency" with which the proposition $p \in L$ will be true when repeatedly tested (D'Espagnat, 1971).

(*Subjective*) ω is a *model for coherent belief* encoding all of our current information about the system $\mathscr{S}$. Thus, if $p \in L$, then $\omega(p)$ measures our current "degree of belief," on a scale from 0 to 1, in the truth of the proposition p (Jaynes, 1989).

(*Propensity*) $\omega(p)$ is a measure on a scale from 0 to 1 of the "propensity" of the system $\mathscr{S}$ to produce the outcome 1(= true) when the proposition p is tested (Popper, 1959).

(*Mathematical*) ω is a mathematical artifact that may be of use in making inferences about $\mathscr{S}$ using data secured by making measurements on $\mathscr{S}$ (Kolmogorov, 1933).

For $\omega \in \Omega$, we define the *support* of ω by

$$\mathrm{supp}(\omega) = \{p \in L \mid \omega(p) > 0\}.$$

If $S = \mathrm{supp}(\omega)$, then $1 \in S$ and, for all $p, q \in L$ with $p \perp q$,

$$p \oplus q \in S \quad \text{if and only if } p \in S \text{ or } q \in S. \tag{10}$$

A subset S of L such that Eq. (10) holds is called a *support* in L. In general, there are supports $S \subseteq L$ that are not of the form $\mathrm{supp}(\omega)$ for $\omega \in \Omega$; those that are of this form are called *stochastic supports*. The set-theoretic union of supports is again a support, and it follows that the collection of all supports in L forms a complete lattice under set-theoretic inclusion.

3.4 Cartesian and Tensor Products of Orthoalgebras

If L_1 and L_2 are orthoalgebras, the Cartesian product $L_1 \times L_2$ becomes an orthoalgebra under the obvious componentwise operations. If L_1 is identified with $L_1 \times \{0\}$ and L_2 is identified with $\{0\} \times L_2$ in $L_1 \times L_2$, then every element in $L_1 \times L_2$ can be written uniquely in the form $p \oplus q$ with $p \in L_1$ and $q \in L_2$. This construction generalizes the superselected direct sum of standard quantum logics. Just as is the case for standard quantum logics, $\Omega(L_1 \times L_2)$ is isomorphic in a natural way to the convex hull of $\Omega(L_1)$ and $\Omega(L_2)$.

A construction for the tensor product $L_1 \otimes L_2$ of orthoalgebras based on Foulis and Randall (1981) can be found in Lock (1981). The factors L_1 and L_2 are embedded in the orthoalgebra $L_1 \otimes L_2$ by mappings $p \to p \otimes 1$ and $q \to 1 \otimes q$ for $p \in L_1, q \in L_2$ in such a way that $p \otimes 1$ and $1 \otimes q$ are compatible and have a conjunction $(p \otimes 1)\&(1 \otimes q) = p \otimes q$. Furthermore, elements of the form $p \otimes q$ generate $L_1 \otimes L_2$. If $\alpha \in \Omega(L_1)$ and $\beta \in \Omega(L_2)$, there is a unique $\gamma = \alpha\beta \in \Omega(L_1 \otimes L_2)$ such that $\gamma(p \otimes q) = \alpha(p)\beta(q)$ for all $p \in L_1, q \in L_2$. A probability measure on $L_1 \otimes L_2$ of the form $\alpha\beta$ is said to be *factorizable*, and a convex combination of factorizable probability measures is said to be *separable* (Kläy, 1988). The existence of probability measures on $L_1 \otimes L_2$ that are not separable seems to be a characteristic feature of the tensor product of non-Boolean orthoalgebras.

3.5 The Logic of a Physical System

We are now in a position to summarize the *quantum logic approach* to the study of *physical systems* (quantum-mechanical or

not). The *basic postulate of quantum logic* for a physical system $\mathscr{S}$ is as follows:

> *(Logic postulate) The set* L *of all two-valued, experimentally testable propositions for* $\mathscr{S}$ *has the structure of an orthoalgebra such that every simultaneously testable set of propositions forms a compatible subset of* L *and every finite compatible subset of* L *is a simultaneously testable set of propositions. If* $\mathrm{p}, \mathrm{q} \in \mathrm{L}$ *with* $\mathrm{p} \perp \mathrm{q}$, *and if* p, q, *and* $\mathrm{p} \oplus \mathrm{q}$ *are tested simultaneously, then at most one of the propositions* p, q *will be true, and* $\mathrm{p} \oplus \mathrm{q}$ *will be true if and only if either* p *or* q *is true.*

We refer to L as the *logic* of the system $\mathscr{S}$.

It is customary to assume that there is a *state space* Ψ associated with the physical system $\mathscr{S}$. The elements $\psi \in \Psi$ are called *states*, and, at any given moment, $\mathscr{S}$ is presumed to be *in* one and only one state $\psi \in \Psi$. A state is supposed to encode all available information about the consequences of performing tests or making measurements on $\mathscr{S}$ when $\mathscr{S}$ is in that state.

Whereas the truth or falsity of an experimental proposition $p \in L$ can be determined by a suitable test, it may or may not be possible to determine the current state $\psi \in \Psi$ of $\mathscr{S}$ by a test or measurement; however, it may be possible to bring $\mathscr{S}$ into a state ψ by means of a suitable *state-preparation procedure*. The state of the system $\mathscr{S}$ can change under the action of a *dynamical law*, under a *state collapse* when an observer tests a proposition or measures an observable, because a *state-preparation procedure* is executed, or simply by virtue of a spontaneous *state transition*.

A connection between the state space Ψ for $\mathscr{S}$ and its logic L is effected as follows:

> *(Stochastic postulate) Each state* $\psi \in \Psi$ *determines a corresponding probability measure* ω_ψ *on* L *in such a way that, for* $\mathrm{p} \in \mathrm{L}$, $\omega_\psi(\mathrm{p})$ *is the probability that the proposition* p *is true when tested with the system* $\mathscr{S}$ *in the state* ψ.

In the stochastic postulate, the probability measure ω_ψ can be interpreted in any of the four ways (frequency, subjective, propensity, mathematical) suggested in Sec. 3.3. In what follows, we denote by Σ the subset of $\Omega(L)$ consisting of all probability measures of the form ω_ψ for $\psi \in \Psi$, and we refer to each $\omega_\psi \in \Sigma$ as a *probability state* for the system $\mathscr{S}$. It is customary to identify the state ψ with the corresponding probability state ω_ψ and to speak of the elements in Σ as *states* for $\mathscr{S}$. Although this custom can lead to philosophical and mathematical difficulties (what if $\phi, \psi \in \Psi, \phi \neq \psi$, and yet $\omega_\phi = \omega_\psi$?), we shall follow it in the interests of simplicity.

Let $\omega \in \Sigma$ be a state and let $p \in L$ be an experimental proposition for the physical system $\mathscr{S}$. We say that p is *possible, impossible,* or *certain* in the state ω if $p \in \mathrm{supp}(\omega)$, $p \notin \mathrm{supp}(\omega)$, or $p' \notin \mathrm{supp}(\omega)$, respectively. If both p and p' are possible, we say that p is *contingent* in the state ω. The state space Σ is said to be *unital* if every nonzero $p \in L$ is certain in at least one state $\omega \in \Sigma$.

If $\Lambda \subseteq \Sigma$ is a set of states, then a state $\omega \in \Sigma$ is said to be a *superposition* of the states in Λ if

$$\mathrm{supp}(\omega) \subseteq \cup\{\mathrm{supp}(\lambda) | \lambda \in \Lambda\}.$$

The *superposition closure* of Λ is defined to be the set Λ^{sp} of all superpositions of states in Λ. If $\Lambda = \Lambda^{sp}$, then Λ is called *superposition closed*. A state ω is *pure* if the set $\{\omega\}$ is superposition closed. If ω is a pure state, Λ is a set of pure states, and $\omega \in \Lambda^{sp}$, then ω is a *coherent superposition* of the states in Λ. In what follows, we denote by $\mathscr{L}$ the set of

all superposition-closed subsets of Σ. Note that $\mathscr{L}$ is closed under set-theoretic intersection, and hence, it forms a complete lattice under set-theoretic inclusion.

It has long been a tenet of natural philosophy that affiliated with a physical system $\mathscr{S}$ is a class $\mathscr{A}$ of *attributes* or *properties*. At any given moment, some of these attributes may be *actual*, while the others are only *potential*. The attributes of $\mathscr{S}$ that are always actual are its *intrinsic* attributes; those that can be either actual or potential are its *accidental* attributes. The charge of an electron is one of its intrinsic attributes, whereas the attribute "spin up in the z direction" is accidental.

To each attribute $A \in \mathscr{A}$ there corresponds a set $\Lambda_A \subseteq \Sigma$ consisting precisely of those states ω such that A is actual whenever $\mathscr{S}$ is in the state ω. A heuristic argument, which we omit here, indicates that Λ_A should be superposition closed, so that $\Lambda_A \in \mathscr{L}$. Similar arguments suggest that every element of $\mathscr{L}$ corresponds in this way to an attribute, and thus lead us to our third postulate:

> *(Attribute postulate) Each attribute* A *determines a corresponding superposition-closed subset* Λ_{A} *of the state space* Σ *such that* A *is actual if and only if the system* $\mathscr{S}$ *is in a state* $\omega \in \Lambda_{\mathrm{A}}$; *furthermore, every* $\Lambda \in \mathscr{L}$ *has the form* Λ_{A} *for some* $\mathrm{A} \in \mathscr{A}$.

Just as we identified states with probability states, we propose to identify elements Λ of the complete lattice $\mathscr{L}$ with attributes of the system $\mathscr{S}$. (Note that, as a perhaps undesirable consequence, all of the intrinsic attributes of $\mathscr{S}$ become identified with the superposition-closed subset Σ itself.) Thus, we shall refer to the complete lattice $\mathscr{L}$ as the *attribute lattice* for the system $\mathscr{S}$.

If $\Lambda, \Gamma \in \mathscr{L}$ are attributes of $\mathscr{S}$, then $\Lambda \subseteq \Gamma$ if and only if Γ is actual whenever Λ is actual. Furthermore, the attribute $\Lambda \cap \Gamma \in \mathscr{L}$ corresponds to a bona fide *conjunction* of the attributes Λ and Γ in the sense that $\Lambda \cap \Gamma$ is actual if and only if both Λ and Γ are actual. However, the least upper bound of Λ and Γ in $\mathscr{L}$ is $(\Lambda \cup \Gamma)^{sp}$, and it can be actual in states in which neither Λ nor Γ is actual. Following Aerts (1982), we say that the attributes Λ and Γ are *separated by a superselection rule* if $\Lambda \cup \Gamma$ is superposition closed, so that the least upper bound of Λ and Γ in $\mathscr{L}$ corresponds to a bona fide *disjunction* of the attributes Λ and Γ.

3.6 The Canonical Mapping

We continue our discussion of the physical system $\mathscr{S}$ subject to the logic, stochastic, and attribute postulates of Sec. 3.5.

Von Neumann (1955, p. 249) writes,

> Apart from the physical quantities . . . , there exists another category of concepts that are important objects of physics – namely the properties of the states of the system $\mathscr{S}$.

Furthermore, he goes on to identify these properties (or attributes) with the projections in the standard quantum logic L affiliated with the quantum-mechanical system $\mathscr{S}$.

In the more general situation under discussion, it is also possible to relate propositions $p \in L$ and properties (i.e., attributes) $\Lambda \in \mathscr{L}$. For $p \in L$, define

$$[p] = \{\omega \in \Omega \mid \omega(p) = 1\}.$$

We claim that $[p]$ is superposition closed. Indeed, suppose that $\alpha \in [p]^{sp}$, but that $\alpha \notin [p]$. Then $\alpha(p) \neq 1$, and so $\alpha(p') > 0$, $p' \in \mathrm{supp}(\alpha)$, and so $p' \in \mathrm{supp}(\omega)$ for some $\omega \in [p]$. But, then, $\omega(p') > 0$, and so $\omega(p) < 1$, contradicting $\omega \in [p]$. Thus $p \to [p]$

provides a mapping from experimental propositions $p \in L$ to attributes $[p] \in \mathscr{L}$. We refer to $p \to [p]$ as the *canonical mapping* (Foulis et al., 1983).

An attribute of the form $[p]$ is called a *principal attribute*; the principal attributes are those that can be identified with experimental propositions as von Neumann did. It is not difficult to show that every attribute is an intersection (i.e., a conjunction) of (possibly infinitely many) principal attributes. The state space Σ is unital if and only if $[p] = 0$ implies that $p = 0$.

Evidently, $p, q \in L$ with $p \leq q$ implies that $[p] \subseteq [q]$. If the converse holds, so that $[p] \subseteq [q]$ implies that $p \leq q$, then $\mathscr{S}$ is said to have a *full* set of states. If $\mathscr{S}$ has a full set of states and every attribute is principal, then the logic L is isomorphic to the attribute lattice $\mathscr{L}$ – this is precisely what happens for a standard quantum logic and it accounts for von Neumann's identification of projections and properties.

3.7
Critique of Quantum Logic

Quantum logic is a relatively young subject, it is still under vigorous development, and many consequences of the epistemological and mathematical insights that it has already provided have yet to be exploited. Quantum-logical techniques involving the tensor product have already cast some light on the well-known Einstein–Podolsky–Rosen paradox (Kläy, 1988), and it is hoped that they will also clarify some of the other classical paradoxes (Wigner's friend, Schrödinger's cat, etc.). The problem of hidden variable can be formulated, understood, and studied rigorously in terms of quantum logics (Greechie and Gudder, 1973). Quantum-logical techniques have enhanced our understanding of group-theoretic imprimitivity methods and the role of superselection rules (Piron, 1976), and ideas related to quantum logic have been used to help unravel the measurement problem (Busch et al., 1991)

There is a strong possibility that unrestricted orthoalgebras are too general to serve as viable models for quantum logic. Some orthoalgebras are extremely "pathological" and thus may be suitable only for the construction of counterexamples. It seems likely that only an appropriately specialized class of orthoalgebras, e.g., unital orthoalgebras, might prove to be adequate as models for a general logic of experimental propositions.

The main drawback of quantum logic is already evident in the standard quantum logic L of a Hilbert space $\mathscr{H}$: In the passage from the wave functions ψ in $\mathscr{H}$ to the projections $P \in L$, all phase information is lost. The lost information becomes critical when sequential measurements – e.g., iterated Stern–Gehrlach spin resolutions (Wright, 1978) – are to be performed. There are at least two ways to restore the lost information, both of which are currently being studied. One can introduce complex-valued *amplitude functions* on the logic L (Gudder, 1988), or one can introduce a general mathematical infrastructure called a *manual* or *test space* (Randall and Foulis, 1973; Foulis, 1989) that can carry phase information and that gives rise to orthoalgebras as derived structures in much the same way that Hilbert spaces give rise to the standard quantum logics.

4
The Logician's Approach

We now present an approach to quantum logic more closely aligned with that

of standard logical techniques. In the preceding section, we gave an axiomatic approach to orthoalgebras, the most general mathematical structures currently used as models for quantum logic. This section deals with quantum logics by using the methods of logical tradition. In so doing, we will speak of *abstract quantum logics*. As we have seen, *standard quantum logic* is identified with the complete orthomodular lattice of the projections on a separable Hilbert space of dimension at least three over the complex number field. Thus standard quantum logic is a particular kind of semantic model for a form of abstract quantum logic. Generally, a logic **L** can be determined as a triple $\langle FL, \vdash, \models \rangle$, consisting of a formal language *FL*, a *proof-theoretic consequence* relation, and a *semantic*-(or *model-theoretic*) *consequence* relation. For the sake of simplicity, we will consider only sentential languages, generated by an alphabet containing

1. a denumerably infinite sequence of *atomic sentences* (i.e., sentences whose proper parts are not sentences),
2. a finite sequence of primitive logical connectives.

The set of the *sentences* of the language *FL* is the smallest set that contains the atomic sentences and is closed under the logical connectives.

The proof-theoretic concept of consequence $\vdash$ for **L** is defined by referring to a *calculus* (a set of *axioms* and of *rules*) that, in turn, determines a notion of *proof from a set of premises to a conclusion*. A sentence β is called a *proof-theoretic consequence* of a sentence $\alpha (\alpha \vdash \beta)$ if and only if (hereafter abbreviated as iff) there is a proof where ω is the premise and β the conclusion. The semantic-consequence relation $\models$ refers to a class of possible interpretations (*models*) of the language, which render any sentence "more or less" *true* or *false*. A sentence β is called a *semantic consequence* of $\alpha (\alpha \models \beta)$ iff in any possible model of the language, β is at least as true as α.

The two consequence relations $\vdash$ and $\models$ are *reciprocally adequate* iff they are equivalent. In other words: for any sentences α, β:

$$\alpha \vdash \beta \quad \text{iff} \quad \alpha \models \beta.$$

The "if arrow" represents the *soundness* property of the logic, whereas the "only if arrow" is the semantic *completeness* property.

Naturally, a logic can be characterized by different consequence relations that turn out to be equivalent. A logic **L** is called *axiomatizable* iff it admits a proof-theoretic relation, where the notion of proof is decidable. Further, **L** is called *decidable* iff the proof-theoretic-consequence relation $\vdash$ is decidable.

5
Algebraic and Possible-World Semantics

In the logical tradition, logics can be generally characterized by means of two privileged kinds of semantics: an algebraic semantics, or a possible-world semantics (called also Kripkean semantics).

These semantics give different answers to the question: What does it mean to interpret a formal language? In the algebraic semantics, the basic idea is that interpreting a language essentially means associating to any sentence an abstract *truth value* or, more generally, an *abstract meaning:* an element of an algebraic structure. Hence, generally, an algebraic model for a logic **L** will have the form

$$\mathfrak{M} = \langle \mathscr{A}, v \rangle,$$

where $\mathscr{A}$ is an algebraic structure belonging to a class $\mathfrak{M}$ of structures satisfying a given set of conditions and v transforms sentences into elements of $\mathscr{A}$, preserving the logical form (in other words, logical constants are interpreted as operations of the structure). We will consider only structures where a binary relation $\leq$ (possibly a partial order) is defined. On this basis, the semantic-consequence relation is defined as follows:

DEFINITION 5.1 – β is a *semantic consequence* of $\alpha(\alpha \models \beta)$ iff for any model $\mathfrak{M} = \langle \mathscr{A}, v \rangle$, $v(\alpha) \leq v(\beta)$ (in other words, the abstract meaning of α precedes the abstract meaning of β).

In the possible-world semantics, instead, one assumes that interpreting a language essentially means associating to any sentence α the set of the *possible worlds* (or *situations*) where α holds: This set, that represents the *extensional meaning* of α, is called the *proposition associated to* α (simply, the *proposition of* α). Hence, generally, a Kripkean model for a logic **L** will have the form:

$$\mathfrak{M} = \langle I, \mathbf{R}_i, \mathbf{o}_j, \Pi, v \rangle,$$

where the meanings of the symbols are as follows.

1. I is a nonempty set of possible worlds possibly correlated by relations in the sequence $\mathbf{R}_i$ and operations in the sequence $\mathbf{o}_j$. In most cases, we have only one relation R, called the *accessibility* relation.
2. Π is a set of sets of possible worlds, representing possible propositions of sentences. Any proposition and the total set of propositions Π must satisfy convenient closure conditions that depend on the particular logic.
3. v transforms sentences into propositions preserving the logical form.

A world i is said to *verify* a sentence $\alpha(i \models \alpha)$ iff $i \in v(\alpha)$.

On this basis, the Kripkean semantic-consequence relation is defined as follows:

DEFINITION 5.2 – β is a *semantic consequence* of $\alpha(\alpha \models \beta)$ iff for any model $\mathfrak{M} = \langle I, \mathbf{R}_i, \mathbf{o}_j, \Pi, v \rangle$ and for any world $i \in I$,

$$\text{if } i \models \alpha \quad \text{then } i \models \beta$$

(in other words: whenever α is verified, also β is verified).

In both semantics, a sentence α is called a *logical truth* ($\models \alpha$) iff α is the consequence of any sentence β.

An interesting variant of Kripkean semantics is represented by the *many-valued possible-world semantics*, founded on a generalization of the notion of proposition. As we have seen, in the standard possible-world semantics, the proposition of a sentence α is a set of worlds: the worlds where α holds. This automatically determines the set of the worlds where α does not hold (the "meaning" of the negation of α). Intermediate truth values are not considered. In the many-valued possible-world semantics, instead, one fixes, at the very beginning, a set of truth values $V \subseteq [0, 1]$ and any proposition is represented as a function X that associates to any truth value $r \in [0, 1]$ a convenient set of possible worlds (the worlds where our proposition holds with truth value r). As a consequence, the total set of propositions Π turns out to behave like a family of *fuzzy* subsets of I.

Classical logic (CL) can be characterized both in the algebraic and in the Kripkean semantics. Algebraically, it is

determined by the class of all algebraic structures $\langle \mathscr{A}, v \rangle$, where $\mathscr{A}$ is a Boolean algebra and v interprets the classical connectives (negation, conjunction, disjunction) as the corresponding Boolean operations (complement, meet, join). In the framework of Kripkean semantics, instead, CL is characterized by the class of all models $\langle I, R, \Pi, v \rangle$, where

1. the accessibility relation R is the identity relation (in other words, any world is accessible only to itself);
2. the set of the possible propositions Π is the set of all subsets of I;
3. v interprets the classical connectives as the corresponding set-theoretic operations.

6 Orthodox Quantum Logic

In the abstract quantum-logical universe, a privileged element is represented by *orthodox quantum logic* (QL), first described "as a logic" by Birkhoff and von Neumann (Birkhoff and von Neumann, 1936). QL is a singular point in the class of all logics that are weaker than classical logic. Many logical and metalogical problems concerning QL have been solved. However, some questions seem to be stubbornly resistant to being resolved.

6.1 Semantic Characterizations of QL

Similarly to classical logic, QL can be characterized both in the algebraic and in the Kripkean semantics. The language of QL contains the two primitive connectives $\neg$ (not), $\owedge$ (and). Disjunction is supposed to be metalinguistically defined via De Morgan's law:

$$\alpha \owvee \beta := \neg(\neg\alpha \owedge \neg\beta).$$

A conditional connective can be defined as the "Sasaki hook":

$$\alpha \rightarrow \beta := \neg\alpha \owvee (\alpha \owedge \beta).$$

DEFINITION 6.1.1 – An *algebraic model* of QL is a pair $\mathfrak{M} = \langle \mathscr{A}, v \rangle$, where

1. $\mathscr{A} = \langle A, \leq, ', \mathbf{1}, \mathbf{0} \rangle$ is an orthomodular lattice;
2. v (the interpretation function) interprets the connective $\neg$ as the operation $'$, the connective $\owedge$ as the lattice-meet $\wedge$:
   a. $v(\alpha) \in A$ for any atomic sentence α.
   b. $v(\neg\beta) = v(\beta)'$.
   c. $v(\beta \owedge \gamma) = v(\beta) \wedge v(\gamma)$.

DEFINITION 6.1.2 – A sentence α is called *true* in a model $\langle \mathscr{A}, v \rangle$ iff $v(\alpha) = 1$. Accordingly, we will have that β is a *consequence of* α *in the algebraic semantics* of QL ($\alpha \models^{A}_{\mathrm{QL}} \beta$) iff $v(\alpha) \leq v(\beta)$ in any model $\langle \mathscr{A}, v \rangle$ based on an orthomodular lattice $\mathscr{A}$. Further, α is a *quantum-logical truth in the algebraic semantics* ($\models^{A}_{\mathrm{QL}} \alpha$) iff α is true in any algebraic model of QL.

As a consequence of the orthomodular property, a semantic version of a "deduction lemma" can be proved:

LEMMA 6.1.1 – $\alpha \models^{A}_{\mathrm{QL}} \beta$ iff $\models^{A}_{\mathrm{QL}} \alpha \rightarrow \beta$. In other words, $\rightarrow$ represents a "good" conditional connective: $\alpha \rightarrow \beta$ is logically true iff β is a consequence of α.

DEFINITION 6.1.3 – A *Kripkean model* of QL has the form $\mathfrak{M} = \langle I, R, \Pi, v \rangle$, where the following conditions held:

1. The accessibility relation R is reflexive and symmetric [we will also write

$i \not\perp j$ for Rij; and $i \perp j$ for not Rij. Moreover, if $X \subseteq I$, we will write $i \perp X$ for $\forall j \in X(i \perp j)$].

A *possible proposition* of $\mathfrak{M}$ is a *maximal* set X of worlds, which contains all and only those worlds whose accessible worlds are accessible to at least one element of X. In other words, $i \in X$ iff $\forall j \not\perp i, \exists k \not\perp j$ with $k \in X$.

For any $X \subseteq I$, let $X^{\textcircled{\perp}} := \{i \in I | i \perp X\}$. One can prove that $X^{\textcircled{\perp}}$ is a possible proposition for any $X \subseteq I$; X is a possible proposition iff $X = X^{\textcircled{\perp}}$ $\textcircled{\perp}$; $\emptyset$ and I are possible propositions; if X, Y are possible propositions, then $X \cap Y$ is a possible proposition.

2. Π is a set of possible propositions closed under, I, $\textcircled{\perp}$, $\cap$.
3. Π is orthomodular: $X \cap (X \cap (X \cap Y)^{\textcircled{\perp}})^{\textcircled{\perp}} \subseteq Y$, for any $X, Y \in \Pi$.
4. **a**. $v(\alpha) \in \Pi$, for any atomic sentence α;
   b. $v(\neg\beta) = v(\beta)^{\textcircled{\perp}}$;
   c. $v(\beta \textcircled{\wedge} \gamma) = v(\beta) \cap v(\gamma)$.

DEFINITION 6.1.4 – A sentence α is called *true* in a model $\mathfrak{M} = \langle I, R, \Pi, v\rangle$ iff α is verified by any world $i \in I$.

Accordingly, we will have that β is a *consequence of* α *in the Kripkean semantics* of QL ($\alpha \models^K_{\mathrm{QL}} \beta$) iff for any Kripkean model $\mathfrak{M} = \langle I, R, \Pi, v\rangle$ of QL and for any world i, if $i \models \alpha$ then $i \models \beta$. Further, α is a *quantum-logical truth in the Kripkean semantics* for QL ($\models^K_{\mathrm{QL}} \alpha$) iff α is true in any Kripkean model of QL.

The algebraic and the Kripkean semantics for QL turn out to characterize the same logic:

THEOREM 6.1.1 – $\alpha \models^A_{\mathrm{QL}} \beta$ *iff* $\alpha \models^K_{\mathrm{QL}} \beta$. This permits us to write simply $\alpha \models_{\mathrm{QL}} \beta$ instead of $\alpha \models^A_{\mathrm{QL}} \beta$ and $\alpha \models^K_{\mathrm{QL}} \beta$.

Both the algebraic and the Kripkean models of QL admit of Hilbert-space exemplifications, which are the basis for the physical interpretations. Let $\mathscr{H}$ be the separable complex Hilbert space associated to a physical system $\mathscr{S}$. An algebraic model $\langle \mathscr{A}, v\rangle$ can be constructed by taking as $\mathscr{A}$ the standard quantum logic based on $\mathscr{H}$ – in other words, the orthomodular lattice of the projections on $\mathscr{H}$; whereas v will follow the intended physical meaning of the atomic sentences. At the same time, a Kripkean model $\mathfrak{M} = \langle I, R, \Pi, v\rangle$ can be constructed by putting $I =$ the set of the pure states (represented by normalized vectors ψ of $\mathscr{H}$), $R =$ the non-orthogonality relation between pure states, and $\Pi =$ the set of the possible propositions, which are uniquely determined by the closed subspaces of $\mathscr{H}$. v will follow the physical meaning of the atomic sentences. It turns out that the propositions of the model correspond to superposition-closed subsets of the pure-state space. We will call this kind of models *Hilbertian models* of QL.

A question arises: Is QL characterized by the class of all algebraic Hilbertian models? The answer is negative as proved by Greechie (1981). For instance, there is a complicated sentence of QL (corresponding to the so-called *orthoarguesian law*) that is true in all Hilbertian models, and not true in some QL models. Let us call *Hilbertian quantum logic* (HQL) the logic that is semantically characterized by the class of all Hilbertian models. Apparently, HQL is stronger than QL. Hence, abstract quantum logic turns out to be definitely more general with respect to its physical and historical origin. The axiomatizability of HQL is still an open problem.

DEFINITION 6.1.5 – A sentence α is called *semantically consistent* iff for any β, $\alpha \not\models_{\mathrm{QL}} \beta \textcircled{\wedge} \neg\beta$ (in other words, no contradiction is a semantic consequence of α).

One can show that α is semantically consistent iff there is at least one algebraic model $\langle \mathscr{A}, \nu \rangle$ such that $\nu(\alpha) \neq \mathbf{0}$ iff there exists at least one Kripkean model $\mathfrak{M} = \langle I, R, \Pi, \nu \rangle$ and at least one world i such that $i \models \alpha$.

6.2 An Axiomatization of QL

QL is an axiomatizable logic. Many axiomatizations are known: in the Hilbert–Bernays style (Hardegree, 1979), in the natural deduction, and in the sequent style (Gibbins, 1985; Nishimura, 1980). We will present here a calculus (Goldblatt, 1974; Dalla Chiara, 1986) that represents a kind of "logical copy" of orthomodular lattices. Our calculus (that has no axioms) is determined as a set of rules. Any rule has the form

$$\frac{\alpha_1 \vdash \beta_1, \ldots, \alpha_n \vdash \beta_n}{\alpha \vdash \beta}.$$

(If β_1 is inferred from $\alpha_1, \ldots, \beta_n$ is inferred from α_n, then β can be inferred from α.) The configurations $\alpha_1 \vdash \beta_1, \ldots, \alpha_n \vdash \beta_n$ represent the *premises* of the rule, while $\alpha \vdash \beta$ is the *conclusion*. An *improper rule* is a rule whose set of premises is empty. Instead of

$$\frac{\emptyset}{\alpha \vdash \beta},$$

we will write $\alpha \vdash \beta$.

The rules of QL are as follows:

R1 $\alpha \vdash \alpha$ (identity).
R2 $\dfrac{\alpha \vdash \beta \ \beta \vdash \gamma}{\alpha \vdash \gamma}$ (transitivity).
R3 $\alpha \barwedge \beta \vdash \alpha$.
R4 $\alpha \barwedge \beta \vdash \beta$.
R5 $\dfrac{\gamma \vdash \alpha \ \gamma \vdash \beta}{\gamma \vdash \alpha \barwedge \beta}$.
R6 $\alpha \vdash \neg\neg\alpha$ (weak double negation).
R7 $\neg\neg\alpha \vdash \alpha$ (strong double negation).
R8 $\dfrac{\alpha \vdash \beta}{\neg\beta \vdash \neg\alpha}$ (contraposition).
R9 $\alpha \barwedge \neg(\alpha \barwedge \neg(\alpha \barwedge \beta)) \vdash \beta$ (orthomodularity).

DEFINITION 6.2.1 – A *proof* is a finite sequence of configurations $\alpha \vdash \beta$ where any element of the sequence is either an improper rule or the conclusion of a proper rule whose premises are previous elements of the sequence.

DEFINITION 6.2.2 – β is a *proof-theoretic consequence* of α (or *provable* from α) ($\alpha \vdash_{\mathrm{QL}} \beta$) iff there is a proof whose last configuration is $\alpha \vdash \beta$.

DEFINITION 6.2.3 – β is a *proof-theoretic consequence* of a set of sentences T (or *provable* from T) ($T \vdash_{\mathrm{QL}} \beta$) iff T includes a finite subset $\{\alpha_1, \ldots, \alpha_n\}$ such that $\alpha_1, \barwedge \cdots \barwedge \alpha_n \vdash_{\mathrm{QL}} \beta$.

DEFINITION 6.2.4 – A set of sentences T is called *contradictory* if $T \vdash_{\mathrm{QL}} \beta \barwedge \neg\beta$ for some sentence β; *noncontradictory*, otherwise. A sentence α is contradictory if $\{\alpha\}$ is contradictory; noncontradictory, otherwise.

The proof-theoretic and the semantic-consequence relations turn out to be equivalent. Namely, a soundness and a completeness theorem can be proved:

THEOREM 6.2.1 – *Soundness*

$$\text{If } \alpha \vdash_{\mathrm{QL}} \beta \quad \text{then } \alpha \models_{\mathrm{QL}} \beta.$$

THEOREM 6.2.2 – *Completeness*

$$\text{If } \alpha \models_{\mathrm{QL}} \beta \quad \text{then } \alpha \vdash_{\mathrm{QL}} \beta.$$

As a consequence, one obtains the result that a sentence is noncontradictory iff it is semantically consistent.

A characteristic "anomaly" of QL is the violation of a metalogical condition, which is satisfied not only by CL but also by a large class of nonclassical logics. This condition is represented by the *Lindenbaum property*, according to which any noncontradictory set of sentences T can be extended to a noncontradictory and *complete* set T^* such that for any sentence α, either $\alpha \in T^*$ or $\neg\alpha \in T^*$. The set $T := \{\neg(\alpha \to (\beta \to \alpha))\}$ (which contains the negation of the *a fortiori* principle) represents an example of a noncontradictory set that cannot be extended to a noncontradictory and complete set. The set T is noncontradictory, because in some models $\langle \mathscr{A}, v\rangle$: $v(\neg(\alpha \to (\beta \to \alpha))) \neq \mathbf{0}$. For instance, take $\langle \mathscr{A}, v\rangle$ based on the orthomodular lattice of the closed subspaces of $\mathbb{R}^2$, where $v(\alpha)$ and $v(\beta)$ are two nonorthogonal unidimensional subspaces. However, one can easily check that $v(\neg(\alpha \to (\beta \to \alpha))) = \mathbf{1}$ is impossible. Hence, $\neg(\alpha \to (\beta \to \alpha))$ cannot belong to a noncontradictory and complete set T^*, which would trivially admit a model $\langle \mathscr{A}, v\rangle$ such that $v(\beta) = \mathbf{1}$, for any $\beta \in T^*$. From an intuitive point of view, the failure of the Lindenbaum property represents a very strong incompleteness result. The *tertium non datur* principle breaks down at the very deep level: There are theories that are intrinsically incomplete, even *in mente Dei*.

Among the questions that are still unsolved, let us mention at least the following:

1. Is QL decidable?
2. Does QL admit the *finite-model property*? In other words, if a sentence is not a quantum-logical truth, is there any finite model where our sentence is not verified? A positive answer to the finite-model property would automatically provide a positive answer to the decidability question, but not vice versa.
3. Is the set of all possible propositions in the Kripkean *canonical* model of QL orthomodular? (The worlds of the canonical model are all the noncontradictory and deductively closed sets of sentences T, whereas two worlds T and T' are accessible iff whenever T contains a sentence α, T' does not contain its negation $\neg\alpha$.) This problem is correlated to the critical question whether any orthomodular lattice is embeddable into a complete orthomodular lattice. Only partial answers are known.

7
Orthologic and Unsharp Quantum Logics

By dropping the orthomodular condition both in the algebraic and in the Kripkean semantics, one can characterize a weaker form of quantum logic, which is usually called *orthologic* or *minimal quantum logic* (MQL). This logic turns out to be more "tractable" from a metalogical point of view: It satisfies the finite-model property; consequently, it is decidable (Goldblatt, 1974). A calculus that represents an adequate axiomatization for MQL can be, naturally, obtained by replacing the orthomodular rule **R9** of our QL calculus with the weaker Duns Scotus rule $\alpha \barwedge \neg\alpha \vdash \beta$ (*ex absurdo sequitur quodlibet*: Any sentence is a consequence of a contradiction).

A less investigated form of quantum logic is represented by *paraconsistent quantum logic* (PQL) (Dalla Chiara and Giuntini, 1989), which is a weak example of an unsharp quantum logic, possibly violating the noncontradiction and the excluded-middle principles. As we will see, unsharp quantum logics represent natural

abstractions from the *unsharp approaches* to quantum theory. Algebraically, PQL is characterized by the class of all models based on an involutive lattice $\langle A, \leq, ', \mathbf{1}, \mathbf{0}\rangle$, with smallest element **0** and largest element **1**. Equivalently, in the Kripkean semantics, PQL is characterized by the class of all models $\langle I, R, \Pi, v\rangle$, where R is a symmetric, not necessarily reflexive, relation, and Π behaves like in the MQL case. Differently from QL and MQL, a world i of a PQL model may verify a contradiction. Since R is generally not reflexive, it may happen that $i \in v(\beta)$ and $i \perp v(\beta)$. Hence: $i \models \beta \wedge \neg\beta$. An adequate axiomatization for PQL can be obtained by dropping the orthomodular rule **R9** in our QL calculus. Like MQL, also PQL satisfies the finite-model property and consequently is decidable.

Interesting unsharp extensions of PQL are the *Brouwer–Zadeh logics* first investigated by Cattaneo and Nisticò (1989). A characteristic of these logics is a splitting of the connective "not" into two forms of negation: a fuzzylike negation, which gives rise to a paraconsistent behavior, and an intuitionisticlike negation. The fuzzy "not" represents a weak negation, which inverts the truth values truth and falsity, satisfies the double-negation principle, but generally violates the noncontradiction and the excluded-middle principles. The second "not" is a stronger negation, a kind of necessitation of the fuzzy "not." As a consequence, the language of the Brouwer–Zadeh logics is an extension of the QL language, with two primitive negations: $\neg$ represents the fuzzy "not," whereas $\sim$ is the intuitionistic "not." On this basis, a necessity operator can be defined in terms of the two negations:

$$L\alpha := \sim \neg\alpha.$$

In other words: "necessarily α" means the intuitionistic negation of the fuzzy negation of α. A possibility operator is then defined in terms of L and $\neg$:

$$M\alpha := \neg L \neg\alpha.$$

We will consider two forms of Brouwer–Zadeh logics: BZL (weak Brouwer–Zadeh logic) and BZL^3, which represents a form of three-valued quantum logic. Both logics admit of Hilbert-space exemplifications. Algebraically, BZL is characterized by the class of all models $\mathscr{M} = \langle \mathscr{A}, v\rangle$, where $\mathscr{A} = \langle A, \leq, ', ^{\sim}, \mathbf{1}, \mathbf{0}\rangle$ is a Brouwer–Zadeh lattice (simply a BZ lattice). In other words:

1. **a**. $\langle A, \leq, ', \mathbf{1}, \mathbf{0}\rangle$ is an involutive lattice with smallest element **0** and largest element **1**.
   b. $^{\sim}$ behaves like an intuitionistic complement:

   $$a \wedge a^{\sim} = \mathbf{0}.$$
   $$a \leq a^{\sim\sim}.$$
   $$\text{If } a \leq b, \quad \text{then } b^{\sim} \leq a^{\sim}.$$

   c. The following relation holds between the fuzzy and the intuitionistic complement:

   $$a^{\sim'} = a^{\sim\sim}.$$

   d. The *regularity condition* holds:

   $$a \wedge a' \leq b \vee b'.$$

2. v interprets the fuzzy negation $\neg$ as the fuzzy complement $'$, and the intuitionistic negation $\sim$ as the intuitionistic complement $^{\sim}$.

The logic BZL, which can be equivalently characterized also by a Kripkean semantics, is axiomatizable and decidable (Giuntini, 1991). The modal operators of

BZL behave similarly to the corresponding operators of the famous modal system S_5. For instance, $LL\alpha$ is equivalent to $L\alpha$; and $LM\alpha$ is equivalent to $M\alpha$.

The three-valued BZL^3 can be naturally characterized by a kind of many-valued possible-world semantics. The intuitive idea can be sketched as follows: One supposes that interpreting a language means associating to any sentence two *domains of certainty*: the domain of possible worlds where the sentence holds, and the domain of possible worlds where the sentence does not hold. All the other worlds are supposed to associate an intermediate truth value (*indetermined*) to our sentence. The models of this semantics will be called *models with positive and negative domains* (shortly, *ortho-pair models*).

Briefly, an ortho-pair model has the form $\mathscr{M} = \langle I, R, \Pi, v\rangle$, where

1. I is a nonempty set of worlds and R (the accessibility relation) is reflexive and symmetric (like in the Kripkean characterization of QL).

The possible propositions (in the sense of our definition of the Kripkean model for QL) will be here called *simple propositions*. The set Σ of all simple propositions gives rise to an ortholattice; let us indicate by $\#, \sqcap, \sqcup$ the lattice operations defined on Σ.

2. A *possible proposition* of $\mathscr{M}$ is any pair $\langle X_1, X_0\rangle$, where X_1, X_0 are simple propositions such that $X_1 \subseteq X_0^{\#}$ (in other words: X_1, X_0 are orthogonal). The following operations and relations are defined on the set of all possible propositions:
   a. the fuzzy complement
   $$\langle X_1, X_0\rangle' = \langle X_0, X_1\rangle;$$
   b. the intuitionistic complement
   $$\langle X_1, X_0\rangle^{\sim} = \langle X_0, X_0^{\#}\rangle;$$
   c. the propositional conjunction
   $$\langle X_1, X_0\rangle \wedge \langle Y_1, Y_0\rangle = \langle X_1 \sqcap Y_1, X_0 \sqcup Y_0\rangle;$$
   d. the order relation
   $$\langle X_1, X_0\rangle \leq \langle Y_1, Y_0\rangle \text{ iff } X_1 \subseteq Y_1 \text{ and } Y_0 \subseteq X_0.$$
3. Π is a set of possible propositions closed under $\wedge, ', {}^{\sim}$, and $\mathbf{0} := \langle \emptyset, I\rangle$.
4. v (the interpretation function) maps sentences into propositions in Π and interprets the connectives $\textcircled{\wedge}, \sim, \neg$ as the corresponding operations.

The other basic semantic definitions are like in the algebraic semantics. One can show that in any ortho-pair model the set of propositions has the structure of a BZ lattice. As a consequence, the logic BZL^3 is at least as strong as BZL. In fact, one can prove that BZL^3 is properly stronger than BZL. As a counterexample, let us consider an instance of the fuzzy excluded middle and an instance of the intuitionistic excluded middle applied to the same sentence α. One can easily check:

$$\alpha \textcircled{\vee} \neg\alpha \models_{\mathrm{BZL}^3} \alpha \textcircled{\vee} \sim \alpha \quad \text{and}$$
$$\alpha \textcircled{\vee} \sim \alpha \models_{\mathrm{BZL}^3} \alpha \textcircled{\vee} \neg\alpha.$$

However, generally

$$\alpha \textcircled{\vee} \neg\alpha \not\models_{\mathrm{BZL}} \alpha \textcircled{\vee} \sim \alpha.$$

Also BZL^3 is axiomatizable (Cattaneo et al., 1993) and can be characterized by means of an algebraic semantics.

8
Hilbert-Space Models of the Brouwer–Zadeh Logics

Hilbert-space models of both BZL and BZL^3 can be obtained in the framework of the *unsharp* (or *operational*) *approach* to quantum theory that was first proposed by Ludwig (1954) and developed (among others) by Kraus (1983), Davies (1976), Busch et al. (1991), and Cattaneo and Laudisa (1994). One of the basic ideas of this approach is a "liberalization" of the mathematical counterpart for the intuitive notion of "experimentally testable proposition." As we have seen, in orthodox Hilbert-space quantum mechanics, experimental propositions are mathematically represented as projections P on the Hilbert space $\mathscr{H}$ corresponding to the physical system $\mathscr{S}$ under investigation. If P is a projection representing a proposition and W is a density operator representing a state of $\mathscr{S}$, the number $\mathrm{tr}(WP)$ represents the probability value that the system $\mathscr{S}$ in the state W verifies P (*Born probability*). However, projections are not the only operators for which a Born probability can be defined. Let us consider the class $\mathscr{E}(\mathscr{H})$ of all linear bounded operators D such that for any density operator W,

$$\mathrm{tr}(WD) \in [0, 1].$$

It turns out that $\mathscr{E}(\mathscr{H})$ properly includes the set $L(\mathscr{H})$ of all projections on $\mathscr{H}$. In a sense, the elements of $\mathscr{E}(\mathscr{H})$ represent a "maximal" possible notion of experimental proposition, in agreement with the probabilistic rules of quantum theory. In the framework of the unsharp approach, the elements of $\mathscr{E}(\mathscr{H})$ have been called *effects*. An important difference between projections and proper effects is the following: Projections can be associated to *sharp* propositions having the form "the value for the observable A lies in the exact Borel set F," whereas effects may represent also *fuzzy* propositions like "the value of the observable A lies in the fuzzy Borel set F." As a consequence, there are effects D that are different from the null projection $\mathbb{0}$ and that are verified with certainty by no state [for any W, $\mathrm{tr}(WD) \neq 1$]. A limit case is represented by the *semitransparent effect* $\frac{1}{2}\mathbb{1}$ (where $\mathbb{1}$ is the identity operator), to which any state W assigns probability value $\frac{1}{2}$.

The class of all effects of $\mathscr{H}$ gives rise to a structure $\langle \mathscr{E}(\mathscr{H}), \leq, ', ^{\sim}, \mathbf{1}, \mathbf{0} \rangle$ which is a BZ poset (not a BZ lattice!). In other words, $\leq$ is a partial order with largest element $\mathbf{1}$ and smallest element $\mathbf{0}$, while the fuzzy and the intuitionistic complement ($'$ and $^{\sim}$) behave like in the BZ lattices. The relation and the operations of the effect-structure are defined as follows:

1. $D_1 \leq D_2$ iff for any density operator W, $\mathrm{tr}(WD_1) \leq \mathrm{tr}(WD_2)$.
2. $\mathbf{1} = \mathbb{1}$.
3. $D' = \mathbb{1} - D$.
4. $D^{\sim}$ is the projection $P_{\mathrm{Ker}(D)}$ into the subspace $\mathrm{Ker}(D)$, consisting of all vectors that are transformed by the operator D into the null vector.
5. $\mathbf{0} = \mathbf{1}'$.

In the particular case where D is a projection, it turns out that $D' = D^{\sim}$. In other words, the fuzzy and the intuitionistic complement coincide for sharp propositions. One can show that any BZ poset can be embedded into a complete BZ lattice [for the *MacNeille completion* (Birkhoff, 1967) of a BZ poset is a complete BZ lattice (Giuntini, 1991)]. As a consequence, the MacNeille completions of the effect-BZ posets represent natural Hilbert-space models for the logic BZL.

As to BZL[3], Hilbert-space models $\langle I, R, \Pi, \nu\rangle$ in the ortho-pair semantics can be constructed as follows:

1. I and R are defined like in the Kripkean Hilbertian models of QL. The simple propositions turn out to be in one-to-one correspondence to the set of the projections of $\mathscr{H}$.
2. Π is the set of all possible propositions. Any effect D can be transformed into a proposition $f(D) = \langle X_1^D, X_0^D\rangle$, where

$$X_1^D := \{\psi \in I | \mathrm{tr}(P_\psi D) = 1\} \quad \text{and}$$
$$X_0^D := \{\psi \in I | \mathrm{tr}(P_\psi D) = 0\}$$

   (P_ψ is the projection onto the unidimensional subspace spanned by the vector ψ). In other words, X_1^D and X_0^D represent respectively the positive and the negative domain of D (in a sense, the extensional meaning of D in the model). The map f turns out to preserve the order relation and the two complements.
3. The interpretation function ν follows the intuitive physical meaning of the atomic sentences.

9 Partial Quantum Logics

So far we have considered only examples of abstract quantum logics, where conjunctions and disjunctions are supposed to be always defined. However, as we have seen, the experimental and the probabilistic meaning of conjunctions of incompatible propositions in quantum theory has been often put in question. How do we construct logics where we admit that conjunctions and disjunctions are possibly meaningless? For instance, how do we give a natural semantic characterization for a logic corresponding to the class of all orthoalgebras or to the class of all orthomodular posets? Let us call these logics respectively *weak partial quantum logic* (WPaQL) and *strong partial quantum logic* (SPaQL). Are WPaQL and SPaQL axiomatizable?

9.1 Algebraic Semantics for WPaQL

The language of WPaQL contains two primitive connectives: the negation $\neg$, and the exclusive disjunction $\boxplus$ (*aut*). A conjunction is metalinguistically defined, via De Morgan's law:

$$\alpha \boxdot \beta := \neg(\neg\alpha \boxplus \neg\beta).$$

The intuitive idea underlying our semantics for WPaQL is the following: Disjunctions and conjunctions are considered "legitimate" from a mere linguistic point of view. However, semantically, a disjunction $\alpha \boxplus \beta$ will have the *intended meaning* only in the "well-behaved cases" (where the values of α and β are orthogonal in the corresponding orthoalgebra). Otherwise, $\alpha \boxplus \beta$ will have any meaning whatsoever (generally not connected with the meanings of α and β). A similar semantic "trick" is used in some standard treatments of the description operator ι ("the unique individual that satisfies a given property") in classical model theory.

DEFINITION 9.1.1 – An *algebraic model* of WPaQL is a pair $\mathscr{M} = \langle \mathscr{A}, \nu\rangle$, where

1. $\mathscr{A} = \langle \mathrm{A}, \oplus, \mathbf{1}, \mathbf{0}\rangle$ is an orthoalgebra.
2. ν (the interpretation function) satisfies the following conditions:
   $\nu(\alpha) \in A$, for any atomic sentence α;
   $\nu(\neg\beta) = \nu(\beta)'$, where $'$ is the orthocomplement operation that is defined in $\mathscr{A}$;

$$\nu(\beta \boxplus \gamma) = \begin{cases} \nu(\beta) \oplus \nu(\gamma), \text{ if} \nu(\beta) \oplus \nu(\gamma) \text{is defined} \\ \text{in} \mathscr{A}, \text{ any element otherwise.} \end{cases}$$

Accordingly, we will have that

$$\alpha \models_{\text{WPaQL}} \beta \quad \text{iff in any WPaQL model}$$
$$\mathscr{M} = \langle \mathscr{A}, \nu \rangle, \nu(\alpha) \leq \nu(\beta),$$

where $\leq$ is the partial order relation defined in $\mathscr{A}$.

9.2 An Axiomatization of Partial Quantum Logics

The logic WPaQL is axiomatizable. We present here a calculus that is obtained as a natural transformation of our QL calculus.

The rules of WPaQL are as follows:

R1 $\alpha \vdash \alpha$ (identity).

R2 $\dfrac{\alpha \vdash \beta \beta \vdash \gamma}{\alpha \vdash \gamma}$ (transitivity).

R3 $\alpha \vdash \neg\neg\alpha$ (weak double negation).

R4 $\neg\neg\alpha \vdash \alpha$ (strong double negation).

R5 $\dfrac{\alpha \vdash \beta}{\neg\beta \vdash \neg\alpha}$ (contraposition).

R6 $\beta \vdash \alpha \boxplus \neg\alpha$ (excluded middle).

R7 $\dfrac{\alpha \vdash \neg\beta\alpha \boxplus \neg\alpha \vdash \alpha \boxplus \beta}{\neg\alpha \vdash \beta}$ (unicity of negation).

R8 $\dfrac{\alpha \vdash \neg\beta\alpha \vdash \alpha_1\alpha_1 \vdash \alpha\beta \vdash \beta_1\beta_1 \vdash \beta}{\alpha \boxplus \beta \vdash \alpha_1 \boxplus \beta_1}$ (weak substitutivity).

R9 $\dfrac{\alpha \vdash \neg\beta}{\alpha \boxplus \beta \vdash \beta \boxplus \alpha}$ (weak commutativity).

R10 $\dfrac{\beta \vdash \neg\gamma\alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \vdash \neg\beta}$.

R11 $\dfrac{\beta \vdash \neg\gamma\alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \boxplus \beta \vdash \neg\gamma}$.

R12 $\dfrac{\beta \vdash \neg\gamma\alpha \vdash \neg(\beta \boxplus \gamma)}{\alpha \boxplus (\beta \boxplus \gamma) \vdash (\alpha \boxplus \beta) \boxplus \gamma}$.

R13 $\dfrac{\beta \vdash \neg\gamma\alpha \vdash \neg(\beta \boxplus \gamma)}{(\alpha \boxplus \beta) \boxplus \gamma) \vdash \alpha \boxplus (\beta \boxplus \gamma)}$.

(**R10–R13** require a weak associativity.)

The other basic proof-theoretic definitions are given like in the QL case. Some derivable rules of the calculus are the following:

D1 $\dfrac{\alpha \vdash \beta}{\beta \vdash \alpha \boxplus \neg(\alpha \boxplus \neg\beta)}$.

D2 $\dfrac{\alpha \vdash \beta}{\alpha \boxplus \neg(\alpha \boxplus \neg\beta) \vdash \beta}$.

D3 $\dfrac{\alpha \vdash \neg\gamma\beta \vdash \neg\gamma\alpha \boxplus \gamma \vdash \beta \boxplus \gamma\beta \boxplus \gamma \vdash \alpha \boxplus \gamma}{\alpha \vdash \beta}$.

D4 $\dfrac{\alpha \vdash \neg\beta\alpha \vdash \gamma\beta \vdash \gamma\gamma \vdash \alpha \boxplus \beta}{\alpha \boxplus \beta \vdash \gamma}$.

The proof-theoretic and the semantic-consequence relations for the logic WPaQL are reciprocally adequate. Namely, a soundness and a completeness theorem can be proved.

THEOREM 9.2.1 – *Soundness*

If $\alpha \vdash_{\text{WPaQL}} \beta$ then $\alpha \models_{\text{WPaQL}} \beta$.

THEOREM 9.2.2 – *Completeness*

If $\alpha \models_{\text{WPaQL}} \beta$ then $\alpha \vdash_{\text{WPaQL}} \beta$.

As to strong partial quantum logic (SPaQL), an axiomatization can be obtained by adding to our WPaQL calculus the following rule:

R14 $\dfrac{\alpha \vdash \neg\beta \; \alpha \vdash \gamma \; \beta \vdash \gamma}{\alpha \boxplus \beta \vdash \gamma}$.

Semantically, the models of SPaQL will be based on orthoalgebras $\mathscr{A} = \langle A, \oplus, \mathbf{1}, \mathbf{0} \rangle$, satisfying the following condition: If defined, $a \oplus b$ is the sup of a and b. As we have seen in Sec. 3.1, this condition is necessary and sufficient in order to make the orthoposet induced by the orthoalgebra $\mathscr{A}$ an orthomodular poset. The soundness and the completeness theorems for SPaQL (with respect to this semantics) can be proved similarly to the case of WPaQL.

10
Critique of Abstract Quantum Logics

Do abstract quantum logics represent "real" logics or should they rather be regarded as mere extrapolations from particular algebraic structures that arise in the mathematical formalism of quantum mechanics? Different answers to this question have been given in the history of the logicoalgebraic approach to quantum theory. According to our analysis, the logical *status* of abstract quantum logics can be hardly put in question. These logics turn out to satisfy all the canonical conditions that the present community of logicians require in order to call a given abstract object a *logic*: syntactical and semantical descriptions, proofs of soundness and completeness theorems, and so on.

Has the quantum-logical research definitely shown that "logic is empirical"? At the very beginning of the history of quantum logic, the thesis according to which the choice of the "right" logic to be used in a given theoretic situation may depend also on experimental data appeared a kind of extremistic view, in contrast with the traditional description of logic as "an *a priori* and analytical science." These days, an empirical position in logic is no more regarded as a "daring heresy." At the same time, we are facing a new difficulty: As we have seen, quantum logic is not unique. Besides orthodox quantum logic, different forms of partial and unsharp quantum logics have been developed. In this situation, one can wonder whether it is still reasonable to look for the most adequate abstract logic that should faithfully represent the structures arising in the quantum world.

A question that has been often discussed concerns the compatibility between quantum logic and the mathematical formalism of quantum theory, based on classical logic. Is the quantum physicist bound to a kind of "logical schizophrenia"? At first sight, the compresence of different logics in one and the same theory may give a sense of uneasiness. However, the splitting of the basic logical operations (negation, conjunction, disjunction,. . .) into different connectives with different meanings and uses is now a well-accepted logical phenomenon that admits consistent descriptions. As we have seen, classical and quantum logic turn out to apply to different sublanguages of quantum theory that must be sharply distinguished.

Glossary

Abstract Quantum Logic: A logic $\langle FL, \vdash, \models \rangle$, where the proof-theoretic and the semantic-consequence relations violate some characteristic classical principles like the distributivity of conjunction and disjunction.

Adjoint: If A is a bounded operator on a Hilbert space, then the adjoint of A is the unique bounded operator A^* that satisfies $\langle A\psi|\phi\rangle = \langle \psi|A^*\phi\rangle$ for all vectors ψ, ϕ in the space.

Algebraic Model of a Language: A pair $\langle \mathscr{A}, \nu \rangle$ consisting of an algebraic structure $\mathscr{A}$ and of an interpretation function ν that transforms the sentences of the language into elements of $\mathscr{A}$, preserving the logical form.

Algebraic Semantics: The basic idea is that interpreting a formal language means associating to any sentence an element of an algebraic structure.

Attribute: One of a class of properties affiliated with a physical system. At any

given moment some of the attributes of the system may be actual, while others are only potential.

Axiomatizable Logic: A logic is axiomatizable when its concept of proof is decidable.

Boolean Algebra: An orthocomplemented lattice L that satisfies the distributive law $p \wedge (q \vee r) = (p \wedge q) \vee (p \wedge r)$ for all $p, q, r \in L$.

Borel Set: A (real) Borel set is a set that belongs to the smallest collection of subsets of the real numbers $\mathbb{R}$ that contains all open intervals and is closed under the formation of complements and countable unions.

Brouwer–Zadeh Lattice (or Poset): A lattice (or poset) L with smallest element $\mathbf{1}$ and largest element $\mathbf{0}$, equipped with a regular involution $'$ (a fuzzylike complement), and an intuitionisticlike complement $^{\sim}$, subject to the following conditions for all $p, q \in L$: (i) $p \leq q \Rightarrow q^{\sim} \leq p^{\sim}$, (ii) $p \wedge p^{\sim} = \mathbf{0}$, (iii) $p \leq p^{\sim\sim}$, (iv) $p^{\sim\prime} = p^{\sim\sim}$.

Brouwer–Zadeh Logic: A logic that is characterized by the class of all models based on Brouwer–Zadeh lattices.

Compatible: A set C of elements in an orthoalgebra L is a compatible set if there is a Boolean subalgebra B of L such that $C \subseteq$ B.

Complete Lattice: A lattice in which every subset has a least upper bound, or join, and a greatest lower bound, or meet.

Completeness: A logic $\langle FL, \vdash, \models \rangle$ is (semantically) complete when all the semantic consequences are proof-theoretic consequences (if $\alpha \models \beta$ then $\alpha \vdash \beta$).

Conjunction (in an Orthoalgebra): If L is an orthoalgebra and $p, q \in L$, then p and q have a conjunction $p \,\&\, q \in L$ if there is a Boolean subalgebra B of L with $p, q \in B$, and the meet $p \wedge_B q$ of p and q in B is independent of the choice of B, in which case $p \,\&\, q = p \wedge_B q$.

Decidable Logic: A logic is decidable when its proof-theoretic consequence relation is decidable.

Density Operator: A self-adjoint, nonnegative, trace-class operator W on a Hilbert space, such that $\mathrm{tr}(W) = 1$.

Direct Sum (or Cartesian Product) of Hilbert Spaces: The direct sum of the Hilbert spaces $\mathscr{H}$ and $\mathscr{K}$ is the Hilbert space $\mathscr{H} \oplus \mathscr{K}$ consisting of all ordered pairs (x, y) with $x \in \mathscr{H}, y \in \mathscr{K}$, and with coordinatewise vector operations. The inner product is defined by $\langle (x_1, y_1) | (x_2, y_2) \rangle = \langle x_1 | x_2 \rangle + \langle y_1 | y_2 \rangle$.

Disjunction (in an Orthoalgebra): If L is an orthoalgebra and $p, q \in L$, then p and q have a disjunction $p + q$ if there is a Boolean subalgebra B of L with $p, q \in B$ and the join $p \vee_B q$ of p and q in B is independent of the choice of B, in which case $p + q = p \vee_B q$.

Dynamics: The evolution in time of the state of a physical system.

Effect: A linear bounded operator A of a Hilbert space such that for any density operator W, $\mathrm{tr}(WA) \in [0, 1]$. In the unsharp approach to quantum mechanics, effects represent possible experimental propositions.

Experimental Proposition: A proposition whose truth value can be determined by conducting an experiment.

Greatest vs Maximal: If L is a partially ordered set (poset) and $X \subseteq L$, then an element $b \in X$ is a greatest element of X if $x \leq b$ for all $x \in X$. An element $b \in X$ is a maximal element of X if there exists no element $x \in X$ with $b < x$.

Involution: A mapping $p \to p'$ on a poset L satisfying the following conditions for all $p, q \in L$: (i) $p \leq q \Rightarrow q' \leq p'$, (ii) $p'' = p$.

Involutive Lattice (or Poset): A lattice (or poset) with smallest element **0** and largest element **1**, equipped with an involution.

Join: If L is a partially ordered set (poset) and $p, q \in L$, then the join (or least upper bound) of p and q in L, denoted by $p \vee q$ if it exists, is the unique element of L satisfying the following conditions: (i) $p, q \leq p \vee q$ and (ii) $r \in L$ with $p, q \leq r \Rightarrow p \vee q \leq r$.

Kripkean Model of a Language: A system $\langle I, R_1, \ldots, R_m, o_1, \ldots, o_n, \Pi, \nu\rangle$ consisting of a set I of possible worlds, a (possibly empty) sequence of world relations $R_1, \ldots, R_m$ and of world operations $o_1, \ldots, o_n$, a family Π of subsets of I (called the propositions), and an interpretation function ν that transforms the sentences of the language into propositions, preserving the logical form.

Kripkean Semantics: The basic idea is that interpreting a formal language means associating to any sentence the set of the possible worlds where the sentence holds. This set is called also the proposition associated to the sentence.

Lattice: A partially ordered set (poset) in which every pair of elements p, q has a least upper bound, or join, $p \vee q$ and a greatest lower bound, or meet, $p \wedge q$.

Least vs Minimal: If L is a partially ordered set (poset) and $X \subseteq L$, then an element $a \in X$ is a least element of X if $a \leq x$ for all $x \in X$. An element $a \in X$ is a minimal element of X if there exists no element $x \in X$ with $x < a$.

Lindenbaum Property: A logic satisfies the Lindenbaum property when any noncontradictory set of sentences T can be extended to a noncontradictory and *complete* set T^* (such that T^* contains, for any sentence of the language, either the sentence or its negation). Abstract quantum logics generally violate the Lindenbaum property.

Logic: According to the tradition of logical methods, a logic can be described as a system $\langle FL, \vdash, \models\rangle$ consisting of a *formal language*, a *proof-theoretic-consequence* relation $\vdash$ (based on a notion of *proof*), and a *semantic-consequence* relation $\models$ (based on a notion of *model* and of *truth*).

MacNeille Completion of a Brouwer–Zadeh Lattice: Let L be a Brouwer–Zadeh lattice (or poset). For $X \subseteq L$, let $X' = \{p \in L | \forall q \in X, p \leq q'\}$, $X^{\sim} = \{p \in L | \forall q \in X,\ p \leq q^{\sim}\}$, and $P(L) = \{X \subseteq L | X = X''\}$. The structure $\mathscr{P}(L) = \langle P(L), \subseteq, ', ^{\sim}, \{\}'', L\rangle$ is called the MacNeille completion of L. $\mathscr{P}(L)$ is a complete Brouwer–Zadeh lattice and L is embeddable into $\mathscr{P}(L)$ via the mapping $p \to \langle p]$, where $\langle p] = \{q \in L | q \leq p\}$.

Meet: If L is a partially ordered set (poset) and $p, q \in L$, then the meet (or greatest lower bound) of p and q, denoted by $p \wedge q$ if it exists, is the unique element of L satisfying the following conditions: (i) $p \wedge q \leq p, q$ and (ii) $r \in L$ with $r \leq p, q \Rightarrow r \leq p \wedge q$.

Minimal Quantum Logic: A logic that is semantically characterized by the class of

all models based on orthocomplemented lattices.

Modular Lattice: A lattice L satisfying the modular law: $p \leq r \Rightarrow p \vee (q \wedge r) = (p \vee q) \wedge r$ for all $p, q, r \in L$.

Nonnegative Operator: A self-adjoint operator A on a Hilbert space $\mathscr{H}$ such that $\langle A\psi|\psi\rangle \geq 0$ for all vectors $\psi \in \mathscr{H}$.

Observable or Dynamical Variable: A numerical variable associated with a physical system the value of which can be determined by conducting a test, a measurement, or an experiment on the system.

Orthoalgebra: A mathematical system consisting of a set L with two special elements $\mathbf{0}$, $\mathbf{1}$ and equipped with a relation $\perp$ such that, for each pair $p, q \in L$ with $p \perp q$, an orthogonal sum $p \oplus q \in L$ is defined subject to the following conditions for all $p, q, r \in L$: (i) $p \perp q \Rightarrow q \perp p$ and $p \oplus q = q \oplus p$, (ii) $p \perp q$ and $(p \oplus q) \perp r \Rightarrow q \perp r, p \perp (q \oplus r)$ and $p \oplus (q \oplus r) = (p \oplus q) \oplus r$, (iii) $p \in L \Rightarrow$ there is a unique $p' \in L$ such that $p \perp p'$ and $p \oplus p' = \mathbf{1}$, and (iv) $p \perp p \Rightarrow p = \mathbf{0}$.

Orthocomplementation: A mapping $p \to p'$ on a poset L with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$ satisfying the following conditions for all $p, q \in L$: (i) $p \vee p' = \mathbf{1}$, (ii) $p \wedge p' = \mathbf{0}$, (iii) $p \leq q \Rightarrow q' \leq p'$, and (iv) $p'' = p$.

Orthocomplemented Lattice (or Poset): A lattice (or poset) equipped with an orthocomplementation $p \to p'$.

Orthodox Quantum Logic: A logic that is semantically characterized by the class of all algebraic models based on orthomodular lattices. Standard quantum logic is a particular model of orthodox quantum logic.

Orthogonal: If L is an orthocomplemented poset and $p, q \in L$, then p is orthogonal to q, in symbols $p \perp q$, if $p \leq q'$.

Orthomodular Lattice: An orthocomplemented lattice L satisfying the orthomodular law: For all $p, q \in L$, $p \leq q \Rightarrow q = p \vee (q \wedge p')$.

Orthomodular Poset: An orthoalgebra L such that, for all $p, q \in L$, $p \perp q \Rightarrow p \oplus q = p \vee q$.

Paraconsistent Quantum Logic: A logic that is semantically characterized by the class of all models based on involutive lattices with smallest element $\mathbf{0}$ and largest element $\mathbf{1}$.

Partially Ordered Set (or Poset): A set L equipped with a relation $\leq$ satisfying the following conditions for all $p, q, r \in L$: (i) $p \leq p$, (ii) $p \leq q$ and $q \leq p \Rightarrow p = q$, (iii) $p \leq q$ and $q \leq r \Rightarrow p \leq r$.

Probability Measure: A function $\omega : L \to [0, 1] \subseteq \mathbb{R}$ on an orthoalgebra L such that $\omega(\mathbf{0}) = 0, \omega(\mathbf{1}) = 1$, and, for all $p, q \in L$ with $p \perp q, \omega(p \oplus q) = \omega(p) + \omega(q)$.

Projection: An operator P on a Hilbert space that is self-adjoint ($P = P^*$) and idem-potent ($P = P^2$).

Pure State: A state ψ is pure if the set $\{\psi\}$ consisting only of that state is superposition closed. In Hilbert-space quantum mechanics, the pure states are precisely the vector states.

Quantum Logic: The study of the formal structure of experimental propositions affiliated with a quantum physical system, or

any mathematical model (e.g., an orthoalgebra) representing such a structure.

Regular Involution: An involution $'$ on a poset L that satisfies the regularity condition: For all $p, q \in L$: $p \leq p'$ and $q \leq q' \Rightarrow p \leq q'$. If L is a lattice, then an involution is regular iff it satisfies the Kleene condition: For all $p, q \in L$: $p \wedge p' \leq q \vee q'$.

Soundness: A logic $\langle FL, \vdash, \models \rangle$ is sound when all the proof-theoretic consequences are semantic consequences (if $\alpha \vdash \beta$ then $\alpha \models \beta$).

Spectral Measure: A mapping from real Borel sets into projection operators on a Hilbert space that maps the empty set into $\mathbb{0}$, maps $\mathbb{R}$ into $\mathbb{1}$, and maps the union of a disjoint sequence of real Borel sets into the least upper bound (join) of the corresponding projection operators.

Spectral Theorem: The theorem establishing a one-to-one correspondence between (not necessarily bounded) self-adjoint operators A on a Hilbert space $\mathscr{H}$ and spectral measures $E \to P_E$ on $\mathscr{H}$ such that, if $P_\lambda = P_{(-\infty,\lambda]}$ for all $\lambda \in \mathbb{R}$, then $A = \int_{-\infty}^{\infty} \lambda dP_\lambda$.

Spectrum: If A is a (not necessarily bounded) self-adjoint operator on a Hilbert space and $E \to P_E$ is the corresponding spectral measure, then a real number λ belongs to the spectrum of A if $P_{(\lambda-\epsilon,\lambda+\epsilon)} \neq \mathbb{0}$ for all $\varepsilon > 0$.

Standard Quantum Logic: The complete orthomodular lattice L of all projection operators on a Hilbert space. For $P, Q \in L$, $P \leq Q$ is defined to mean that $P = PQ$ and the orthocomplement of P is defined by $P' = \mathbb{1} - P$.

State: The state of a physical system encodes all information concerning the results of conducting tests or measuring observables on the system. It is usually assumed that, corresponding to each state ψ of the system, there is a probability measure ω_ψ on the logic L of the system such that $\omega_\psi(p)$ is the probability that the experimental proposition $p \in L$ is true when the system is in the state ψ.

State Space: The set of all possible states of the physical system.

Strong Partial Quantum Logic: A logic that is semantically characterized by the class of all models based on orthomodular posets.

Superselection Rule: A rule that determines the possible states of a physical system. The usual quantum-mechanical superselection rules state that only vector states that commute with a certain set of pairwise orthogonal projections (i.e., projections onto superposition sectors) represent possible states of the systems.

Support: If W is a density operator on a Hilbert space $\mathscr{H}$, then the support of W is defined to be the set $\mathrm{supp}(W)$ of all projection operators P on $\mathscr{H}$ such that $\mathrm{tr}(WP) \neq 0$. More generally, if ω is a probability measure on an orthoalgebra L, then $\mathrm{supp}(\omega) = \{p \in L|_\omega(p) \neq 0\}$.

Superposition Closed: A set of states is superposition closed if it contains all of its own superpositions.

Superposition in Hilbert Space: For a Hilbert space $\mathscr{H}$, if (W_α) is a family of vector states determined by the corresponding family (ψ_α) of normalized vectors in $\mathscr{H}$, and if ψ is a normalized linear combination of the vectors in this family, then the vector state W determined by ψ is

a (coherent) superposition of the family (W_α). If (W_α) is an arbitrary family of density operators on $\mathscr{H}$ and (t_α) is a corresponding family of nonnegative real numbers such that $\Sigma_\alpha t_\alpha = 1$, then $W = \Sigma_\alpha t_\alpha W_\alpha$ is an (incoherent) superposition (or mixtures) of the family (W_α).

Superposition in an Orthoalgebra: A probability measure ω on an orthoalgebra L is a superposition of a family (ω_α) of probability measures on L if $\mathrm{supp}(\omega) \subseteq \cup_\alpha \mathrm{supp}(\omega_\alpha)$.

Tensor Product of Hilbert Spaces: If $\mathscr{H}$ and $\mathscr{K}$ are Hilbert spaces, then the tensor product $\mathscr{H} \otimes \mathscr{K}$ is a Hilbert space together with a mapping $(x, y) \to x \otimes y \in \mathscr{H} \otimes \mathscr{K}$ for $x \in \mathscr{H}, y \in \mathscr{K}$ that is separately linear in each argument and has the property that if (ψ_i) is an orthonormal basis for $\mathscr{H}$ and (ϕ_j) is an orthonormal basis for $\mathscr{K}$, then $(\psi_i \otimes \phi_j)$ is an orthonormal basis for $\mathscr{H} \otimes \mathscr{K}$.

Trace: The trace of an operator A on a Hilbert space $\mathscr{H}$ is defined by $\mathrm{tr}(A) = \Sigma_{\psi \in B} \langle A\psi | \psi \rangle$, where B is an orthonormal basis for $\mathscr{H}$, provided that the series converges.

Trace Class: An operator A on a Hilbert space $\mathscr{H}$ belongs to the trace class if $\mathrm{tr}(A) = \Sigma_{\psi \in B} \langle A\psi | \psi \rangle$ converges absolutely, where B is an orthonormal basis for $\mathscr{H}$.

Unsharp Quantum Logics: Examples of paraconsistent logics where the noncontradiction principle is generally violated.

Vector State: A probability measure on the standard quantum logic L of a Hilbert space $\mathscr{H}$ determined by a normalized vector $\psi \in \mathscr{H}$ and assigning to each projection operator $P \in L$ the probability $\langle P\psi | \psi \rangle$.

Weak Partial Quantum Logic: A logic that is semantically characterized by the class of all models based on orthoalgebras.

List of Works Cited

Aerts, D. (1982), "Description of Many Physical Entities without the Paradoxes Encountered in Quantum mechanics," *Found. Phys.* **12**, 1131–1170.

Beltrametti, E., Cassinelli, G. (1981), "The Logic of Quantum Mechanics," in: Gian-Carlo Rota (Ed.), Encyclopedia of Mathematics and its Applications, vol. 15, Reading, MA: Addison-Wesley.

Bennett, M., Foulis, D. (1990), "Superposition in Quantum and Classical Mechanics," *Found. Phys.* **20**, 733–744.

Birkhoff, G. (1967), *Lattice Theory*, 3rd ed., Providence, RI: American Mathematical Society Colloquium Publications, XXV.

Birkhoff, G. (1961), "Lattices in Applied Mathematics," in: R. P. Dilworth (Ed.), *Lattice Theory*, Proceedings of the Second Symposium in Pure Mathematics of the American Mathematical Society, Providence, RI: American Mathematical Society.

Birkhoff, G., von Neumann, J. (1936), "The Logic of Quantum Mechanics," *Ann. Math.* **37**, 823–843.

Busch, P., Lahti, P., Mittelstaedt, P. (1991), *The Quantum Theory of Measurement*, Lecture Notes in Physics, New Series m2, Berlin/Heidelberg/New York: Springer Verlag.

Cattaneo, G., Dalla Chiara, M. L., Giuntini, R. (1993), "Fuzzy-Intuitionistic Quantum Logics," *Studia Logica* **52**, 1–24.

Cattaneo, G., Laudisa, F. (1994), "Axiomatic Unsharp Quantum Mechanics (from Mackey to Ludwig and Piron)," *Found. Phys.* **24**, 631–684.

Cattaneo, G., Nisticò, G. (1989), "Brouwer-Zadeh Posets and Three-Valued Łukasiewicz Posets," *Fuzzy Sets Systems* **33**, 165–190.

Dalla Chiara, M. L. (1986), "Quantum Logic," in: D. Gabbau, F. Guenthner (Eds.), *Handbook*

of Philosophical Logic, III, Dordrecht, Netherlands: Reidel.

Dalla Chiara, M. L., Giuntini, R. (1989), "Paraconsistent Quantum Logics," *Found. Phys.* **19**, 891–904.

Davies, E. B. (1976), *Quantum Theory of Open Systems*, New York: Academic.

D'Espagnat, B. (1971), *Conceptual Foundations of Quantum Mechanics*, Menlo Park, NJ: Benjamin.

Foulis, D. (1962), "A note on Orthomodular Lattices," *Port. Math.* **21**, 65–72.

Foulis, D. (1989), "Coupled Physical Systems," *Found. Phys.* **7**, 905–922.

Foulis, D., Randall, C. (1981), "Empirical Logic and Tensor Products," in: *Interpretations and Foundations of Quantum Theory*, Grundlagen der Exakt Naturwissenschaften, Vol. 5, Mannheim/Wien: Bibliographisches Institut.

Foulis, D., Greechie, R., Rüttimann, G. (1992), "Filters and Supports in Orthoalgebras," *Int. J. Theor. Phys.* **31**, 789–807.

Foulis, D., Randall, C., Piron, C. (1983), "Realism, Operationalism, and Quantum Mechanics," *Found. Phys.* **8**, 813–842.

Gibbins, P. (1985), "A User-Friendly Quantum Logic," *Log. Anal.* **112**, 353–362.

Giuntini, R. (1991), "A Semantical Investigation on Brouwer-Zadeh Logic," *J. Philos. Log.* **20**, 411–433.

Gleason, A. (1957), "Measures of Closed Subspaces of a Hilbert Space," *J. Math. Mech.* **6**, 885–893.

Goldblatt, R. (1974), "Semantic Analysis of Orthologic," *J. Philos. Log.* **3**, 19–35.

Greechie, R. (1981), "A Non-standard Quantum Logic with a Strong Set of States," in: E. Beltrametti, B. van Fraassen (Eds.), *Current Issues in Quantum Logic*, Ettore Majorana International Science Series, vol. 8, New York: Plenum.

Greechie, R., Gudder, S. (1973), "Quantum Logics," in: C. Hooker (Ed.), *Contemporary Research in the Foundations and Philosophical of Quantum Theory*, Dordrecht, Netherlands: Reidel.

Gudder, S. (1988), *Quantum Probability*, San Diego: Academic.

Hardegree, G. (1979), "The Conditional in Abstract and Concrete Quantum Logic," in: C. Hooker (Ed.), *The Logic-Algebraic Approach to Quantum Mechanics, II*, The University of Western Ontario Series in Philosophy of Science, vol. 5, Dordrecht, Netherlands: Reidel.

Hardegree, G., Frazer, P. (1981), "Charting the Labyrinth of Quantum Logics," in: E. Beltrametti, B. van Fraassen (Eds.), *Current Issues in Quantum Logic*, Ettore Majorana International Science Series, vol. 8, New York: Plenum.

Husimi, K. (1937), "Studies on the Foundation of Quantum Mechanics I," *Proc. Phys. Math. Soc. Jpn.* **19**, 766–789.

Jauch, J. (1968), *Foundations of Quantum Mechanics*, Reading, MA: Addison-Wesley.

Jaynes, E. (1989), in: R. D. Rosenkrantz (Ed.), *Papers on Probability, Statistics, and Statistical Physics*, Dordrecht/Boston/London: Kluwer.

Kalmbach, G. (1983), *Orthomodular Lattices*, New York: Academic.

Keller, H. (1980), "Ein Nicht-klassischer Hilbertscher Raum," *Math. Z.* **172**, 41–49.

Kläy, M. (1988), "Einstein-Podolsky-Rosen Experiments: The Structure of the Probability Space, I," *Found. Phys. Lett.* **1**, 205–244.

Kläy, M., Randall, C., Foulis, D. (1987), "Tensor Products and Probability Weights," *Int. J. Theor. Phys.* **26**, 199–219.

Kolmogorov, A. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechung*, Berlin: Springer; translation (1956), *Foundations of Probability*, New York: Chelsea.

Kraus, K. (1983), *States, Effects and Operations*, Lecture Notes in Physics vol. 190, Berlin/Heidelberg/New York: Springer-Verlag.

Lock, P., Hardegree, G. (1984a), "Connections among Quantum Logics, Part 1, Quantum Propositional Logics," *Int. J. Theor. Phys.* **24**, 43–53.

Lock, P., Hardegree, G. (1984b), "Connections among Quantum Logics, Part 2, Quantum Event Logics," *Int. J. Theor. Phys.* **24**, 55–61.

Lock, R. (1981), *Constructing the Tensor Product of Generalized Sample Spaces*, Ph.D. Thesis, University of Massachusetts, Amherst.

Loomis, L. (1955), *The Lattice Theoretic Background of the Dimension Theory of Operator Algebra*, Memoirs of the American Mathematical Society vol. 18.

Ludwig, G. (1954), *Die Grundlagen der Quantenmechanik*, Berlin: Springer; translation (1983), *Foundations of Quantum Mechanics, I*, Berlin: Springer.

Mackey, G. (1957), "Quantum Mechanics and Hilbert Space," *Am. Math. Monthly* **64**, 45–57.

Mackey, G. (1963), *The Mathematical Foundations of Quantum Mechanics*, New York: Benjamin.

Maeda, S. (1955), "Dimension Functions on Certain General Lattices," *J. Sci. Hiroshima Univ.* **A19**, 211–237.

Mittelstaedt, P. (1978), *Quantum Logic*, Dordrecht: Reidel.

Murray, F., von Neumann, J. (1936), "On Rings of Operators," *Ann. Math.* **37**, 116–229.

Nishimura, H. (1980), "Sequential Method in Quantum Logic," *J. Symbol. Log.* **45**, 339–352.

Piron, C. (1976), in: A. Wightman (Ed.), *Foundations of Quantum Physics*, Mathematical Physics Monograph Series, Reading, MA: Benjamin.

Popper, K. (1959), *Br. J. Phil. Sci.* **10**, 25.

Pták, P., Pulmannová, S. (1991), *Orthomodular Structures as Quantum Logics*, Dordrecht/Boston/London: Kluwer.

Randall, C. H., Foulis, D. J. (1973), "Operational Statistics II, Manuals of Operation and Their Logics," *J. Math. Phys.* **14**, 1472–1480.

Randall, C., Foulis, D. (1978), "New Definitions and Theorems," University of Massachusetts Mimeographed Notes.

Stone, M. (1932), "On One Parameter Unitary Groups in Hilbert Space," *Ann. Math.* **33**, 643–648.

von Neumann, J. (1932), *Grundlagen der Quantenmechanik*, Berlin: Springer.

von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics*, translation of von Neumann (1932), Princeton, NJ: Princeton Univ. Press.

von Neumann, J. (1960), *Continuous Geometry*, Princeton, NJ: Princeton Univ. Press.

Wick, G., Wightman, A., Wigner, E. (1952), "Intrinsic Parity of Elementary Particles," *Phys. Rev.* **88**, 101–105.

Wilbur, W. J. (1977), "On Characterizing the Standard Quantum Logics," *Trans. Am. Math. Soc.* **233**, 265–282.

Wright, R. (1978), "Spin Manuals," in: A. R. Marlow (Ed.), *Mathematical Foundations of Quantum Theory*, New York: Academic.

Special Functions

Charlie Harper
Department of Physics, California State University, Hayward, California, USA

Introduction ***476***
1 The Sturm–Liouville Theory ***477***
1.1 Introduction *477*
1.2 Hermitian Operators and Their Eigenvalues *480*
1.3 Orthogonality Condition and Completeness of Eigenfunctions *481*
1.4 Orthogonal Polynomials and Functions *482*
1.4.1 Recurrence Formula for Orthogonal Polynomials *482*
1.4.2 Rodrigues Formulas *482*
1.4.3 The Generating Function *483*
1.5 Factorization of the Sturm–Liouville Equation *484*
1.6 Connection with Local Lie-Group Theory *488*
2 The Hypergeometric Function, $_2F_1(a, b, c; z)$ ***490***
2.1 Properties of the Hypergeometric Differential Equation *490*
2.2 Properties of F(a,b,c;z) *491*
2.3 Helmholtz's Differential Equation in Spherical Coordinates *494*
2.4 Associated Legendre Functions and Legendre Polynomials *495*
2.5 The Radial Equation *497*
3 The Confluent Hypergeometric Function, $_1F_1(a, c; x)$ ***499***
3.1 More on Hermite Polynomials *499*
3.2 More on the Laguerre and Associated Laguerre Polynomials *501*
4 Helmholtz's Differential Equation in Cylindrical Coordinates ***503***
4.1 Solutions of Bessel's Differential Equation *504*
4.2 Bessel Functions of the First Kind *504*
4.3 Neumann Functions *504*
4.4 Hankel Functions *504*
4.5 Modified Bessel Functions *505*
4.6 Spherical Bessel Functions *506*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

5 Other Special Functions used in Mathematical Physics ***507***
5.1 Some Other Special Functions – Type 1 *509*
5.2 Some Other Special Functions – Type 2 *509*
Glossary ***510***
List of Works Cited ***511***
Further Reading ***511***

Introduction

This article is devoted to the theory and applications of a set higher transcendental functions that arise naturally in mathematical physics. These higher transcendental functions are referred to as special functions, and they arise

1. when solving, in certain curvilinear coordinate systems, partial differential equations that are defined by physical problems and/or
2. when finding eigenfunctions and eigenvalues of differential operators.

The partial-differential-equations approach to special functions involves use of the separation-of-variables method and either the Frobenius–Fuchs power series solutions of one or more resulting ordinary differential equations or an Infeld–Hull type factorization procedure for finding eigenfunctions and eigenvalues of second-order ordinary linear differential equations.

The focus is on a class of physical problems whose differential-equation formulation involves the Laplacian operator, ∇^2. The resulting partial differential equations include the Laplace equation, the heat-conduction (diffusion) equation, the mechanical wave-motion equation, and the Schrödinger wave equation. After application of the separation-of-variables method, the resulting time-independent parts of all these partial differential equations may be written in the form of the Helmholtz differential equation, $\nabla^2 u + k^2 u = 0$. Problems involving the Helmholtz differential equation in spherical coordinates lead to spherical harmonics, Legendre polynomials and associated Legendre functions, Laguerre and associated Laguerre polynomials, and spherical Bessel functions. Problems modeled by use of Helmholtz's differential equation in cylindrical coordinates involve the various types of Bessel functions. Solutions of the Schrödinger wave equation for a linear harmonic oscillator are expressed in terms of Hermite polynomials.

Special functions such as Hermite polynomials, Legendre polynomials and associated Legendre functions, spherical harmonics, Laguerre and associated Laguerre polynomials, and Bessel functions are widely used in mathematical physics and are the main focus of this article; these special functions are special cases of the hypergeometric function ${}_2F_1(a, b, c; z)$ or confluent hypergeometric function ${}_1F_1(a, c; z)$. There exist other useful special functions in mathematical physics that are not expressible in terms of ${}_2F_1$ or ${}_1F_1$. The functions ${}_2F_1$ and ${}_1F_1$ may be developed from the following main view-points:

1. ordinary differential equations and the Frobenius–Fuchs power series method,
2. factorization of ordinary differential equations, and

3. representation theory of local Lie groups.

In a lecture course, *The Application of Group Theory to the Special Functions of Mathematical Physics* (unpublished lecture notes, Princeton University, Princeton, NJ, 1955; see Talman, 1968), Wigner pointed out that certain classes of special functions arise as matrix elements of the representations of local Lie groups. Since Wigner's work, many other group-theoretical approaches to special functions have been developed. The purposes of these various group-theoretical approaches are to show unity (or demonstrate a central foundation) among the extremely large number of special functions and to derive their various known basic properties. It is important, however, to note that there exist no single approach to special functions that unites all special functions and illuminates all of their various properties.

Section 1 is devoted to a treatment of the Sturm–Liouville theory and orthogonal polynomials, since these concepts provide insight into properties of solutions of the second-order ordinary linear differential equations in which we are interested. The factorization method of Infeld and Hull and its connection to a group-theoretical foundation of certain special functions are summarized in Secs. 1.5 and 1.6.

A discussion of the properties of the hypergeometric differential equation and the hypergeometric function ${}_2F_1$ is given in Sec. 2. The Legendre polynomials and associated Legendre functions are special examples of ${}_2F_1$, and their important properties are developed in Sec. 2. Also, some discussion of the Chebyshev polynomial, Gegenbauer polynomial, and Jacobi polynomial is given in Sec. 2.

Section 3 is devoted to special functions that are special cases of the confluent hypergeometric function, ${}_1F_1$. These special functions include the various Bessel functions, Laguerre and associated Laguerre polynomials, and the Hermite polynomials. Also, discussions of some properties of the confluent hypergeometric differential equation and properties of confluent hypergeometric function are given in Sec. 3.

In Sec. 4, a comprehensive treatment of the Helmholtz differential equation in cylindrical coordinates and the resulting Bessel functions is given. A summary of certain other special functions used in mathematical physics is given in Sec. 5.

1 The Sturm–Liouville Theory

1.1 Introduction

Linear operators are basic to linear differential equations, and the solutions of each of the differential equations we will consider form a vector space. Hence, we begin with the essentials of the theory of linear operators.

A vector space (also known as a linear space or linear manifold) V is a nonempty set of elements $\{\psi_i\}$, called vectors (here vector is used in an abstract mathematical sense), for which the operations of addition and multiplication by a scalar are valid. Addition is a rule that assigns an element for $\psi_1 + \psi_2$ in V for every pair of elements ψ_1 and ψ_2 in V. The operation of multiplication by a scalar is a rule that assigns an element for $a\psi_1$ in V to each complex a and each ψ_1 in V. In addition, a zero element and negative elements exist, and the associative and commutative properties of addition are valid.

A function T that transforms (maps) vectors in V into a vector space W, $T : V \to W$, is called a linear transformation if $T(\psi_1 + \psi_2) = T(\psi_1) + T(\psi_2)$ and $T(a\psi_1) = aT(\psi_1)$ are valid for all vectors and all real scalars in V. The case $T : V \to V$ is called a linear operator on V. A real number λ is an eigenvalue (characteristic value) of the linear operator T if there is a nonzero vector ψ (called the eigenvector of T) in V for which $T(\psi) = \lambda\psi$.

In mathematical physics, the linear operator is normally a differential operator, the eigenvalue equation is a differential equation, and the eigenfunctions (solutions) form a vector space and satisfy certain imposed boundary conditions.

For physical problems, the three basic types of boundary conditions are

1. Dirichlet, specification of the solution at each point on the boundary;
2. Neumann, specification of the normal derivative of the solution at each point on the boundary; and
3. Cauchy, specification of both Dirichlet and Neumann conditions.

Typically, the actual physical problem is used as a guide for the formulation of boundary conditions. Sometimes, however, it is difficult to formulate the appropriate boundary conditions for a problem. Hence, it is important to understand what conditions are appropriate for a particular type of differential equation (see, e.g., ANALYTIC METHODS).

The Laplace equation, the time-independent heat-conduction (diffusion) equation, and the time-independent mechanical wave equation may be put into the form of the Helmholtz differential equation. The Helmholtz differential equation, $\nabla^2 u + k^2 u = 0$ for constant k^2, is an eigenvalue equation where $-k^2$ is the eigenvalue, u is the eigenfunction and is subject to boundary conditions, and ∇^2 is the operator. The time-independent Schrödinger wave equation contains the operator ∇^2 and is an eigenvalue equation of the form $\hat{H}\psi = E\psi$ where $\hat{H} = (-\hbar^2/2m)\nabla^2 + V(x, y, z)$.

The separation-of-variables method applied to Helmholtz's differential equation and to the time-independent Schrödinger's wave equation in various coordinate systems leads to ordinary differential equations that may be written in the following general form:

$$\frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u + \lambda\rho(x)u = 0. \tag{1}$$

The parameter λ is a separation constant (in some cases, more than one separation constant may appear; we will focus on the case of one separation constant). Equation (1) is the well-known Sturm–Liouville equation, and it may be written in the following operator form:

$$\mathscr{L}(u) + \lambda\rho(x)u = 0. \tag{2}$$

The Liouville operator, a linear operator, is defined by

$$\mathscr{L}(u) = \frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u. \tag{3}$$

For the general differential operator $M(u) = p(x)u'' + r(x)u' + q(x)u$, the operator $\bar{M}(u) = (pu)'' - (ru)' + qu$ is defined as the adjoint of $M(u)$. Note that $M(u) = \bar{M}(u)$ when $p' = r$, and $M(u)$ is said to be a self-adjoint operator in this case. On applying the general definition for the adjoint of an operator to Eq. (3), we find that $\mathscr{L}(u) = \bar{\mathscr{L}}(u)$, which means that the Liouville operator is a self-adjoint operator. In fact, it can be shown that every second-order differential operator can be transformed to the self-adjoint

form (see Courant and Hilbert, 1953, p. 279).

In the Sturm–Liouville equation, the function $\rho(x)$ [$w(x)$ and $r(x)$ are also used] is called the density or weight. This name for $\rho(x)$ is related to the historical origin of the Sturm–Liouville equation, which involved finding the solution for the one-dimensional mechanical wave equation, $[p(x)u_x]_x = \rho(x)u_{tt}$, representing the motion of a nonhomogeneous string. In the mechanical wave equation, $u(x,t)$ is the displacement of the string from its equilibrium position, $p(x)$ is proportional to the modulus of elasticity, and $\rho(x)$ is the mass per unit length of the string. Separation of variables leads to the ordinary differential equations $(pX')' + \lambda\rho X = 0$ and $\ddot{T} + \lambda T = 0$, where λ is the separation constant and $u(x,t)$ is assumed to equal the product $X(x)T(t)$, with typical boundary conditions given by $X(a) = X(b)$ and $p(a)X'(a) = p(b)X'(b)$.

In Eq. (1), the functions $p(x)$, $q(x)$, and $\rho(x)$ are assumed to be real, continuous with continuous derivatives, and nonzero in the region of interest, $[a,b]$. Moreover, it is assumed that $p(x)$ and $\rho(x)$ are always positive in $[a,b]$. The Sturm–Liouville equation is a generalized form of the usual eigenvalue equation since the eigenvalue is multiplied by the density function $\rho(x)$, which may be different from unity. The sign convention for $q(x)$ in Eqs. (1) and (3) conforms to the usage of Courant and Hilbert; some authors use a plus sign for $q(x)$ in these equations. The function $u(x)$ is subject to appropriate boundary conditions. With appropriate substitutions, the following differential equations are among the list of important differential equations in mathematical physics that may be put in the Sturm–Liouville form: Legendre and associated Legendre, Laguerre and associated Laguerre, Schrödinger equation for the linear harmonic oscillator, and Bessel. Hence, a study of the general properties of the Sturm–Liouville equation is extremely useful in mathematical physics. A summary of the relations between the differential equations for many important special functions and the Sturm–Liouville equation is given in Table 1.

Tab. 1 Relation to the Sturm–Liouville equation

$$\frac{d}{dx}\left\{p(x)\frac{du}{dx}\right\} - q(x)u + \lambda\rho(x)u = 0$$

Equation	$p(x)$	$q(x)$	$\rho(x)$	λ
Legendre, $P_n(x)$	$1-x^2$	0	1	$n(n+1)$
Associated Legendre, $P_n^m(x)$	$1-x^2$	$\frac{m^2}{1-x^2}$	1	$n(n+1)$
Laguerre, $L_n(x)$	xe^{-x}	0	e^{-x}	n
Associated Laguerre, $L_n^k(x)$	$x^{k+1}e^{-x}$	0	$x^k e^{-x}$	$n-k$
Bessel, $J_n(x)$, $Y_n(x)$, $H_n^{(1)}, \ldots$	x	n^2/x	x	1
Hermite, $H_n(x)$	e^{-x^2}	0	e^{-x^2}	$2n$
Quantum oscillator, $\psi_n(x)$	1	x^2	1	λ
Jacobi, $P_n^{(\alpha,\beta)}(x)$; $\alpha, \beta > -1$	$\frac{1-x^2}{(1-x)^{-\alpha}(1+x)^{-\beta}}$	0	$\frac{(1-x)^{\alpha}}{(1+x)^{-\beta}}$	$n(n+\alpha+\beta+1)$
Chebyshev, $T_n(x)$	$(1-x^2)^{1/2}$	0	$(1-x^2)^{-1/2}$	n^2
Gegenbauer, $C_n^{(\alpha)}(x)$; $\alpha > -\frac{1}{2}$	$(1-x^2)^{\alpha+1/2}$	0	$(1-x^2)^{\alpha-1/2}$	$n(n+2\alpha)$

The problem of determining the dependence of the eigenfunction $u(x)$ on the eigenvalue λ and of the eigenvalue λ on the boundary conditions imposed on $u(x)$ is often referred to as the Sturm–Liouville problem. The Sturm–Liouville problem is important for problems in both classical and quantum theory. Sturm–Liouville theory unites properties of the solutions of second-order ordinary linear differential equations related to

1. Hermitian and self-adjoint operators;
2. reality of eigenvalues of Hermitian and self-adjoint operators;
3. orthogonality and completeness of eigenfunctions;
4. degeneracy of eigenvalues (if N linearly independent eigenfunctions correspond to the same eigenvalue, then the eigenvalue is said to be N-fold degenerate); and
5. the fact that eigenvalues of the Sturm–Liouville equation form a discrete set of values such that $\ldots \lambda_1 \leq \lambda_2 \leq \lambda_3 \ldots$.

These properties are important in the study of problems that lead to each of the differential equations we analyze in this article.

1.2 Hermitian Operators and Their Eigenvalues

Consider two twice-differentiable functions u_i and u_j. By use of Eqs. (1) and (2), we obtain

$$u_i^* \mathscr{L}(u_j) - [\mathscr{L}(u_i)]^* u_j = \frac{d}{dx}\left\{p\left(u_i^* \frac{du_j}{dx} - u_j \frac{du_i^*}{dx}\right)\right\}. \quad (4)$$

In Eq. (4), the asterisk is used to denote complex conjugate of the respective functions. Integrating both sides of Eq. (4) over the range of interest yields

$$\int_a^b \{u_i^* \mathscr{L}(u_j) - [\mathscr{L}(u_i)]^* u_j\} dx = \left\{p\left(u_i^* \frac{du_j}{dx} - u_j \frac{du_i^*}{dx}\right)\right\}_{x=b} - \left\{p\left(u_i^* \frac{du_j}{dx} - u_j \frac{du_i^*}{dx}\right)\right\}_{x=a}.$$

Note that the above equation results from the fact that $\mathscr{L}$ is self-adjoint. The operator $\mathscr{L}$ is said to be Hermitian if the following end-point boundary conditions are imposed on the two functions and their derivatives.

$$\left\{p\left(u_i^* \frac{du_j}{dx} - u_j \frac{du_i^*}{dx}\right)\right\}_{x=b} = \left\{p\left(u_i^* \frac{du_j}{dx} - u_j \frac{du_i^*}{dx}\right)\right\}_{x=a}. \quad (5)$$

By use of the boundary conditions in Eq. (5), the Hermitian relation may be written as

$$\int_a^b u_i^* \mathscr{L}(u_j) dx = \int_a^b [\mathscr{L}(u_i)]^* u_j \, dx. \quad (6)$$

Thus far, the Liouville operator has been assumed to be real. In quantum mechanics, operators are generally complex (for example, the x component of the linear momentum operator is given by $p_x = -i\hbar\partial/\partial x$), and it is assumed that wave functions satisfy the boundary conditions in Eq. (5). The Hermitian condition in quantum mechanics for linear operator $\hat{A}$ takes the form

$$\int_{\text{all space}} \psi_i^* \hat{A} \psi_j d\tau = \int_{\text{all space}} (\hat{A}\psi_i)^\dagger \psi_j d\tau. \quad (7)$$

An arbitrary linear operator may be put in matrix form, and the notation $\hat{A}^\dagger$ means

1. interchange rows with columns and
2. take the complex conjugate of each element

(this process is called the Hermitian conjugate); in this connection, note that $(\hat{A}\psi_i)^\dagger = \psi_i^* \hat{A}^\dagger$. When an operator satisfies the condition $\hat{A} = \hat{A}^\dagger$, the operator is said to be Hermitian. In the bra and ket vector notation, Eq. (7) becomes $\langle \psi_i | \hat{A}\psi_j \rangle = \langle \hat{A}\psi_i | \psi_j \rangle$.

For solution $u_i = u_j = u$ in Eq. (6) and use of Eq. (2), we obtain

$$\int_a^b [u^* \mathscr{L}(u) - u\mathscr{L}(u^*)]dx = (\lambda - \lambda^*) \int_a^b \rho(x) u^* u dx = 0. \quad (8)$$

The result in Eq. (8) means that eigenvalues of Hermitian operators are real, $\lambda = \lambda^*$.

1.3 Orthogonality Condition and Completeness of Eigenfunctions

By use of Eq. (2) for distinct eigenfunctions u_i and u_j with distinct eigenvalues, the Hermitian relation in Eq. (6) may be written as

$$\int_a^b \{u_i^* \mathscr{L}(u_j) - [\mathscr{L}(u_i)]^* u_j\} dx = \int_a^b [u_i^*(-\lambda_j \rho u_j) + \lambda_i^* \rho u_i^* u_j] dx = (\lambda_i^* - \lambda_j) \int_a^b u_i^* u_j \rho(x) dx = 0.$$

Since $\lambda_i \neq \lambda_j$ and λ_i is real, the above equation implies that

$$\int_a^b u_i^*(x) u_j(x) \rho(x) dx = 0. \quad (9)$$

Equation (9) shows that eigenfunctions corresponding to distinct eigenvalues are orthogonal in the interval $[a,b]$ with respect to the weight function $\rho(x)$.

An orthonormal set of Sturm–Liouville eigenfunctions, $\{u_k(x)\}$, forms a complete set of functions (Courant and Hilbert, 1953, Chap. 6, Sec. 3). This completeness property means that the following equation is valid for any function $f(x)$ that is at least piecewise continuous in the interval $[a,b]$:

$$\lim_{n\to\infty} \int_a^b \left| f(x) - \sum_{k=1}^n c_k u_k(x) \right|^2 \rho(x) dx = 0. \quad (10)$$

The notation $|\ldots|^2$ means the product of the enclosed quantity and its complex conjugate, $|z|^2 = z^* z$. Equation (10) and the orthogonality relation in Eq. (9) lead to

$$f(x) = \sum_{k=1}^\infty c_k u_k(x) \quad \text{where } c_k = \int_a^b f(x) u_k^*(x) \rho(x) dx \quad (11)$$

and

$$\sum_{k=1}^\infty |c_k|^2 = \int_a^b |f(x)|^2 \rho(x)\, dx. \quad (12)$$

Equation (12) is referred to as the completeness relation. By use of Eq. (11) with the appropriate orthogonality relation and weight, one may obtain series expansions for $f(x)$ in terms of any complete set of orthogonal polynomials (or orthogonal functions); for example, the Fourier series, Legendre series, and Hermite series may be written respectively as

$$f(x) = \sum_{n=-\infty}^\infty c_n e^{in\pi x/l}$$

where

$$c_n = \frac{1}{2l}\int_{-l}^{l} f(x)e^{-in\,\pi x/l}\,dx, \quad -l \le x \le l;$$

$$f(x) = \sum_{n=0}^{\infty} c_n P_n(x)$$

where

$$c_n = \frac{2n+1}{2}\int_{-1}^{1} f(x)P_n(x)\,dx,$$
$$-1 \le x \le 1;$$

and

$$f(x) = \sum_{n=0}^{\infty} c_n H_n(x)$$

where

$$c_n = \frac{1}{2^n n!\sqrt{\pi}}\int_{-\infty}^{\infty} f(x)H_n(x)e^{-x^2}\,dx,$$
$$-\infty \le x \le \infty.$$

1.4 Orthogonal Polynomials and Functions

The power-series solutions of many second-order ordinary linear differential equations in mathematical physics such as the Legendre, Laguerre, and Hermite differential equations may be written as an orthogonal polynomial plus an orthogonal function. The set of real polynomials $\{f_n(x)\}$ is said to be orthogonal with respect to the weight function $\rho(x)$ over the interval $[a,b]$ if the following relations are valid:

$$\int_a^b \rho(x)f_n(x)f_m(x)dx = \begin{cases} 0 & m \ne n \\ h_n^2 & m = n \end{cases} \tag{13}$$

If $h_n^2 = 1$ for all n, then the system of orthogonal polynomials in Eq. (13) is said to be orthonormal. Note that the weight determines the system of polynomials up to a constant factor (the value of h_n) in each polynomial; the specification of this constant factor for each polynomial is referred to a standardization (or standard convention). Summarized in Table 2 are some frequently used orthogonality relations.

Tab. 2 Some orthogonality relations

$\int_{-1}^{1} P_n(x)P_k(x)dx = \frac{2}{2n+1}\delta_{nk}$
$\int_{-1}^{1} P_n^m(x)P_k^m(x)dx = \frac{2}{2n+1}\frac{(n+m)!}{(n-m)!}\delta_{nk}$
$\int_0^{\infty} e^{-x}L_n(x)L_m(x)dx = \delta_{nm}$
$\int_0^{\infty} e^{-x}x^k L_n^k(x)L_m^k(x)dx = \frac{(n!)^3}{(n-k)!}\delta_{nm}$
$\int_{-\infty}^{\infty} e^{-x^2}H_n(x)H_m(x)dx = 2^n n!\sqrt{\pi}\delta_{nm}$

1.4.1 Recurrence Formula for Orthogonal Polynomials

In many cases, three consecutive orthogonal polynomials f_{n+1}, f_n, and f_{n-1} satisfy a recurrence formula of the form

$$A_n f_{n+1}(x) = (B_n + C_n x)f_n(x) - D_n f_{n-1}(x). \tag{14}$$

The recurrence relations for some of the frequently used orthogonal polynomials are listed in Table 3.

1.4.2 Rodrigues Formulas

Formulas involving the nth derivative of an elementary function that can be used to generate orthogonal polynomials

Tab. 3 Recurrence relations for some polynomials
$A_n f_{n+1}(x) = (B_n + C_n x)f_n(x) - D_n f_{n-1}(x)$.

$f_n(x)$	A_n	B_n	C_n	D_n
$P_n(x)$	$n+1$	0	$2n+1$	n
$P_n^m(x)$	$n-m+1$	0	$2n+1$	$n+m$
$L_n(x)$	$n+1$	$2n+1$	-1	n
$L_n^k(x)$	$n+1$	$2n+k+1$	-1	$n+k$
$H_n(x)$	1	0	2	$2n$

f_n are called Rodrigues formulas. It can be shown that the Rodrigues formulas for many orthogonal polynomials may be combined into the following general Rodrigues formula:

$$f_n = \frac{1}{a_n \rho(x)} \frac{d^n\{\rho(x)[g(x)]^n\}}{dx^n}. \tag{15}$$

The quantity $g(x)$ is a polynomial whose coefficients are independent of n, and the factor a_n is determined by the standardization of the orthogonal polynomial system. The Leibnitz formula for the nth derivative of a product should be used to evaluate the right-hand side of Eq. (15); this formula is

$$\frac{d^n}{dx^n}\{A(x)B(x)\} = \sum_{s=0}^{n} \left\{ \frac{n!}{(n-s)!s!} \frac{d^{n-s}}{dx^{n-s}} A(x) \frac{d^s}{dx^s} B(x) \right\}. \tag{16}$$

The Rodrigues formulas for some important orthogonal polynomials are given in Table 4.

1.4.3 The Generating Function

The function $\mathscr{G}(x,t)$ is said to be a generating function of the sequence of functions $\{f_n(x)\}$ if the $f_n(x)$ are, up to a constant, the coefficients of t^n in the expansion of $\mathscr{G}(x,t)$ in powers of t. Such an expansion is valid for orthogonal polynomials and most orthogonal functions, and it may be written in the form

$$\mathscr{G}(x, t) = \sum_{n=0}^{\infty} a_n f_n(x) t^n. \tag{17}$$

In Eq. (17), the a_n are independent of x and t. The generating functions for certain orthogonal polynomials that will be used in other parts of this article are given in Table 5.

Tab. 5 Some generating functions

$$\mathscr{G}(x, t) = \sum_{n=0}^{\infty} a_n f_n(x) t^n;\ R = \sqrt{1 - 2xt + t^2}.$$

$f_n(x)$	a_n	$\mathscr{G}(x,t)$
$P_n(x)$	1	R^{-1}
$L_n(x)$	1	$(1-t)^{-1}\exp\{-xt/(1-t)\}$
$L_n^k(x)$	1	$(1-t)^{-(k+1)}\exp[xt/(t-1)]$
$H_n(x)$	$\frac{1}{n!}$	$\exp(2xt - t^2)$
$P_n^{(\alpha,\beta)}(x)$	$2^{-\alpha-\beta}$	$R^{-1}(1-t+R)^{-\alpha}(1+t+R)^{-\beta}$
$C_n^{(\alpha)}(x)$	1	$R^{-2\alpha}$
$T_n(x)$	2	$1 + (1-t^2)/R^2$

Tab. 4 Rodrigues formula for some polynomials

$$f_n = \frac{1}{a_n \rho(x)} \frac{d^n\{\rho(x)[g(x)]^n\}}{dx^n}.$$

$f_n(x)$	a_n	$g(x)$	$\rho(x)$
$P_n(x)$	$2^n n!$	$x^2 - 1$	1
$L_n^k(x)$	$n!$	x	$x^k e^{-x}$
$H_n(x)$	$(-1)^n$	1	e^{-x^2}
$P_n^{(\alpha,\beta)}(x)$	$(-1)^n 2^n n!$	$1 - x^2$	$(1-x)^\alpha(1+x)^\beta$
$C_n^{(\alpha)}(x)$	$(-1)^n 2^n n! \frac{\Gamma(2\alpha)\Gamma(\alpha+n+\frac{1}{2})}{\Gamma(\alpha+\frac{1}{2})\Gamma(n+2\alpha)}$	$1 - x^2$	$(1-x^2)^{\alpha-1/2}$
$T_n(x)$	$(-1)^n 2^{n+1} \frac{\Gamma(n+\frac{1}{2})}{\sqrt{\pi}}$	$1 - x^2$	$(1-x^2)^{-1/2}$

1.5 Factorization of the Sturm–Liouville Equation

The factorization method for finding eigenfunctions and corresponding eigenvalues of a large class of Schrödinger-type equations (forms of the Sturm–Liouville equation) was introduced by Schrödinginer (1940); in a clearly written paper, Infeld and Hull (1951) further developed the method. By use of the factorization method, a second-order differential equation is factored (transformed) into a product of first-order differential operators, which results in a pair of first-order differential equations that are equivalent to the original second-order differential equation. The form of the potential function determines if the factorization method will be successful. In terms of the forms of the potential function, Infeld and Hull classified factorizations into six general factorization types. Many examples of the various factorization types are presented in the paper by Infeld and Hull. We give a brief overview of the factorization method because it is

1. a useful and straightforward method for finding eigenfunctions (in many cases, involving special functions) and corresponding eigenvalues for important differential equations in mathematical physics, and
2. equivalent to a local Lie-group theoretical analysis involving differential operators that leads to a group-theoretical approach to special functions.

In this overview of the factorization method, we closely follow the Infeld–Hull approach; using a plus sign for the second term in the Sturm–Liouville equation, we find that it may be transformed (see, e.g., Courant and Hilbert, 1953, p. 292) into the following general differential equation in standard form:

$$\frac{d^2y}{dx^2} + r(x, m)y + \lambda y = 0;\ m = 0, 1, 2, \ldots. \tag{18}$$

In Eq. (18), the function $r(x,m)$ contains the dynamical information, such as the potential energy, that characterizes the particular problem. The basic idea of the factorization method is that Eq. (18) can be either factored directly as a single differential equation or factored into a pair of differential equations of the following forms:

$$H^{+}_{m+1}H^{-}_{m+1}y(\lambda, m) = [\lambda - L(m+1)]\, y(\lambda, m) \tag{19}$$

and

$$H^{-}_{m}H^{+}_{m}y(\lambda, m) = [\lambda - L(m)]y(\lambda, m), \quad \text{where } H^{\pm}_{m} = k(x, m) \pm \frac{d}{dx}. \tag{20}$$

The condition that must be satisfied by $k(x,m)$ and $L(m)$ is obtained by carrying out the indicated operations in Eqs. (19) and (20), comparing the result with the original differential equation, and eliminating $r(x,m)$; this condition is

$$k^2(x, m+1) - k^2(x, m) + \frac{dk(x, m+1)}{dx} + \frac{dk(x, m)}{dx} = L(m) - L(m+1). \tag{21}$$

The fundamental idea of the factorization method is established in the five theorems that we now summarize.

THEOREM I: If $y(\lambda, m)$ is a solution of the original differential equation, then $y(\lambda, m+1) = H^{-}_{m+1}y(\lambda, m)$ and $y(\lambda, m-1) = H^{+}_{m}y(\lambda, m)$ are also solutions corresponding to the same λ.

THEOREM II: (mutual adjointness of the operators): For functions g and f satisfying the end-point conditions,

$$\int_a^b g(H_m^- f)dx = \int_a^b (H_m^+ g)f\,dx.$$

THEOREM III: (satisfying the boundary conditions): If $y(\lambda, m)$ is quadratically integrable over the entire range of x and $L(m)$ is an increasing function of $m(0 < m)$, then the raising operation, $y(\lambda, m+1) = H_{m+1}^- \times y(\lambda, m)$, produces a function that is also quadratically integrable and vanishes at the end points. If $L(m)$ is a decreasing function of $m(0 < m)$, then the lowering operation, $y(\lambda, m-1) = H_m^+ y(\lambda, m)$, produces a function that is also quadratically integrable and vanishes at the end points.

THEOREM IV: Class I solutions – $L(m)$ is an increasing function of the integer m. If $0 < m \leq M$ and $\lambda \leq$ the larger of $L(M)$ and $L(M+1)$, then a necessary condition for a square integrable solution is that $\lambda = \lambda_l = L(l+1)$, where l is an integer, and $m = 0, 1, \ldots, l$ and $H_{l+1}^- y(\lambda, l) \equiv 0$. Other solutions are given by

$$y_l^{m-1} = [L(l+1) - L(m)]^{-1/2} \times \left[k(x, m) + \frac{d}{dx}\right] y_l^m.$$

Class II solutions – $L(m)$ is a decreasing function of the integer m. If $0 \leq m \leq M$ and $\lambda \leq L(0)$, then a necessary condition for square integrable solutions is that $\lambda = \lambda_l = L(l)$, where l is an integer, and $m = l, l+1, \ldots$ and $H_l^+ y(\lambda, l) \equiv 0$. Other solutions are given by

$$y_l^{m+1} = [L(l) - L(m+1)]^{-1/2} \times \left[k(x, m+1) - \frac{d}{dx}\right] y_l^m.$$

THEOREM V: (normalization of solutions): The $H^{\pm}$ operators preserve the normalization of the eigenfunctions.

The basic task now is to find $k(x,m)$ and $L(m)$ corresponding to a given $r(x,m)$. Infeld and Hull have shown that finding $k(x,m)$ and $L(m)$ is accomplished by use of one of the following six operator types.

Type A solutions are related to spherical harmonic functions and other eigenfunctions that are special cases of the hypergeometric function, ${}_2F_1$; we have

$$r(x, m) = -\frac{a^2(m+c)(m+c+1) + d^2 + 2ad(m+c+\frac{1}{2})\cos a(x+p)}{\sin^2 a(x+p)},$$

$$k(x, m) = a(m+c)\cot a(x+p) + \frac{d}{\sin a(x+p)},$$

and

$$L(m) = a^2(m+c)^2$$

where a, c, d, and p are constants.

Type B solutions are related to associated Laguerre and Laguerre functions and other eigenfunctions that are special cases of the confluent hypergeometric function, ${}_1F_1$; here

$$r(x, m) = -d^2 e^{2ax} + 2ad(m+c+\tfrac{1}{2})e^{ax},$$

$$k(x, m) = de^{ax} - m - c,$$

and

$$L(m) = -a^2(m+c)^2.$$

Type C solutions are related to confluent hypergeometric functions, with

$$r(x, m) = -\frac{(m+c)(m+c+1)}{x^2} - \frac{b^2x^2}{4} + b(m-c),$$

$$k(x, m) = \frac{m+c}{x} + \frac{bx}{2},$$

and

$$L(m) = -2bm + \frac{b}{2}.$$

Type D solutions are related to generalization of the Hermite polynomials; here

$$r(x, m) = -(bx + d)^2 + b(2m + 1),$$
$$k(x, m) = bx + d,$$

and

$$L(m) = -2bm.$$

Type E solutions are related to hypergeometric functions, with

$$r(x, m) = -\frac{m(m+1)a^2}{\sin^2 a(x+p)} - 2aq \cot a(x+p),$$
$$k(x, m) = ma \cot a(x+p) + \frac{q}{m},$$

and

$$L(m) = a^2 m^2 - \frac{q^2}{m^2}.$$

Type F solutions are related to Laguerre polynomials, with

$$r(x, m) = -\frac{2q}{x} - \frac{m(m+1)}{x^2},$$
$$k(x, m) = \frac{m}{x} + \frac{q}{m},$$

and

$$L(m) = -\frac{q^2}{m^2}.$$

EXAMPLE 1: Find the eigenvalues and eigenfunctions of the associated Legendre differential equation by use of the factorization method.

Solution: By use of the transformation $u = P \sin^{1/2} \theta$, the associated Legendre differential equation [the theta equation in spherical coordinates, Eq. (49)],

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta \frac{dP}{d\theta}\right) + \left[\lambda - \frac{m^2}{\sin^2\theta}\right] P = 0, \qquad (22)$$

reduces to

$$\frac{d^2u}{d\theta^2} - \frac{m^2 - \frac{1}{4}}{\sin^2\theta} u + (\lambda + \tfrac{1}{4})u = 0. \qquad (23)$$

On comparing Eq. (23) with the general equation, Eq. (18), in standard form, we find that it is a Type A factorization (see Theorem IV) where the parameters are given by $a = 1, c = -\frac{1}{2}, d = 0, p = 0, x = \theta$, and $\lambda \Rightarrow \lambda + \frac{1}{4}$; the required quantities for Type A factorization reduce to $r(\theta, m) = -(m^2 - \frac{1}{4})/\sin^2\theta$, $L(m) = (m - \frac{1}{2})$, and $k(\theta, m) = (m - \frac{1}{2})\cot\theta$. Since this is a Class I problem, $\lambda + \frac{1}{4} = L(l+1)$ or $\lambda = l(l+1)$ for $l = 0, 1, 2, \ldots l \geq m$. The corresponding eigenfunctions (solutions) of Eq. (23) are obtained by use of Theorem IV for Class I solutions, $H_l^+ \gamma(\lambda, l) \equiv 0$; for normalization constants C_l and D_{lm}, the solutions are given by

$$u_l^l = C_l \sin^{(l+1/2)}\theta \qquad \text{and}$$
$$u_l^{m-1} = D_{lm}\left\{(m - \tfrac{1}{2})\cot\theta + \frac{d}{d\theta}\right\} u_l^m.$$

EXAMPLE 2: By use of a modified factorization method, find the eigenvalues and eigenfunctions of the quantum mechanical linear harmonic oscillator.

Solution: The linear harmonic oscillator problem was used by Schrödinger in his original work on factorization; it is a Type D problem, but $r(x,m)$ does not depend on m. For the linear harmonic oscillator, it is simpler to use the modified version of factorization involving the idea

of raising and lowering operators; this modified factorization method is used in many quantum mechanics textbooks and is summarized here. The time-independent Schrödinger equation for the linear harmonic oscillator with potential energy $V(x) = kx^2/2 = m\omega^2 x^2/2$ is given by

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + \frac{m\omega^2 x^2}{2}\psi = E\psi. \qquad (24)$$

The change of variable $\xi = (m\omega/k)^{1/2}x$ reduces Eq. (24) to the dimensionless form

$$\frac{d^2\psi}{d\xi^2} - \xi^2\psi + \lambda\psi = 0, \quad \text{where } \lambda = \frac{2E}{\hbar\omega}. \qquad (25)$$

By use of the factorization (of differential operators) method, Eq. (25) may be written as

$$H^+H^-\psi = (\lambda+1)\psi \text{ and}$$
$$H^-H^+\psi = (\lambda-1)\psi,$$
$$\text{where } H^\pm = \xi \pm \frac{d}{d\xi}. \qquad (26)$$

The above factored equations are equivalent to Eq. (25). The basic idea of this factorization method is that ψ_n and λ_n can be generated from known ψ_0 and λ_0. If ψ_0 satisfies the original equation (factored equations) Eq. (26) with eigenvalue λ_0, then we may write

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi_0 = (-\lambda_0 + 1)\psi_0. \qquad (27)$$

By use of the identity

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi$$
$$= 2\psi - \left(\frac{d}{d\xi} + \xi\right)\left(\xi - \frac{d}{d\xi}\right)\psi,$$

Eq. (27) reduces to

$$\left(\frac{d}{d\xi} - \xi\right)\left(\frac{d}{d\xi} + \xi\right)\psi_1 = \lambda_1\psi_1. \qquad (28)$$

To obtain Eq. (28), we made the following substitutions:

$$\psi_1(\xi) = \left(\xi - \frac{d}{d\xi}\right)\psi_0 \text{ and } \lambda_1 = \lambda_0 + 1. \qquad (29)$$

In a similar manner, we obtain

$$\psi_{-1}(\xi) = \left(\xi + \frac{d}{d\xi}\right)\psi_0 \text{ and}$$
$$\lambda_{-1} = \lambda_0 - 1. \qquad (30)$$

The operators in Eqs. (29) and (30) are referred to as raising and lowering operators, respectively. If the above process is repeated n times, we obtain

$$\psi_n(\xi) = \left(\xi - \frac{d}{d\xi}\right)^n \psi_0 \text{ for } \lambda_n = \lambda_0 + n \qquad (31)$$

and

$$\psi_{-n}(\xi) = \left(\xi + \frac{d}{d\xi}\right)^n \psi_0 \text{ for}$$
$$\lambda_{-n} = \lambda_0 - n. \qquad (32)$$

As seen from Eqs. (31) and (32), the eigenvalue (eigenenergy) is bounded from below at λ_0, but it is an arbitrarily large negative value, as seen in Eq. (32). It is required that the negative values of λ_{-n} be terminated (Theorem IV) by setting $\psi_{-1} = 0 = (\xi - d/d\xi)\psi_0$; by use of this termination, we find that $\psi_0 = N\exp(-\xi^2/2)$ is a solution of the original differential equation, Eq. (25). The corresponding lower bound for the eigenvalue is $\lambda_0 = 1$ or $E_0 = \hbar\omega/2$. Hence, we find that the knowledge of ψ_0 and λ_0 allows us to find ψ_n and λ_n; we obtain

$$\psi_n(\xi) = N_n H_n(\xi)e^{-\xi^2/2} \text{ for } \lambda_n = 1 + n \text{ or}$$
$$E_n = \hbar\omega(n + \tfrac{1}{2}). \qquad (33)$$

Equation (33) results from the fact that $(\xi - d/d\xi)^n \exp(-\xi^2/2)$ produces exp $(-\xi^2/2)$ times an nth-order polynomial, $H_n(\xi)$; this polynomial, as is shown in Sec. 2.2, is the Hermite polynomial.

1.6 Connection with Local Lie-Group Theory

Wigner, unpublished Princeton University Lecture Notes in 1955, demonstrated that large classes of special functions arise as matrix elements of the representations of Lie groups such as the groups of rotations in two, three, and four dimensions or the Euclidean groups in two and three dimensions. Talman (1968) extended the work of Wigner. The Wigner group-theoretical approach to special functions shows that

1. addition theorems for special functions result from the group multiplication rule and
2. differential equations giving rise to special functions result from limits of generators.

Other group-theoretical development of special functions exist (see, e.g., Vilenkin, 1968).

The focus in this section is on the approach by Miller (1968), which is based on the recognition of the equivalence between the Infeld–Hull factorization method and the representation theory of complex local Lie groups with the four-dimensional Lie algebras $\mathscr{G}(a,b)$ and the six-dimensional Lie algebra $\mathscr{F}_6$. Local Lie-group theory provides a unifying approach to a large class of special functions and their properties. A comprehensive treatment of group theory is given in GROUP THEORY and in the many excellent books on the subject; we simply summarize essential group concepts needed to outline the Miller group-theoretical approach to special functions.

Local Lie-group theory was developed in the nineteenth century; it is based on the use of local coordinates (needed for a discussion of special functions), and concerned with groups that are analytic only in the neighborhood of the group identity element. A global Lie group involves coordinate-free considerations and is an abstract group.

The Lie algebra (vector space) $\mathscr{G}$ of the local Lie group G is the set of all tangent vectors (generators) at the identity element together with commutation relations $[\alpha, \beta] \in \mathscr{G}$ defined for all tangent vectors $\alpha, \beta \in \mathscr{G}$. Lie's three fundamental theorems and their converses together with the Taylor expansion theorem provide a mechanism for constructing the Lie algebra associated with a Lie group. For local transformation groups, which are important in our analysis, the commutation relations involve Lie derivatives, $[L_\alpha, L_\beta]$, or generalized Lie derivatives, $[D_\alpha, D_\beta]$.

A four-dimensional complex Lie algebra $\mathscr{G}(a,b)$ with basis $\mathscr{J}^\pm, \mathscr{J}^3$, and $\mathscr{E}$ is defined by the commutation relations

$$\begin{aligned} &[\mathscr{J}^+, \mathscr{J}^-] = 2a\mathscr{J}^3 - b\mathscr{E}; \\ &[\mathscr{J}^3, \mathscr{J}^\pm] = \pm\mathscr{J}^\pm; \text{ and} \\ &[\mathscr{J}^\pm, \mathscr{E}] = [\mathscr{J}^3, \mathscr{E}] = 0. \end{aligned} \tag{34}$$

The parameters a and b are a pair of arbitrary complex numbers. For the representation ρ of $\mathscr{G}(a,b)$ on the complex vector space V, we define four operators as follows:

$$\begin{aligned} &J^+ = \rho(\mathscr{J}^+); J^- = \rho(\mathscr{J}^-); \\ &J^3 = \rho(\mathscr{J}^3); \text{ and} \\ &E = \rho(\mathscr{E}). \end{aligned} \tag{35}$$

The four operators in Eq. (35) obey commutation relations similar to those in Eq. (34); they are

$$[J^+, J^-] = 2aJ^3 - bE; [J^3, J^\pm] = \pm J^\pm; \text{ and}$$
$$[J^\pm, E] = [J^3, E] = 0. \tag{36}$$

The object at this stage is that of finding realizations of the irreducible representations $\rho(\alpha)$ of $\mathscr{G}(a,b)$ [where $\alpha \in \mathscr{G}(a, b)$] that form a Lie algebra of analytic differential operators acting on V. An extremely large number of possible solutions of Eq. (36) exists. By use of generalized Lie derivatives and a classification of all such solutions by differential operators in two variables, Miller (1968, Chap. 8) has shown that only generalized forms of the angular momentum operators are important in special function theory. We therefore write

$$J^3 = \frac{\partial}{\partial y};$$
$$J^\pm = e^{\pm y}\left[\pm\frac{\partial}{\partial x} - k(x)\frac{\partial}{\partial y} + j(x)\right]; \text{ and}$$
$$E = \mu. \tag{37}$$

In Eq. (37), μ is a complex constant; the functions $k(x)$ and $j(x)$ are to be determined. The equations in Eq. (37) satisfy all the commutation conditions for the operators in Eq. (36), except $[J^+, J^-] = 2aJ^3 - bE$, which is satisfied when $k(x)$ and $j(x)$ are solutions of the differential equations

$$\frac{dk(x)}{dx} + k(x)^2 = -a^2 \text{ and}$$
$$\frac{dj(x)}{dx} + k(x)j(x) = -\frac{d\mu}{2}. \tag{38}$$

The connection of the group-theoretical analysis with the Infeld–Hull factorization method is related to finding the various solutions of the equations in Eq. (38); these various solutions will now be summarized.

$\mathscr{G}(1,0)$ yields

Type A: $k(x) = \cot(x + p)$ and
$$j(x) = \frac{q}{\sin(x + p)};$$

and

Type B: $k(x) = i$ and $j(x) = qe^{-ix}$.

$\mathscr{G}(0,1)$ yields

Type C': $k(x) = \dfrac{1}{x + p}$ and
$$j(x) = -\frac{\mu(x + p)}{4} + \frac{q}{x + p};$$

and

Type D': $k(x) \equiv 0$ and $j(x) = -\dfrac{\mu x}{2} + q$.

$\mathscr{G}(0,0)$ yields

Type C'': $k(x) = \dfrac{1}{x + p}$ and $j(x) = \dfrac{q}{x + p}$;

and

Type D'': $k(x) \equiv 0$ and $j(x) = q$.

The quantities p and q are complex constants, and $i = \sqrt{-1}$. In the Infeld–Hull notation, Types C′ and C″ are combined and called Type C, and Types D′ and D″ are combined to form Type D. The six-dimensional complex Lie algebra $\mathscr{F}_6$, with generators $\mathscr{P}^\pm$, $\mathscr{P}^3$, $\mathscr{J}^\pm$, and $\mathscr{J}^3$, is defined by appropriate commutation relations for these generators. Because of the difficulty involved in computing matrix elements in the representation theory of $\mathscr{F}_6$, the task of obtaining all of the special-function identities implied by $\mathscr{F}_6$ is incomplete. Some results are obtained by noting that the elements $\mathscr{J}^\pm$ and $\mathscr{J}^3$ generate a subalgebra of $\mathscr{F}_6$ that is isomorphic to $sl(2)$, and the elements $\mathscr{P}^\pm$ and

$\mathscr{P}^3$ generate a three-dimensional abelian subalgebra of $\mathscr{F}_6$. Type A and B operators forming a realization of $sl(2)$ can be extended to Type E and F operators, respectively, forming a realization of $\mathscr{F}_6$ (Miller, 1968, Chap. 6). The connection of these eight types of solutions of Eq. (32) with special functions is now summarized.

$\mathscr{G}(1,0)$: Type A solutions are related to hypergeometric functions, and Type B solutions are related to confluent hypergeometric functions and generalized associated Laguerre functions.

$\mathscr{G}(0,1)$: Type C' solutions are closely related to confluent hypergeometric functions, Laguerre and associated Laguerre functions. Type D' solutions are related to parabolic cylinder functions and Hermite polynomials.

$\mathscr{G}(0,0)$: Type C'' solutions are related to Bessel functions, and Type D'' solutions are related to simple transcendental functions, $e^{\pm i\omega x}$.

$\mathscr{F}_6$: Type E solutions are related to hypergeometric functions. Types A and E solutions have different recurrence relations and the same eigenfunctions. Type F solutions are related to confluent hypergeometric functions. Types B and F solutions have different recurrence relations and the same eigenfunctions.

2 The Hypergeometric Function, ${}_2F_1(a, b, c; z)$

2.1 Properties of the Hypergeometric Differential Equation

Important members of a large subset of special functions are related to a class of functions called hypergeometric functions, which are solutions of the hypergeometric differential equation (also known as Gauss's differential equation). The hypergeometric differential equation has three regular singular points, and it can be shown that any second-order ordinary linear differential equation with three regular singular points can be transformed (reduced) to the hypergeometric differential equation form. The solutions of many physical problems involve special functions that result from solving second-order ordinary linear differential equations with regular singular points. It is, therefore, natural to expect a connection among hypergeometric functions and certain special functions. The hypergeometric differential equation has the form

$$z(1-z)\frac{d^2w}{dz^2}+[c-(a+b+1)z]\frac{dw}{dz} - abw = 0. \tag{39}$$

Note that Eq. (39) has regular singular points at $z = 0,\ 1,\ \infty$. In Eq. (39), parameters a, b, and c are arbitrary complex constants. The hypergeometric differential equation can be solved by use of the Frobenius–Fuchs power series method (see, e.g., ANALYTIC METHODS),

$$w = \sum_{\lambda=0}^{\infty} a_\lambda z^{\lambda+k}, \text{ where } a_0 \neq 0.$$

At the three regular singular points, the respective solutions of the indicial equations are, at $z = 0:\ k = 0$ and $1 - c$; at $z = 1:\ k = 0$ and $c. - a - b$; and at $z = \infty:\ k = a$ and b. In general, the solutions of Eq. (39) are the various forms of the Gauss hypergeometric series; they are $[w(z) = {}_2F_1(a, b, c; z) = F(a, b, c; z)]$

$${}_2F_1(a, b, c; z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{n!(c)_n} z_n. \tag{40}$$

The factorial function in Eq. (40) is defined by

$$(\lambda)_n = \prod_{k=1}^{n}(\lambda + k - 1)$$
$$= \lambda(\lambda+1)(\lambda+2)\ldots(\lambda+n-1)$$
$$= \frac{\Gamma(\lambda+n)}{\Gamma(\lambda)} \tag{41}$$

for $\lambda = a$, b, and c, respectively. This notation for the factorial function is called a Pochhammer symbol. The notation ${}_2F_1$ means that there are two factorial functions in the numerator of the series and one factorial function in the denominator of the series in Eq. (40). The gamma function $\Gamma(\lambda)$ is defined in Sec. 5.2.

For convenience, we will write $F(a,b,c;z)$ for ${}_2F_1(a, b, c; z)$ since the three parameters in the parentheses are sufficient to avoid confusion with the notation for the confluent hypergeometric function. Originally, the notation ${}_2F_1(a, b; c; z)$ was used to represent the hypergeometric function. One, however, finds a variety of combinations of commas and/or semicolons used in the literature to represent the hypergeometric function. We will use a comma to separate constants and a semicolon to separate the variable from the constants.

The hypergeometric series $F(-m, b, -n; z)$ is not defined if $n < m$. The hypergeometric series converges if $|z| < 1$ and diverges if $|z| > 1$. For $z = 1$, the series converges if $c > a + b$, and it converges for $z = -1$ if $c > a + b - 1$. The hypergeometric series becomes a polynomial of degree n in z when a or b equals a negative integer.

Note that $F(a, b, c; 0) = 1$ and $F(a, b, c; z) = F(b, a, c; z)$. Also, note that many elementary transcendental functions may be expressed in terms of a hypergeometric series; two examples are $\ln(1 + z) = zF(1, 1, 2; -z)$ and $(1 + z)^a = F(-a, b, b; -z)$. The geometric series is a special case of the hypergeometric series since

$$F(1, 1, 1; z) = \sum_{n=0}^{\infty} z^n.$$

The numerous properties of $F(a,b,c;z)$ summarized in this article as well as in many other places were developed by Euler and Gauss.

2.2 Properties of F(a,b,c;z)

If none of the numbers c, $c - a - b$, or $a - b$ equals an integer, then two linearly independent solutions of the hypergeometric differential equation exist, and each of the six solutions (two at each of three singular points) may be written in four equivalent forms. These 24 forms are referred to as Kummer's 24 solutions of the Gauss hypergeometric differential equation. One of these forms in each case is as follows:

$$w_1(0) = F(a, b, c; z),$$
$$w_2(0) = z^{1-c}F(a - c + 1, b - c + 1, 2 - c; z),$$
$$w_1(1) = F(a, b, a + b + 1 - c; 1 - z),$$
$$w_2(1) = (1 - z)^{c-a-b} \times F(c - b, c - a, c - a - b + 1; 1 - z),$$
$$w_1(\infty) = z^{-a}F(a, a - c + 1, a - b + 1; z^{-1}),$$

and

$$w_2(\infty) = z^{-b}F(b, b - c + 1, b - a + 1; z^{-1}).$$

The functions $F(a \pm 1, b, c; z)$, $F(a, b \pm 1, c; z)$, and $F(a, b, c \pm 1; z)$ are said

to be contiguous to $F(a,b,c;z)$. Numerous relations between $F(a,b,c;z)$ and any two contiguous functions, of a form similar to

$$(c-a)F(a-1,b,c;z) + (2a-c-az+bz)F(a,b,c;z) + a(z-1)F(a+1,b,c;z) = 0,$$

were developed by Gauss.

When $|z| < 1$, $F(a,b,c;z)$ is analytic, and $F(a,b,c;z)$ has a branch point at $z = 1$ (see Whitaker and Watson, 1947, Sec. 14.53); if a cut is made from 1 to infinity along the real axis, $F(a,b,c;z)$ is analytic throughout the cut plane. Formulas for $F(a,b,c;x)$ for $z = x + iy$ may be obtained from those involving $F(a,b,c;z)$ by use of the following replacements in the cut interval $(-1,1) : z - 1$ by $(1-x)e^{\pm i\pi}$, $z^2 - 1$ by $(1-x^2)e^{\pm i\pi}$, and $z + 1$ by $x + 1$.

The following differentiation and integration relations are valid for ${}_2F_1$.

A differentiation relation for $F(a,b,c;z)$:

$$\frac{d^n F(a,b,c;z)}{dz^n} = \frac{(a)_n(b)_n}{(c)_n} \times F(a+n, b+n, c+n; z).$$

Fundamental integral relation: The form of the fundamental integral relation for the hypergeometric function is

$$F(a,b,c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \times \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt$$

for $\mathrm{Re}(c) > \mathrm{Re}(b) > 0$.

Solutions of the hypergeometric differential equation that are orthogonal polynomials are of particular interest in this article, and polynomial solutions occur when a or b is a negative integer. Examples of the connections of $F(a,b,c;x)$ with some special polynomials of interest are as follows:

$$F(-n, n, \tfrac{1}{2}; x) = T_n(1-2x) \text{ – Chebyshev;}$$

$$F(-n, n+1, 1; x) = P_n(1-2x) \text{ – Legendre;}$$

$$F(-n, n+2\alpha, \alpha+\tfrac{1}{2}; x) = \frac{n!}{(2\alpha)_n} C_n^{(\alpha)}(1-2x) \text{ – Gegenbauer,}$$

and

$$F(-n, \alpha+1+\beta+n, \alpha+1; x) = \frac{n!}{(\alpha+1)_n} P_n^{(\alpha,\beta)}(1-2x) \text{ – Jacobi.}$$

The Gegenbauer (also known as ultraspherical), Legendre and associated Legendre, and Chebyshev polynomials are special cases of the Jacobi polynomial (sometimes called hypergeometric polynomial). Chebyshev (Tschebyscheff, Tchebichef, and Tchebicheff are other spellings found in the literature) polynomials involve solutions of separated equations in spherical, parabolic, and prolate and oblate spheroidal coordinates. Chebyshev polynomials converge rapidly and have the special property that $\mathrm{Max}T_n(x) = +1$ and $\mathrm{Min}T_n(x) = -1$; because of this property, Chebyshev polynomials are useful in numerical analysis. Gegenbauer functions result from separated equations in circular cylinder and spherical coordinates with two regular singular points at ± 1 rather than at 0 and 1.

We now summarize some of the basic properties of the solutions of the Jacobi differential equation (see Table 1 in Sec. 1.1.) and express these solutions in terms of ${}_2F_1$. The solutions of the Jacobi differential equation may be written as

$$y = c_1 P_n^{(\alpha,\beta)}(x) + c_2 Q_n^{(\alpha,\beta)}(x).$$

The quantity $P_n^{(\alpha,\beta)}(x)$ is a polynomial and is called Jacobi polynomial of the first kind. The quantity $Q_n^{(\alpha,\beta)}(x)$ is not a polynomial and is called Jacobi function of the second kind. In terms of hypergeometric functions, we write

$$P_n^{(\alpha,\beta)}(x) = \frac{(\alpha+1)_n}{n!} \times F\left(-n, n+\alpha+\beta+1, \alpha+1; \frac{1-x}{2}\right)$$

and

$$Q_n^{(\alpha,\beta)}(x) = \frac{C(\alpha,\beta)}{(x-1)^{n+\alpha+1}(x+1)^{\beta}} \times F\left(n+1, n+\alpha+1, 2n+\alpha+\beta+2; \frac{2}{1-x}\right).$$

The symbol $C(\alpha, \beta)$ in the above equation represents the quantity

$$C(\alpha,\beta) = \frac{2^{n+\alpha+\beta} \times \Gamma(n+\alpha+1) \times \Gamma(n+\beta+1)}{\Gamma(2n+\alpha+\beta+2) \times (x-1)^{n+\alpha+1}}.$$

The standardization for the Jacobi polynomial is given by

$$P_n^{(\alpha,\beta)}(1) = (\alpha+1)_n/n!.$$

Rodrigues's formula and the generating function for the Jacobi polynomial are given in Tables 4 and 5, respectively. The form of the recursion formula for the Jacobi polynomial is

$$\begin{aligned} &2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta) \\ &\times P_{n+1}^{(\alpha,\beta)}(x) = (2n+\alpha+\beta+1) \\ &\quad \times [(2n+\alpha+\beta)(2n+\alpha+\beta) \\ &\quad \times x + \alpha^2 - \beta^2] \\ &\quad \times P_n^{(\alpha,\beta)}(x) - 2(n+\alpha)(n+\beta) \\ &\quad \times (2n+\alpha+\beta+2) P_{n-1}^{(\alpha,\beta)}(x). \end{aligned}$$

The integral representation of the Jacobi polynomial may be written as

$$P_n^{(\alpha,\beta)}(x) = \frac{1}{2\pi i} \oint_c \frac{1}{2}\left(\frac{t^2-1}{t-x}\right)^n \times \left(\frac{1-t}{1-x}\right)^{\alpha} \left(\frac{1+t}{1+x}\right)^{\beta} dt; \; x \neq \pm 1. \tag{42}$$

The contour in Eq. (42) is a simple closed contour in a positive sense around $t = x$; the points $t = \pm 1$ are outside of the contour. In Eq. (42), the quantities raised to the α and β power are defined to be unity when $t = x$. Graphical illustrations

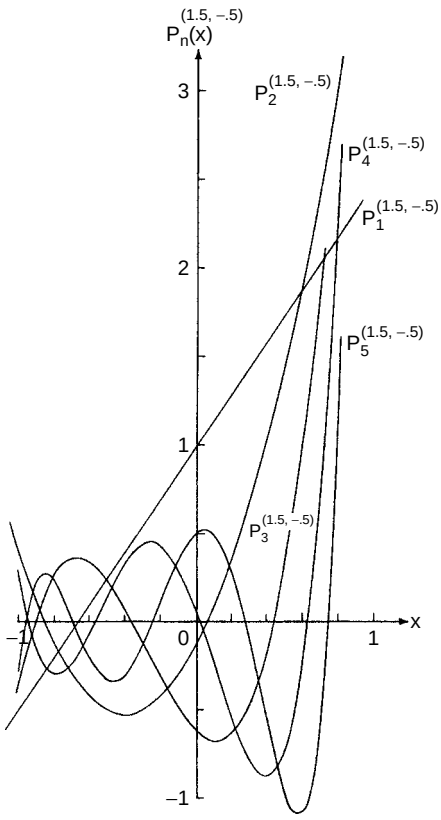


Fig. 1 Jacobi polynomials, $P_n^{(1.5,-0.5)}(x)$ (Abramowitz and Stegun, 1964)

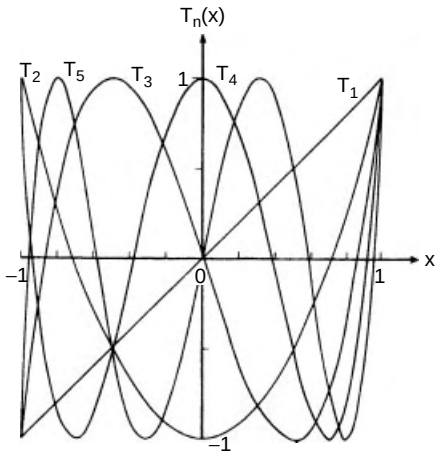


Fig. 2 Chebyshev polynomials, $T_n(x)$ (Abramowitz and Stegun, 1964)

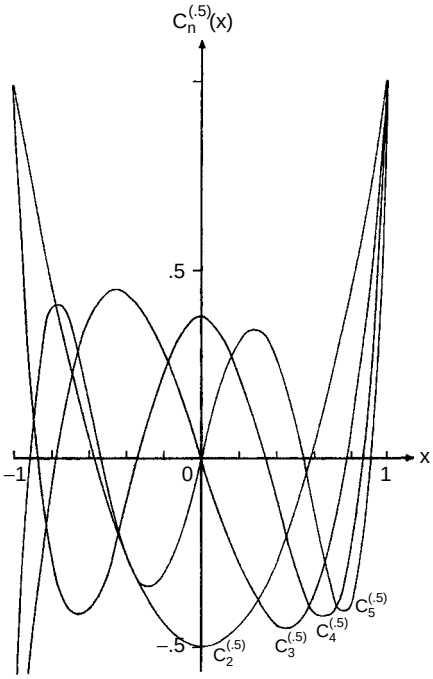


Fig. 3 Gegenbauer polynomials, $C_n^{(0.5)}(x)$ (Abramowitz and Stegun, 1964)

of Jacobi, Chebyshev, and Gegenbauer polynomials are given in Figs. 1–3.

2.3 Helmholtz's Differential Equation in Spherical Coordinates

The Laplacian operator in spherical coordinates (r, θ, ϕ) has the form

$$\nabla^2 = \nabla_r^2 + \frac{1}{r^2}\nabla_{\theta,\phi}^2.$$

The radial and angular parts of the Laplacian are given respectively by

$$\nabla_r^2 = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) \tag{43}$$

and

$$\nabla_{\theta,\phi}^2 = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}. \tag{44}$$

The Helmholtz differential equation in spherical coordinates has the form

$$\left\{\nabla_r^2 + \frac{1}{r^2}\nabla_{\theta,\phi}^2\right\}u + k^2u = 0. \tag{45}$$

The solution of Eq. (45) involves symmetry-type information in terms of the angular parts θ and ϕ, which are common to all problems with spherical symmetry, and dynamical information in terms of the radial part, which characterizes the particular problem under investigation. In this connection, the Helmholtz differential equation reduces to the Laplace differential equation for $k^2 = 0$, to the time-independent heat-conduction (diffusion) or time-independent mechanical wave differential equations for $k^2 = \text{constant}$, and to the time-independent Schrödinger wave equation for $k^2 = (2m/\hbar^2)\{E - V(r)\}$. On separating the angular parts from the radial part in Eq. (45) for $u(r, \theta, \phi) = R(r)Y(\theta, \phi)$, the corresponding differential

equations for $R(r)$ and $Y(\theta, \phi)$ with separation constant λ are

$$\nabla_r^2 R(r) + \left\{k^2 - \frac{\lambda}{r^2}\right\} R(r) = 0 \qquad (46)$$

and

$$\nabla_{\phi,\phi}^2 Y(\theta, \phi) + \lambda Y(\theta, \phi) = 0. \qquad (47)$$

Note that the general solutions of Eq. (47) are independent of the specific problem under investigation but are common to all problems that involve the Laplacian operator in spherical coordinates. The solutions of Eq. (47) with separation constant $-m^2$ are called spherical harmonics (also known as surface harmonics of the first kind), $Y_\lambda^m(\theta, \phi)$. For square integrable solutions, a replacement of the form $\lambda = n(n+1)$ is required. Tesseral harmonics is the name given to $Y_n^m(\theta, \phi)$ when $m < n$, and the term sectoral harmonics is used when $m = n$. Tesseral and sectoral harmonics may be written as $C_n e^{im\phi} P_n^m(\cos\theta)$, where $P_n^m(\cos\theta)$ are associated Legendre functions of the first kind. When $m = 0$, the spherical functions are called Legendre polynomials of the first kind (also known as zonal harmonics and Legendre coefficients).

On substituting Eqs. (43) and (44) into Eqs. (46) and (47), respectively, and separating the variables in Eq. (47) [$Y(\theta, \phi) = \Theta(\theta)\Phi(\phi)$ with separation constant $-m^2$], we obtain the following three ordinary differential equations: radial equation,

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \left[k^2 - \frac{n(n+1)}{r^2}\right] R = 0; \qquad (48)$$

theta equation,

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right) + \left[n(n+1) - \frac{m^2}{\sin^2\theta}\right]\Theta = 0; \qquad (49)$$

and azimuthal equation,

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0. \qquad (50)$$

The general solution of the azimuthal equation is $\Phi = c_1 e^{im\phi} + c_2 e^{-im\phi}$. In solving physical problems, the requirement that $\Phi(\phi)$ be a single-valued function is imposed. That is to say, we require that $\Phi(\phi) = \Phi(\phi + 2\pi)$, which leads to the following acceptable values for $m : m = 0, \pm 1, \pm 2, \ldots$; the solution may be written in the form $\Phi(\phi) = Ae^{im\phi}$. In quantum mechanics, the single-valued function requirement is referred to as Born's periodic boundary condition, and m is the magnetic quantum number.

2.4 Associated Legendre Functions and Legendre Polynomials

Solutions of the theta equation, Eq. (49), involve the Legendre polynomials and associated Legendre functions. The traditional treatment of the theta equation involves introducing a new independent variable x by use of the transformation $x = \cos\theta$; here x is not the usual Cartesian coordinate. The associated Legendre differential equation results from this transformation,

$$(1 - x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + \left\{n(n+1) - \frac{m^2}{1 - x^2}\right\}\Theta = 0. \qquad (51)$$

The Legendre differential equation is the result when m equals zero in Eq. (51); it is given by

$$(1 - x^2)\frac{d^2\Theta}{dx^2} - 2x\frac{d\Theta}{dx} + n(n+1)\Theta = 0. \qquad (52)$$

The Legendre and associated Legendre differential equations can be solved by use of the Frobenius–Fuchs power series method, factorization method, and hypergeometric functions. The two linearly independent solutions of the associated Legendre differential equation are

$$\Theta_n^m(x) = AP_n^m(x) + BQ_n^m(x).$$

The quantities $P_n^m(x)$ and $Q_n^m(x)$ are associated Legendre functions of the first and second kind, respectively; they are related to the hypergeometric functions as follows:

$$P_n^m(z) = \frac{1}{\Gamma(1-m)}\left(\frac{z+1}{z-1}\right)^{m/2} \times F\left(-n, n+1, 1-m; \frac{1-z}{2}\right), \quad |1-z| < 2,$$

and

$$Q_n^m(z) = C^{nm} \times F\left(1+\frac{n}{2}+\frac{m}{2}, \frac{1}{2}+\frac{n}{2}+\frac{m}{2}, n+\frac{3}{2}; \frac{1}{z^2}\right); |z| > 1.$$

The factor C^{nm} in the above equation is given by

$$C^{nm} = e^{im\pi}2^{-n-1}\pi^{1/2} \times \frac{\Gamma(n+m+1)}{\Gamma(n+\frac{3}{2})} z^{-n-m-1} (z^2-1)^{m/2}.$$

Also, the associated Legendre functions $P_n^m(x)$ result when the Legendre polynomials are differentiated m times. That is to say,

$$P_n^m(x) = (1-x^2)^{m/2}\frac{d^m P_n(x)}{dx^m}.$$

The orthogonality relation, recurrence relation, and Rodrigues formula for the associated Legendre functions are given in Tables 2, 3, and 4, respectively. Graphical illustrations of associated Legendre functions are given in Fig. 4.

The Legendre polynomials of the first kind, $P_n(x)$, and Legendre functions of the second kind, $Q_n(x)$, are related to the Jacobi functions as follows:

$$P_n^{(0,0)}(x) = P_n(x) \text{ and } Q_n^{(0,0)}(x) = Q_n(x).$$

In series form, the Legendre polynomials may be written as

$$\Theta_n(x) \equiv P_n(x) = \sum_{r=0}^{N} \frac{(-1)^r(2n-2r)!x^{n-2r}}{2^n r!(n-r)!(n-2r)!}.$$

In the above sum, $N = n/2$ for n even and $N = (n-1)/2$ for n odd. The general solution of the Legendre differential equation has the form $\Theta_n(x) = C_1P_n(x) +$

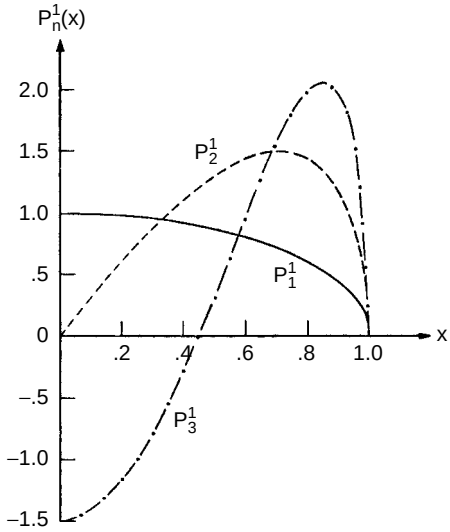


Fig. 4 Associated Legendre functions, $P_n^1(x)$ (Abramowitz and Stegun, 1964)

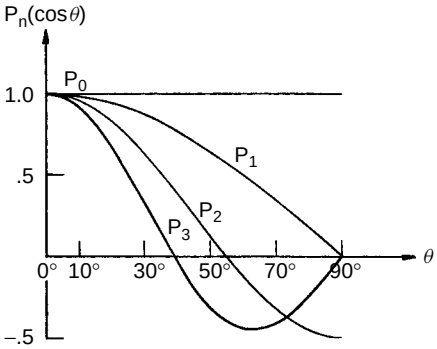


Fig. 5 Legendre polynomials, $P_n(\cos\theta)$ (Abramowitz and Stegun, 1964)

$C_2 Q_n(x)$; Legendre functions of the second kind, $Q_n(x)$, are not polynomials. The orthogonality relation, recurrence relation, Rodrigues formula, and generating function for the Legendre polynomials are given in Tables 2, 3, 4, and 5, respectively.

The Legendre functions of the second kind, $Q_n(x)$, satisfy a recursion relation of the same form as the one for $P_n(x)$. Graphical illustrations of Legendre polynomials are given in Figs. 5–7, and Legendre functions of the second kind are illustrated in Figs. 8 and 9.

2.5 The Radial Equation

The radial equation, Eq. (48), characterizes the dynamical information of specific problems or classes of problems. The Laplace equation results when $k^2 = 0$; for this case, the solution characterizes such steady-state problems as potentials in electrostatics and temperatures in heat conduction. The time-independent Schrodinger equation results when $k^2 = 2\mu\{E - V(r)\}/\hbar^2$ for a class of two-body central-force problems; the reduced mass of such a system is given by $\mu = m_1 m_2/(m_1 + m_2)$, where m_1 and m_2 are the masses of constituent particles one and two, respectively.

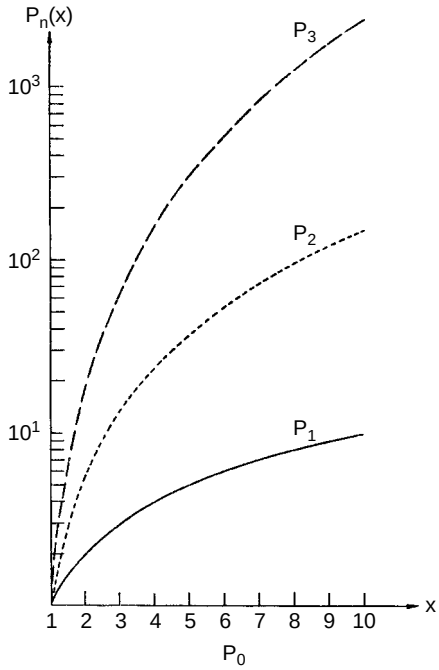


Fig. 6 Legendre polynomials, $P_n(x)$ (Abramowitz and Stegun, 1964)

EXAMPLE 3: Calculate the steady-state temperature distribution $T(r, \theta)$ within a sphere of radius b when the temperature over the surface of the sphere is independent of ϕ. That is to say, $T(b, \theta) = f(\theta)$, where $f(\theta)$ is a known function.

Solution: The general solution of Laplace's differential equation of this problem is independent of ϕ (circular symmetry) and has the form $T(r, \theta) = R(r)\Theta(\theta)$. The radial and theta equations for this problem reduce to

$$r^2 R'' + 2rR' - n(n+1)R = 0$$

and

$$\sin\theta\Theta'' + \cos\theta\Theta' + n(n+1)\sin\theta\Theta = 0.$$

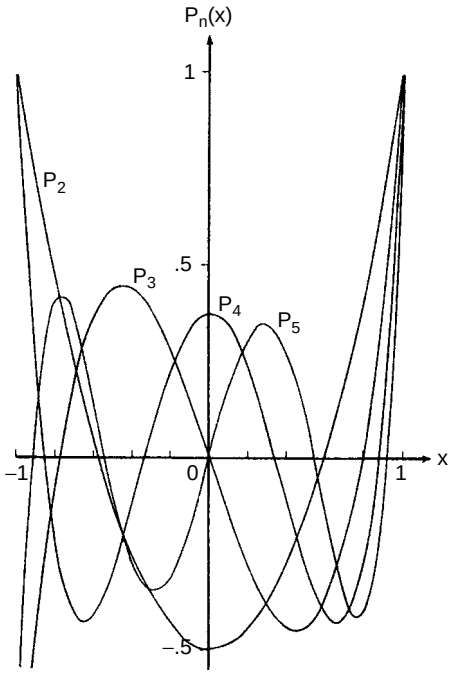


Fig. 7 Legendre polynomials, $P_n(x)$ (Abramowitz and Stegun, 1964)

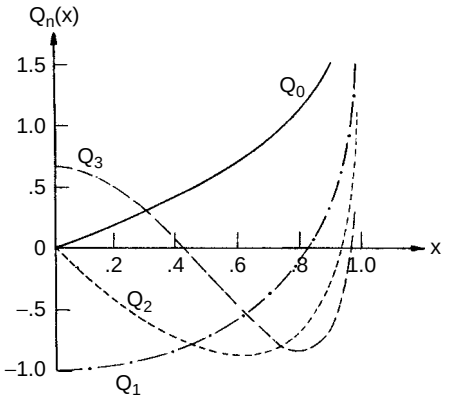


Fig. 8 Legendre functions of the second kind, $Q_n(x)$ (Abramowitz and Stegun, 1964)

The general solution of the radial equation is

$$R(r) = Ar^n + B/r^{n+1}.$$

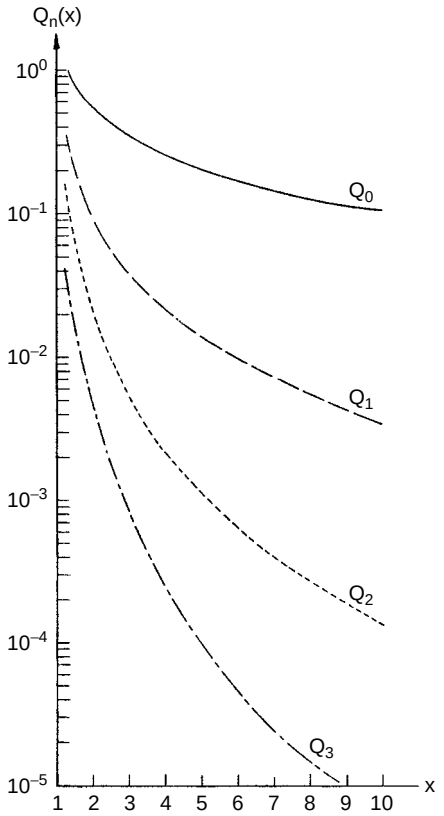


Fig. 9 Legendre functions of the second kind, $Q_n(x)$ (Abramowitz and Stegun, 1964)

Constants A and B are arbitrary, and the general solution of the theta equation (the Legendre differential equation) is $P_n(\cos\theta)$. We set the coefficient B equal to zero since a solution finite at every point within the sphere is required. The general solution of the problem is a superposition of products of radial and theta solutions; we write

$$T(r,\theta) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\theta).$$

The coefficients A_n are obtained by use of the boundary condition $T(b,\theta) = f(\theta)$ and the orthogonality relations for the

Legendre polynomials. The specific form for $f(\theta)$ must be given if the specific values for A_n are required.

3 The Confluent Hypergeometric Function, ${}_1F_1(a, c; x)$

The confluent hypergeometric differential equation (also called the Kummer differential equation) has the form

$$xy'' + (c - x)y' - ay = 0. \tag{53}$$

Equation (53) is obtained from the hypergeometric differential equation by substituting $z = bx$ and letting b approach infinity; this substitution causes a merging or confluence of the two upper singular points. In the confluent hypergeometric differential equation, there is a regular singularity at $x = 0$ and an irregular singularity at $x = \infty$. By use of the power-series method in the neighborhood of $x = 0$, we find that one solution of Eq. (53) has the form (confluent hypergeometric functions or Kummer functions)

$$\begin{aligned} y(x) \equiv {}_1F_1(a, c; x) &= 1 + \frac{ax}{c} \\ &\quad + \frac{a(a+1)x}{2!c(c+1)} + \cdots \\ &= \sum_{n=0}^{\infty} \frac{(a)_n x^n}{(c)_n}. \end{aligned}$$

The confluent hypergeometric series converges for all values of x.

The Bessel functions and modified Bessel functions in terms of ${}_1F_1$ are respectively given by

$$J_n(x) = \frac{e^{-ix}}{n!}\left(\frac{x}{2}\right)^n {}_1F_1(n + \tfrac{1}{2}, 2n + 1; 2ix)$$

and

$$I_n(x) = \frac{e^{-x}}{n!}\left(\frac{x}{n!}\right)^n {}_1F_1(n + \tfrac{1}{2}, 2n + 1; 2x).$$

The Laguerre and associated Laguerre polynomials in terms of ${}_1F_1$ are respectively given by

$$L_n(x) = {}_1F_1(-n, 1; x)$$

and

$$L_n^m(x) = \frac{(n+m)!}{n!m!}\,{}_1F_1(-n, m + 1; x).$$

Hermite polynomials in terms of ${}_1F_1$ have the form

$$H_{2n}(x) = (-1)^n \frac{(2n)!}{n!}\,{}_1F_1(-n, \tfrac{1}{2}; x^2).$$

The error function and complementary error function are respectively defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt,$$

$$\text{where } \operatorname{erf}(\infty) = 1$$

and

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2}\, dt.$$

The Hermite polynomials may be obtained from derivatives of the error function as follows:

$$\frac{d^{n+1}\operatorname{erf}(x)}{dx^{n+1}} = (-1)^n \frac{2}{\sqrt{\pi}} H_n(x) e^{-x^2}.$$

In addition, the error function is related to the confluent hypergeometric function by use of the following equation:

$$\operatorname{erf}(x) = \frac{2x}{\sqrt{\pi}}\,{}_1F_1(\tfrac{1}{2}, \tfrac{3}{2}; -x^2).$$

3.1 More on Hermite Polynomials

The Hermite differential equation,

$$y'' - 2xy' + 2ny = 0 \text{ for } n = \text{constant},$$

is a special case of the Sturm–Liouville differential equation and a special case of the confluent hypergeometric differential equation. The polynomial solution $H_n(x)$ of the Hermite differential equation may be obtained by use of the Frobenius–Fuchs power-series method,

$$y = \sum_{r=0}^{\infty} a_r x^{k+r}, \text{ where } a_0 \neq 0.$$

For standardization $a_n = 2^n$, the solution has the form

$$y_n \equiv H_n(x) = \sum_{j=0}^{N} \frac{(-1)^j n!(2x)^{n-2j}}{j!(n-2j)!}. \tag{54}$$

Equation (54) is the series form of the Hermite polynomials. In Eq. (54), $N = n/2$ for n even and $N = (n-1)/2$ for n odd. The orthogonality relation, recurrence relation, Rodrigues formula, and generating function for $H_n(x)$ are respectively given in Tables 2, 3, 4, and 5. A representative sketch of $H_n(x)$ is given in Fig. 10.

EXAMPLE 4: More on the quantum mechanical linear harmonic oscillator.

Solution: The problem of describing the small oscillation of a mass m attached to the end of a spring with force constant k and potential energy $V(x) = kx^2/2$ can be solved exactly in both classical and quantum theory. This system is referred to as a linear harmonic oscillator and is used to represent and analyze more complex physical systems such as vibrations of individual atoms in molecules and in crystals, and classical and quantum theories of radiation. The solution of Schrodinger's wave equation for the linear harmonic oscillator is expressed in terms of Hermite polynomials $H_n(x)$. The equation to be solved is the one-dimensional wave equation for the linear harmonic oscillator, which has the form

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + \frac{kx^2\psi}{2} = E\psi. \tag{55}$$

Note that Eq. (55) is just the one-dimensional Helmholtz equation for which k^2 equals $2m(E - kx^2/2)/\hbar^2$. Solving a problem in quantum mechanics involves finding the wave functions ψ_n and the corresponding eigenenergies E_n. In dimensionless form, Eq. (55) becomes

$$\frac{d^2\psi}{d\xi^2} + (\lambda - \xi^2)\psi = 0,$$

$$\text{where } \xi = \left(\frac{m\omega}{\hbar}\right)^{1/2} x, \omega^2 = \frac{k}{m},$$

$$\text{and } \lambda = \frac{2E}{\hbar\omega}. \tag{56}$$

On substituting $\lambda = 1 + 2n$ into Eq. (56), we obtain the Weber differential equation, and the transformation $\psi = \exp(-\xi^2/2)\, y(\xi)$ reduces the Weber differential equation to the Hermite differential equation. The transformation equation leading to Hermite's differential equation is motivated by use of the Sommerfeld (1949) polynomial method for solving certain differential equations. According to the Sommerfeld method, the solution of Eq. (56)

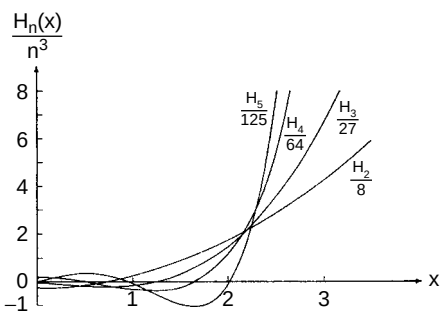


Fig. 10 Hermite polynomials, $H_n(x)/n^3$ (Abramowitz and Stegun, 1964)

is the product of the physically acceptable asymptotic solution times a polynomial. The resulting asymptotic ($|\xi|$ approaches infinity) differential equation is $\psi'' - \xi^2\psi = 0$, and the physically acceptable solution of this differential equation is given by $\exp(-\xi^2/2)$. In this case, the polynomial part of the solution of Eq. (56) comes from solving Hermite's differential equation. The eigenfunctions and eigenenergy for the linear harmonic oscillator, Eq. (56), are respectively given by

$$\psi_n(\xi) = Ne^{-\xi^2/2}H_n(\xi) \text{ and}$$
$$\lambda_n = 1 + 2n = 2E_n/\hbar\omega \text{ or}$$
$$E_n = \hbar\omega(n + \tfrac{1}{2}).$$

The behavior of $\psi_n(\xi)$ for the first six values of n is illustrated in the sketches in Fig. 11.

3.2 More on the Laguerre and Associated Laguerre Polynomials

The Laguerre differential equation, $xy'' + (1 - x^2)y' + ny = 0$ for $n =$ constant, is a special case of the Sturm–Liouville differential equation as well as a special case of the confluent hypergeometric differential equation; its solution may be obtained by relation to ${}_1F_1$ or by use of the Frobenius–Fuchs power-series method,

$$y(x) = \sum_{\lambda=0}^{\infty} a_\lambda x^{k+\lambda} \text{ for } a_0 \neq 0.$$

The indicial equation in the power-series method has a double root at $k = 0$, and the power-series method yields only one of the two linearly independent solutions

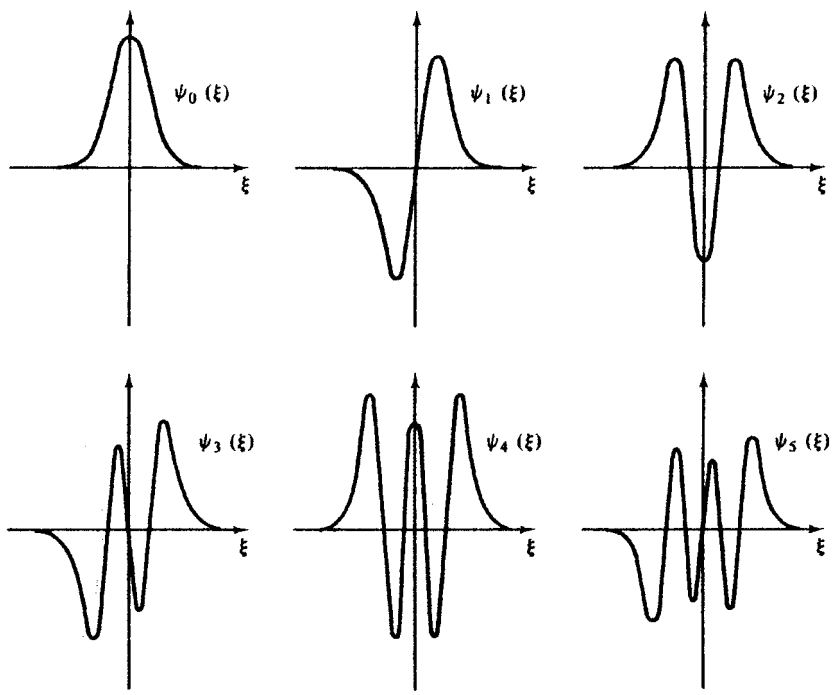


Fig. 11 Linear harmonic oscillator wave functions, $\psi_n(\xi)$

of the Laguerre differential equation; this solution is extremely important in mathematical physics. The finite solutions of the Laguerre differential equation, $L_n(x)$, are called Laguerre polynomials. The solution in series form is

$$L_n(x) = \sum_{\lambda=0}^{n} \frac{(-1)^\lambda n! x^\lambda}{(n-\lambda)!(\lambda!)^2};$$

$$a_0 \equiv 1 \text{ (standardization)}.$$

The orthogonality and recurrence relations for the Laguerre polynomials are given in Tables 2 and 3, respectively; the generating function and Rodrigues formula are respectively given by

$$\frac{\exp[-xt/(1-t)]}{(1-t)} = \sum_{n=0}^{\infty} \frac{L_n(x)t^n}{n!}$$

and

$$L_n(x) = \frac{1}{n!} e^x \frac{d^n}{dx^n}(x^n e^x).$$

Note that $L_n(0) = 1$, $L_0(x) = 1$, $L_1(x) = 1 - x$, and $L_2(x) = 2 - 4x + x^2$. A sketch of Laguerre polynomials is given in Fig. 12.

Note that the kth derivative of the Laguerre differential equation yields the associated Laguerre differential equation,

$$x\frac{d^2 L_n^k(x)}{dx^2} + (k+1-x)\frac{dL_n^k(x)}{dx} + (n-k)L_n^k(x) = 0. \quad (57)$$

In obtaining Eq. (57), the Leibnitz formula, Eq. (16), for finding the kth derivative of a product was used, and we note that the associated Laguerre polynomials $L_n^k(x)$ are related to the Laguerre polynomials by

$$L_n^k(x) = \frac{d^k L_n(x)}{dx^k}.$$

The orthogonality relation, recurrence relation, Rodrigues formula, and generating function for associated Laguerre polynomials are respectively given in Tables 2, 3, 4, and 5.

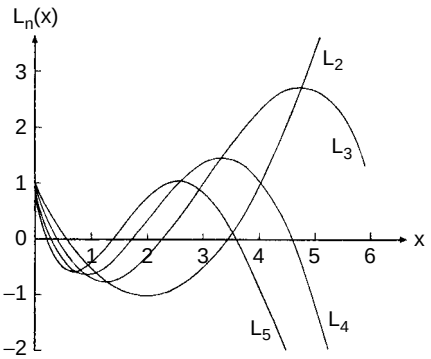


Fig. 12 Laguerre polynomials, $L_n(x)$ (Abramowitz and Stegun, 1964)

EXAMPLE 5: a central-force problem in quantum mechanics.

Solution: The time-independent Schrödinger wave equation is used to study the mechanics of two microscopic particles moving under the influence of a central force (that is to say, the force and potential depend only on the distance between the two particles). This problem provides the basis for the quantum mechanical treatment of a fundamental class of problems such as the rigid rotator, which is of considerable importance in the study of the spectra of diatomic molecules; the theory of the hydrogen atom; and the nonrelativistic theory of the deuteron. For an attractive inverse-square force law, we substitute $k^2 = 2\mu(E - A/r)/\hbar^2$ into the radial equation; the quantity A is a positive constant. Anticipating use of the well-known solution of the associated Laguerre differential equation, the following substitutions are made in

the radial equation for inverse-square force problems:

$$\xi = \beta r, \beta^2 \equiv \frac{8\mu|E|}{\hbar^2}, \text{ and}$$

$$\gamma \equiv \frac{2\mu A}{\beta\hbar^2} = \frac{A}{\hbar}\left(\frac{\mu}{2|E|}\right)^{1/2}. \quad (58)$$

By use of the substitutions in Eq. (58), the radial equation, Eq. (48), reduces to

$$\xi\frac{d^2R}{d\xi^2} + 2\frac{dR}{d\xi} + \left\{\gamma - \frac{\xi}{4} - \frac{l(l+1)}{\xi}\right\}R = 0. \quad (59)$$

The symbol l represents the angular momentum quantum number; it is used instead of n in Eq. (48) since n denotes principal quantum number for this problem. Equation (59) is reduced to the associated Laguerre differential equation by use of the transformation

$$R_{nl}(\xi) = \exp(-\xi/2)\xi^l L_{n+1}^{2l+1}(\xi).$$

The desired solution of the equation to be solved, Eq. (59), is a normalization constant times $R_{nl}(\xi)$. A polynomial solution of the associated Laguerre differential equation is obtained when $\gamma = n = l + k + 1$ for $k = 0, 1, 2, \ldots, n-1$.

EXAMPLE 6: the hydrogen atom: Obtain the eigenfunction and eigenenergy for the hydrogen atom.

Solution: The hydrogen atom represents a two-body central-force problem in quantum mechanics, where the electron and proton are the two particles under investigation. The Coulomb potential is the central potential for the hydrogen atom. Here, the total energy is negative for bound states, $E < 0$, and the attractive potential energy is given by $V = -e^2/4\pi\varepsilon_0 r = -A/r$. The eigenenergy E_n is obtained from Eq. (58); we obtain the usual Bohr result,

$$E_n = -\mu A^2/2\hbar^2 n^2.$$

The corresponding steady-state wave function is the product of solution of the radial part and the angular part, which is $Y_l^m(\sigma, \phi)$ times a normalization constant, C_{nl}; the result is

$$\psi(r, \theta, \phi) = -C_{nl}R_{nl}(\xi)Y_l^m(\theta, \phi). \quad (60)$$

The eigenfunction solution, Eq. (60), is given in most quantum mechanics books, and the normalization constant is obtained in the usual manner.

4 Helmholtz's Differential Equation in Cylindrical Coordinates

Problems in mathematical physics that involve cylindrical geometry are, in general, simpler to solve in cylindrical coordinates (ρ, ϕ, z) than in Cartesian coordinates. The Helmholtz differential equation in cylindrical coordinates has the form

$$\rho\frac{\partial^2 u}{\partial\rho^2} + \frac{\partial u}{\partial\rho} + \frac{1}{\rho}\frac{\partial^2 u}{\partial\phi^2} + \rho\frac{\partial^2 u}{\partial z^2} + k^2\rho u = 0.$$

By use of the separation-of-variables method for $u(\rho, \phi, z) = P(\rho)\Phi(\phi)Z(z)$ with separation constants $-\lambda^2$ and $-n^2$, the following three ordinary differential equations are obtained:

$$\frac{1}{Z}\frac{d^2Z}{dz^2} = -\lambda^2$$

with solution $Z(z) = A\cos(\lambda z) + B\sin(\lambda z)$;

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = -n^2$$

with solution $\Phi(\phi) = C\cos(n\phi) + D\sin(n\phi)$, and

$$\xi^2 \frac{d^2P}{d\xi^2} + \xi\frac{dP}{d\xi} + (\xi^2 - n^2)P = 0$$
$$\text{for } \xi = \alpha\rho \text{ and } k^2 - \lambda^2 \equiv \alpha^2. \quad (61)$$

4.1 Solutions of Bessel's Differential Equation

Equation (61) is Bessel's differential equation, and its solutions are called Bessel (or cylindrical) functions. Bessel's differential equation is solved by use of the power-series method,

$$P(\xi) = \sum_{\lambda=0}^{\infty} a_\lambda \xi^{k+\lambda} \text{ for } a_0 \neq 0.$$

The general solutions of Bessel's differential equation when n is an integer are

$$P(\xi) = AJ_n(\xi) + BN_n(\xi);\ n = \text{integer}.$$

The functions $J_n(\xi)$ and $N_n(\xi)$ are explained in the following sections.

4.2 Bessel Functions of the First Kind

The functions $J_n(\xi)$ are called Bessel functions of the first kind; the series representation, generating function, and recurrence relation for $J_n(\xi)$ are respectively given by

$$J_n(\xi) = \sum_{j=0}^{\infty} \frac{(-1)^j(\xi/2)^{2j+1}}{j!\Gamma(n+j+1)}$$

for standardization

$$a_0 \equiv \frac{1}{2^n\Gamma(n+1)}, \quad (62)$$

$$\exp\left\{\tfrac{1}{2}\xi\left(t - \frac{1}{t}\right)\right\} = \sum_{n=-\infty}^{\infty} J_n(\xi)t^n,$$

and

$$J_{n-1}(\xi) + J_{n+1}(\xi) = \frac{2n}{\xi}J_n(\xi).$$

Note that $J_{-n}(\xi) = (-1)^n J_n(\xi)$ results from use of Eq. (62). The gamma function Γ is defined in Sec. 5.2. The orthogonality relation for the interval $[0,a]$ may be written in the form

$$\int_0^a J_n\left(\beta_{ni}\frac{\xi}{a}\right) J_n\left(\beta_{nj}\frac{\xi}{a}\right)\xi\, d\xi = \begin{cases} 0, & i \neq j, \\ \dfrac{a^2}{2}[J_{n+1}(\beta_{nj})]^2, & i = j. \end{cases} \quad (63)$$

In Eq. (63), $n > -1$, the parameter β_{ni} is the ith zero of J_n, and $0 \leq \xi \leq a$.

The general solutions of Bessel's differential equation when n is not an integer are

$$P(\xi) = CJ_n(\xi) + DJ_{-n}(\xi); \quad n \neq \text{integer}.$$

Sketches of several Bessel functions of the first kind are given in Fig. 13.

4.3 Neumann Functions

The Neumann functions are defined by

$$Y_n(\xi) = N_n(\xi) \equiv \frac{J_n(\xi)\cos n\pi - J_{-n}(\xi)}{\sin n\pi}; \quad n = \text{integer}.$$

The Neumann functions are called Bessel functions of the second kind. L'Hospital's rule should be used to evaluate $N_n(\xi)$. Sketches of several Neumann functions are given in Fig. 13.

4.4 Hankel Functions

Hankel functions of the first and second kind are respectively defined by

$$H_n^{(1)}(\xi) \equiv \frac{i}{\sin n\pi}[e^{-n\pi i}J_n(\xi) - J_{-n}(\xi)] = J_n(\xi) + iN_n(\xi)$$

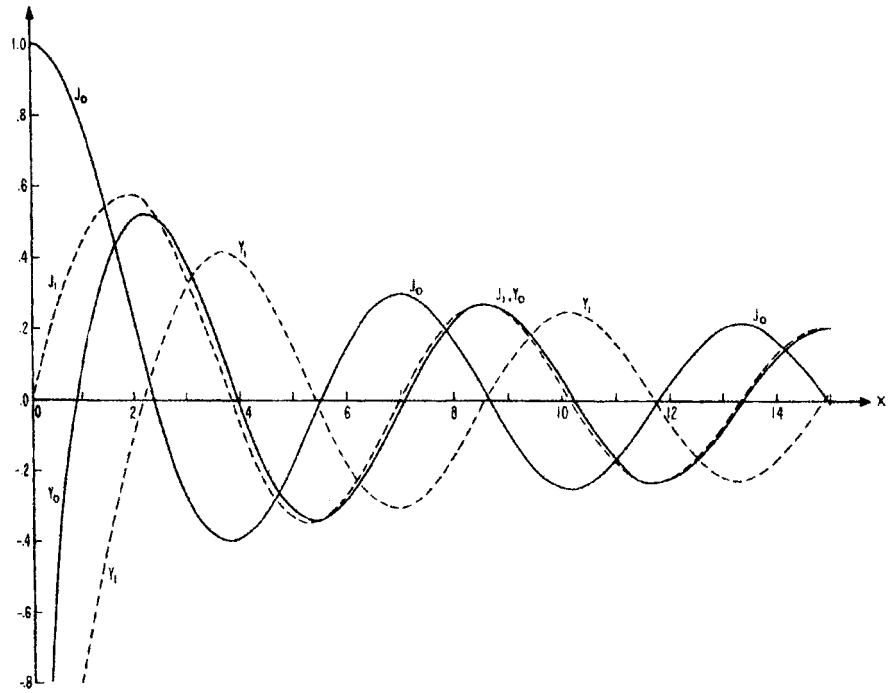


Fig. 13 Bessel functions of the first and second kinds, $J_n(x)$ and $Y_n(x)$ (Abramowitz and Stegun, 1964)

and

$$H_n^{(2)}(\xi) \equiv -\frac{i}{\sin n\pi}[e^{n\pi i}J_n(\xi) - J_{-n}(\xi)]$$
$$= J_n(\xi) - iN_n(\xi).$$

The Hankel functions are independent solutions of the Bessel differential equation, and they are useful in connection with their behavior for large values of ξ since they are infinite at $\xi = 0$. Hankel functions are sometimes referred to as Bessel functions of the third kind.

4.5 Modified Bessel Functions

The modified Bessel differential equation is obtained when ξ is replaced with it in Eq. (61); the result is

$$t^2\frac{d^2P}{dt^2} + t\frac{dP}{dt} - (t^2 + n^2)P = 0. \qquad (64)$$

The solutions of Eq. (64) are called modified Bessel functions of the first kind and are denoted by $I_n(\xi)$; they are given by

$$I_n(\xi) = i^{-n}J_n(i\xi) = \sum_{\lambda=0}^{\infty}\frac{(\xi/2)^{2\lambda+n}}{\lambda!(\lambda+n)!};$$
$$n = \text{integer}.$$

When n is not an integer, $I_n(\xi)$ and $I_{-n}(\xi)$ are linearly independent solutions of the modified Bessel differential equation, Eq. (64). When n is an integer, $I_n(\xi) = I_{-n}(\xi)$. The modified Bessel functions of

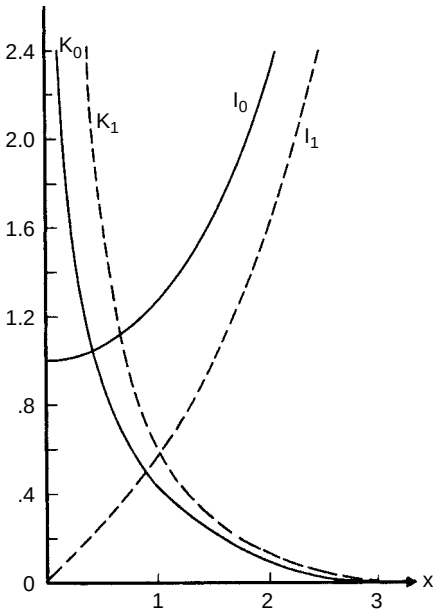


Fig. 14 Modified Bessel functions, $K_n(x)$ and $I_n(x)$ (Abramowitz and Stegun, 1964)

the second kind, $K_n(\xi)$, are defined by

$$K_n(\xi) \equiv \frac{\pi}{2}\left\{\frac{I_n(\xi) - I_{-n}(\xi)}{\sin n\pi}\right\}.$$

The modified Bessel functions of the second kind are well behaved for all values of n. Sketches of several modified Bessel functions are given in Fig. 14.

4.6 Spherical Bessel Functions

Solutions of the radial equation, Eq. (48), for $k^2 =$ constant are obtained by comparing the radial equation with the Bessel differential equation; these solutions are called spherical Bessel functions and have the form

$$R(\xi) = Aj_n(\xi) = \left(\frac{\pi}{2\xi}\right)^{1/2} J_{n+1/2}(\xi) \text{ for } \xi = kr.$$

Spherical Bessel functions are often used in quantum mechanics and in other areas of physics. Sketches of several spherical Bessel functions, spherical Neumann functions, and spherical modified Bessel functions are respectively given in Figs. 15, 16, and 17.

EXAMPLE 7: vibrations of a circular membrane: The displacement $u(r, \theta, t)$ of a stretched circular membrane with mass per unit area μ and under tension T satisfies the two-dimensional mechanical wave equation in plane polar coordinates (r, θ), which may be written in the following form:

$$\frac{1}{r}\left\{\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\frac{1}{r}\frac{\partial u}{\partial \theta}\right)\right\} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2}.$$

The speed of the wave motion is defined by $v = \sqrt{T/\mu}$. Develop the solution of this equation for a vibrating drum head.

Solution: Separating the variables $u(r, \theta, t) = R(r)\Theta(\theta)T(t)$ with separation constants $-\alpha^2$ and $-n^2$ yields the following three ordinary differential equations:

$$\frac{d^2T}{dt^2} + \omega^2 T = 0,$$

with solution $T(t) = A\cos\omega t + B\sin\omega t$, where $\omega^2 \equiv v^2\alpha^2$;

$$\frac{d^2\Theta}{d\theta^2} + n^2\Theta = 0,$$

with solution $\Theta(\theta) = C\cos n\theta + D\sin n\theta$; and

$$\xi^2\frac{d^2R}{d\xi^2} + \xi\frac{dR}{d\xi} + (\xi^2 - n^2)R = 0,$$

where $\xi \equiv \alpha r$. The solution of the last equation is $R(\xi) = EJ_n(\xi) + FN_n(\xi)$ since it is the Bessel differential equation. In addition, it is required that the solution be finite at $\xi = 0$; hence, F is set equal to

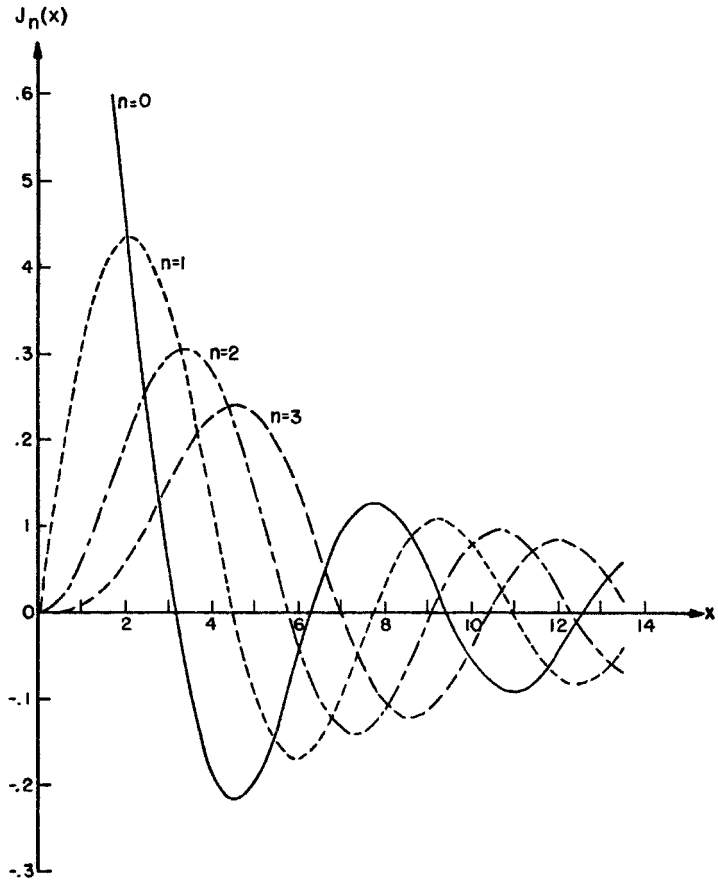


Fig. 15 Spherical Bessel functions of the first kind, $j_n(x)$ (Abramowitz and Stegun, 1964)

zero since $N_n(\xi)$ approaches infinity as ξ approaches zero. The general solution for the motion of the drum head is therefore given by

$$u = [A\cos\omega t + B\sin\omega t] \times [C\cos n\phi + D\sin n\phi]EJ_n(\xi).$$

Since the membrane is fixed (no vibration, $u = 0$) around the edge where $r = b$ (radius of the head), the drum head vibrates in circular modes such that $EJ_n(\xi) = 0$. The nodes are located at $\alpha r = \xi_k$, where ξ_k are the values of ξ for which $J_n(\xi)$ has a zero. A single term in the solution corresponds to a standing wave whose modes are concentric circles, and the complete solution is obtained by summing over all such modes of vibration.

5 Other Special Functions used in Mathematical Physics

As explained in the Introduction of this article, the list of special functions is

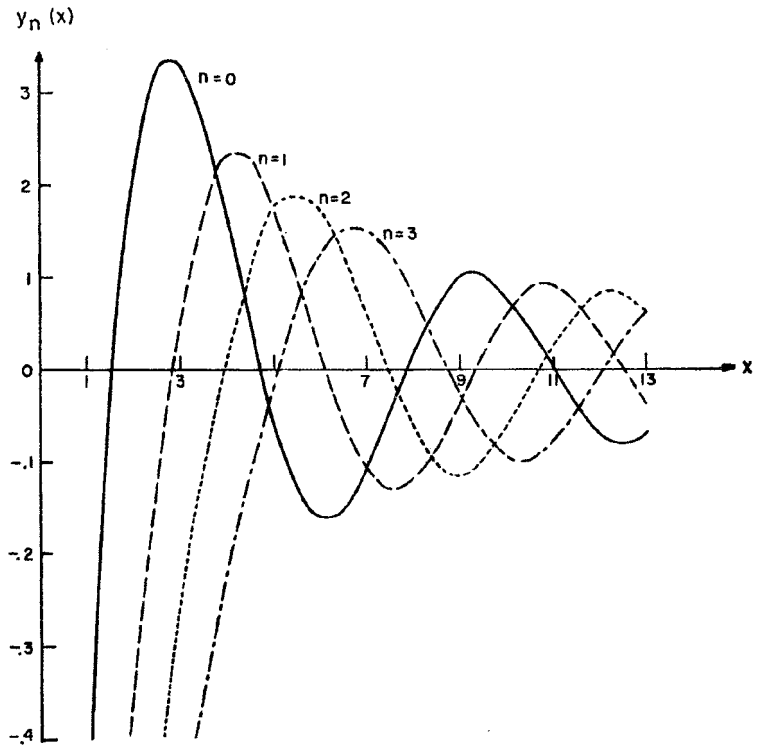


Fig. 16 Spherical Bessel functions of the second kind, $y_n(x)$ (Abramowitz and Stegun, 1964)

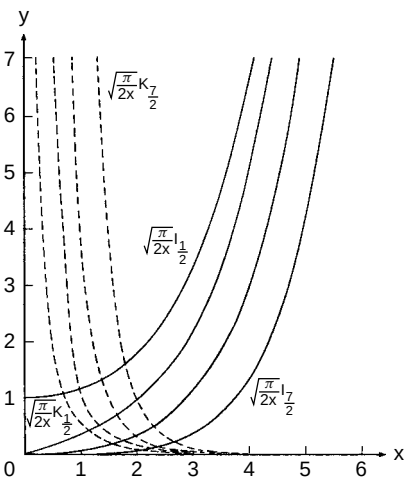


Fig. 17 Modified spherical Bessel functions of the first and second kinds (Abramowitz and Stegun, 1964)

extensive. The *Handbook of Mathematical Functions* (Abramowitz and Stegun, 1964) contains a fairly comprehensive list of special functions. The focus in this article has been on special functions that are widely used in mathematical physics to solve classes of problems whose formulations involve special cases of the Helmholtz differential equation. In general, special functions may be classified as Type 1, those special functions that satisfy a differential equation, or Type 2, special functions that do not satisfy a differential equation; for example, the gamma function is a Type 2 special function. In Secs. 5.1 and 5.2, we summarize some other special functions that are used in mathematical physics.

5.1 Some Other Special Functions – Type 1

Some other special functions that satisfy a differential equation are the following:

1. Airy functions are solutions of the Airy differential equation, which has the form $y'' - xy = 0$; the Airy differential equation characterizes constant-force–type problems in quantum mechanical and in elementary particle physics.
2. Mathieu functions are solutions of the Mathieu differential equation, which has the form $y'' + (a - 2b\cos 2x)y = 0$; the Mathieu differential equation results when a cosine-type potential is substituted into the one-dimensional time-independent Schrödinger wave equation.
3. Parabolic cylinder functions are connected with confluent hypergeometric functions and with Hermite polynomials; they are solutions of differential equations of the form $y'' + (ax^2 + bx + c)y = 0$.

Many bound-state and collision problems in classical and quantum mechanics as well as in other areas of physics involve integrals of the form

$$\int R(x, y)\, dx. \tag{65}$$

When $R(x,y)$ is a rational function of x and y and $y^2 = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, where $a_4 \neq 0$ or $a_4 = 0$ and $a_3 \neq 0$, the integral in Eq. (65) is called an elliptic integral. Complete elliptic integrals of the first and second kind are respectively defined as

$$K(m) = \int_0^{\pi/2} (1 - m\sin^2\theta)^{-1/2} d\theta$$

and

$$E(m) = \int_0^{\pi/2} (1 - m\sin^2\theta)^{1/2} d\theta.$$

Analyzing the motion of a simple pendulum involves an elliptic integral of the first kind. Elliptic integrals can be evaluated directly by use of series expansions or computers. Their importance in mathematical physics, however, is related to their appearance in the solution of physical problems involving certain nonlinear differential equations. Elliptic integrals are special cases of the hypergeometric functions since

$$K(m) = \frac{\pi}{2} F(\tfrac{1}{2}, \tfrac{1}{2}, 1; m) \quad \text{and}$$

$$E(m) = \frac{\pi}{2} F(-\tfrac{1}{2}, \tfrac{1}{2}, 1; m).$$

The *Handbook of Mathematical Functions* (Abramowitz and Stegun, 1964) is a good reference for additional information on elliptic integrals.

5.2 Some Other Special Functions – Type 2

Some special functions used in mathematical physics that do not satisfy a differential equation are Einstein and Debye functions, which are used in representing the specific heats of solids due to lattice vibrations; error function (defined in Sec. 3.1); gamma function; and beta function. The last two functions are widely used in many areas of mathematical physics and are now summarized.

The factorial, $n!$, is defined as

$$n! = n(n-1)\ldots 2 \cdot 1 = \int_0^\infty e^{-t} t^n\, dt$$

for integer values of n. Note that $0! = 1$, and $n! = \pm\infty$ if n equals a negative integer. The gamma function, Γ, is a generalization of

the factorial to cases of noninteger values for n. The Euler definition of the gamma function is

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt \quad \text{for } \mathrm{Re}(z) > 0. \tag{66}$$

In Eq. (66), $\mathrm{Re}(z)$ denotes the real part of $z = x + iy$. Note that $\Gamma(\frac{1}{2}) = \pi^{1/2}$. Evaluating the integral in Eq. (66) by parts yields the recurrence relation for the gamma function, $\Gamma(z+1) = z\Gamma(z)$. If z is a positive integer n, then $\Gamma(z+1)$ equals $n!$.

The gamma function is used to express, in compact form, solutions of many problems of mathematical physics. The gamma function, however, does not satisfy a differential equation that is related to a physical problem; in fact, the gamma function does not satisfy any differential equation with rational coefficients. A sketch of $\Gamma(x)$ for some positive and negative values is given in Fig. 18.

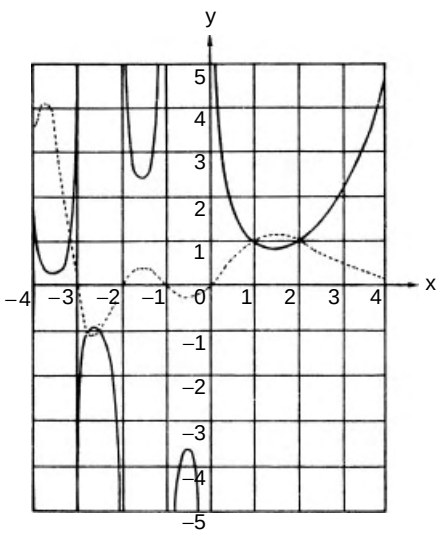


Fig. 18 Gamma function: solid curve, $\Gamma(x)$; dashed curve, $1/\Gamma(x)$ (Abramowitz and Stegun, 1964)

The beta function, $B(x,y)$, is defined by use of an integral, and it involves a simple and useful combination of gamma functions; it has the form

$$B(p,q) \equiv \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \quad \text{for } \mathrm{Re}(p) > 0,\ \mathrm{Re}(q) > 0.$$

Note that $B(p,q) = B(q,p)$. The beta function is frequently used in high-energy particle physics as well as other areas of mathematical physics.

Glossary

Confluent Hypergeometric Function: A solution of the confluent hypergeometric differential equation, which is a second-order ordinary differential equation with a regular singularity at $x = 0$ and an irregular singularity at $x = \infty$. Laguerre and associated Laguerre polynomials and Hermite polynomials are special cases of the confluent hypergeometric function.

Gamma Function: A generalization, to noninteger values of n, of the factorial $n!$.

Generating Function: A function $\mathscr{G}(x,t)$ that, when expanded in a power series with respect to t, contains the functions (set of polynomials) to be generated as coefficients of the parameter t.

Hypergeometric Function: A generalized geometric series that is the solution of the hypergeometric differential equation, which is a second-order ordinary differential equation with regular singular points at 0,1, and infinity. For example,

the Legendre polynomials and associated Legendre functions are special cases of the hypergeometric function.

Rodrigues Formula: A formula used to generate an orthogonal polynomial by taking the *n*th derivative of an elementary function.

Special Functions: Higher transcendental functions used in mathematical physics.

Sturm–Liouville Theory: The theory devoted to determining the dependence of eigenfunctions on eigenvalues and the dependence of eigenvalues on boundary conditions imposed on eigenfunction solutions of the Sturm–Liouville equation.

List of Works Cited

Abramowitz, M., Stegun, I. A. (Eds.) (1964), *Handbook of Mathematical Functions*, New York: Dover Publications.

Courant, R., Hilbert, D. (1953), *Methods of Mathematical Physics*, Vol. 1, New York: Interscience Publishers.

Infeld, L., Hull, T. E. (1951), *Rev. Mod. Phys.* **23**, 21–68.

Miller, W. (1968), *Lie Theory and Special Functions*, New York: Academic.

Schrödinger, E. (1940), *Proc. Roy. Irish Acad.* **A46**, 9.

Sommerfeld, A. (1949), *Partial Differential Equations in Physics*, New York: Academic.

Talman, J. D. (1968), *Special Functions: A Group Theoretic Approach*, New York: W. A. Benjamin, Inc.

Vilenkin, N. J. (1968), *Special Functions and the Theory of Group Representations*, Translations of Mathematical Monographs, Vol. 22, Providence, RI: American Mathematical Society.

Whitaker, E. T., Watson, G. N. (1947), *Modern Analysis*, New York: Macmillan Company.

Further Reading

Arfken, G., Weber, H. J. (1995), *Mathematical Methods for Physicists*, New York: Academic.

Beckmann, P. (1973), *Orthogonal Polynomials for Engineers and Physicists*, Boulder, CO: The Golem Press.

Erdelyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G. (Eds.) (1953), *Higher Transcendental Functions*, Vols. 1, 2, and 3, New York: McGraw-Hill.

Gilmore, R. (1974), *Lie Groups, Lie Algebras, and Some of Their Applications*, New York: Wiley.

Morse, P. M., Feshbach, H. (1953), *Methods of Theoretical Physics*, Vols. 1 and 2, New York: McGraw-Hill.

Sattinger, D. H., Weaver, O. L. (1986), *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, New York: Springer-Verlag.

Wang, Z. X., Guo, D. R. (1989), *Special Functions*, Singapore: World Scientific.

Stochastic Processes

Melvin Lax
Physics Department, City College of the City University of New York, New York, USA, and Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, USA

Introduction ***514***
What Is a Stochastic Process? *514*
Kinds of Dynamical Systems *515*
An Approach to the Spectrum of Stationary Stochastic Processes *516*
1 The Treatment of Stationary Stochastic Processes ***517***
1.1 Correlation Functions and the Regression Theorem *517*
2 Spectral Measurements and Correlations ***519***
2.1 Introduction: An Approach to the Spectrum of Stochastic Processes *519*
2.2 Standard Engineering Definition of Noise *519*
2.3 The Wiener–Khinchin Theorem *521*
2.4 Noise Measurements Using Filters *521*
2.4.1 A Realizable Filter *522*
3 Thermal Noise ***523***
3.1 Johnson Noise *523*
3.2 Equipartition *524*
3.3 Thermodynamic Derivation of Johnson Noise *525*
3.4 Nyquist's Theorem *525*
3.5 Relation between Nyquist and Einstein *527*
4 Shot Noise ***528***
4.1 The Poisson Process *528*
4.2 Pure Shot Noise *529*
4.3 Generalized Characteristic Functions *530*
4.4 Rice's Generalized Campbell Theorem *531*
4.5 Campbell's Theorems *532*
4.6 Equivalence of Shot Noise to the Poisson Process *532*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

4.7 Transit-Time Effects *533*
4.8 Generation–Recombination Process *533*
5 Resistance Modulation Fluctuations *535*
5.1 Conductivity Fluctuations *535*
5.2 A Thermodynamic Treatment of Total Carrier Fluctuations *536*
5.3 Einstein Derivation of Carrier Fluctuations with Traps *537*
5.4 The Spectrum of Resistance Modulation Fluctuations *539*
6 Concentration Fluctuations in Semiconductors *540*
6.1 General Theory of Concentration Fluctuations *540*
6.1.1 Application to Semiconductors with Electrons, Holes, and Traps *541*
6.2 The Influence of Drift and Diffusion on Resistance Modulation Noise *541*
7 Langevin Processes *544*
7.1 Simplicity of Langevin Processes *544*
7.2 Relation to the Fokker–Planck Equation *545*
7.3 An Exactly Solvable Gaussian Example *546*
7.4 Stochastic Integrals: The Ito–Stratonovich Controversy *548*
8 Further Contributions to Stochastic Processes *552*
8.1 Overview *552*
8.2 Random-Walk Problems *552*
8.3 Linear with Time-Dependent Decay *552*
8.4 The Nonlinear (Fokker–Planck) Case: Reaction-Rate Theory *553*
8.5 Stochastic Resonance *556*
8.6 Self-Sustained Oscillators *557*
Glossary *559*
List of Works Cited *559*
Further Reading *562*

Introduction

What Is a Stochastic Process?

If there are n people in a room and we ask what is the probability that any two have the same birthday, we have a pure probability problem since there is no change with time. Person j has birthday b_j, which can take any of 365 values, presumably with equal probability. There are n random variables, each of which has a discrete state space of 365 points. And with a modest effort, one can show that a coincidence of birthdays will acquire a probability more than 1/2 if $n > 25$.

If a man starts with a capital of \$100 and aims to end up with \$1000 by betting a dollar once a minute on coin tosses, we are dealing with a stochastic process because we have a random variable (the bettor's current capital) that changes (makes transitions) with time. The state space is discrete with the integral values from 0 to 1000, and the time variable is also discrete. This problem can be mapped onto a random-walk problem in which the man's capital is indicated by a position on a line, with absorbing barriers at each end.

The total number of electrons in a cell, in a semiconductor, increases by one in a

generation process and decreases by one in a recombination processes. In this case, we have a stochastic process for a single, discrete random variable N as a function of a continuous time.

The current flowing through a resistor is a random process involving a continuous random variable $I(t)$ as a function of a continuous time.

The concept of a stochastic process is, of course, immediately extended to the case of several, or even an infinite number of, random variables. Each cell in a crystal could contain atomic-like states whose occupancy is a random variable, indexed by a discrete triple index. A carrier density, $n(\mathbf{r},t)$, provides a continuous index to an infinite number of continuous-time variables. In this case, carriers can be generated, recombined, or modified by drift and diffusion.

Kinds of Dynamical Systems

The simplest dynamical systems are linear. When the noise is weak, a stable nonlinear system will stray only slightly from a motion, possibly a time-independent state of the nonlinear state. Many important electronic and semiconductor devices are of this nature. We attempt to display some techniques that are universally applicable to fluctuations from a stable nonequilibrium state. We summarize briefly, here and in Sec. 8, problems that violate the requirements just mentioned that permit quasilinear solutions.

We are usually concerned with physical systems in which the noise is intrinsic to the system and related to the dissipative response of the system. There is also a vast literature on the response of nonlinear systems to external noise whose properties are unrelated to the system. See an early summary by Deutsch (1962), and the book by Van Kampen (1992) that provides an overview of stochastic processes with an emphasis on nonlinear processes.

Aside from the distinction between weak and strong noise, there is also a distinction between noise that is white (that is, possesses a flat spectrum) and colored noise whose spectrum is frequency dependent.

If the noise source is white, the problem, even if nonlinear, reduces to a Markoffian one. A Markoff process is a stochastic process whose future depends only on the latest information and is independent of prior information. This absence of memory is guaranteed by the fact that the Fourier transform of a noise constant in frequency is a delta function in time. The zero width of the delta function is the length of the memory time. If $a(t)$ is a random process, and t_j constitute a set of increasing times $t_0 < t_1 < t \cdots < t_n$, then the probability density of $a(t_n)$ conditional on all the values at all the earlier times obeys

$$P(a_n|a_{n-1}, a_{n-2}, \ldots a_0) = P(a_n|a_{n-1}) \tag{1}$$

or a Markoffian process, like a person who remembers only the last thing he was told. Markoff processes then have the simplicity that a general multiple-time probability can be written

$$P(t_n, t_{n-1}, \ldots t_0) = P(t_n|t_{n-1})P(t_{n-1}|t_{n-2}|) \ldots P(t_1|t_0)P(t_0) \tag{2}$$

in terms of an initial probability multiplied by a series of two-time transition probabilities. The latter must obey a self-consistency condition known as the Chapman–Kolmogorov relation to insure the Markoffian nature of the process (see Sec. 7.2).

If a system is linear (or quasilinear, with a weak noise source), analytic solutions

are possible for the probability densities even in the presence of colored noise sources. If the system is nonlinear, but with a white noise source, the probability will obey a generalized Fokker–Planck equation. The problem is thus reduced to (nontrivial) analysis. If the noise sources are also Gaussian, the problem reduces to a standard Fokker–Planck equation (see Sec. 7.2). Then analytic and numerical methods are feasible if the number of variables is small.

A system can be described as stationary if it is stable and none of the parameters describing the system is time dependent. In this case, there will be a steady (time independent) state, and all probability distributions that depend on variables at more than one time will be a function only of the time differences. We added the caveat about stability to exclude such cases as a free particle executing a Brownian motion that permits excursions that will grow indefinitely. Another example is that of an autonomous self-sustained oscillator. The latter has been shown by Lax (1968a, 1968b) always to contain one mode of instability that is usually related simply to an overall phase of the system that can grow at will. (The reason is given in Sec. 8.5.)

Quantum systems also require special handling. By the use of ordered operators, Lax (1968a,b) and Haken and Weidlich (1967) showed that quantum stochastic processes could be exactly expressed in terms of associated c-number ("classical") processes.

An Approach to the Spectrum of Stationary Stochastic Processes

Many experimental articles on semiconductor devices prior to 1960 included a systematic discussion of transport in these devices but an *ad hoc* theory of the associated noise. Lax (1960) recognized that a large class of devices that involved weak fluctuations from a (possibly nonequilibrium) steady state could be treated by a general approach that related noise to transport by an Einstein relation and a regression theorem, outlined as follows:

1. The response of most nonlinear systems to noise sources is sufficiently weak to permit a quasilinear approximation. Generalization of this procedure to self-sustained oscillators including quantum lasers is reviewed elsewhere (Lax, 1991).
2. The time decay of a correlation such as $\langle A(t)B(0)\rangle$ is the same as the time dependence of $\langle A(t)\rangle$, the dissipative relaxation of an initial deviation from the steady state. This is referred to as Onsager's regression hypothesis for deviation from an equilibrium state. Lax (1963, 1968b) proposed it as a theorem for Markoffian systems for fluctuations from a steady, possibly nonequilibrium, steady state. The angular brackets denote an average over an ensemble. In probability, $\langle$something$\rangle$ is often written E[something] and is called an expectation.
3. To the equations of time dependence determined from transport theory, we must add an initial condition, namely the value of $\langle A(0)B(0)\rangle$. For fluctuations from the equilibrium state, the second moments are known thermodynamically (Callen, 1985):

$$\langle A(0)B(0)\rangle \equiv \langle AB\rangle = \frac{-k\partial\langle A\rangle}{\partial F_B}. \quad (3)$$

When regarded as a thermodynamic quantity, it is customary to rewrite $\langle A\rangle$ simply as A. Here, F_B is the force "conjugate" to the variable B in the sense

that the negative of the pressure, $-P$, is conjugate to the volume V. (The negative sign is necessary since pressure decreases volume rather than increasing it.)

Sometimes, we must deal with quasi-equilibrium states. For example, in a semiconductor, the electrons come to equilibrium with each other rapidly. So do the holes. But the radiative and nonradiative processes that produce recombination and generation of electron–hole pairs are slow. It is then customary to introduce quasi-Fermi levels, one for the electrons and one for the holes. In that case, Eq. (3) is useful, even though complete equilibrium is absent. In more complicated cases, one must make use of more detailed knowledge of the noise processes involved.

In Sec. 1 we discuss the above-mentioned approach to obtaining correlation functions for stochastic processes. In Sec. 2 we develop the relation between the Fourier transform of the correlation functions, the ideal noise measurement, and real noise measurements. In Sec. 3 we discuss the set of noise sources that arise from thermal origins. In Sec. 4 we discuss the "shot noise sources" that arise from the discrete nature of the charges that give rise to this noise. In Sec. 5 we discuss examples of these noise sources. In Sec. 6 we discuss noise in homogeneous semiconductors. In Sec. 7 we review the Langevin noise-source approach and describe its relation to the Fokker–Planck approach. Pitfalls in the use of the Ito definition of a stochastic integral are elucidated. In Sec. 8 we discuss a variety of nonlinear examples including random-walk theory, linear noise with time-dependent decay, reaction-rate theory, stochastic resonance, and self-sustained oscillators.

1 The Treatment of Stationary Stochastic Processes

1.1 Correlation Functions and the Regression Theorem

When equilibrium statistics cannot be utilized, the second moments can be obtained by using a generalization of the Einstein relation.

If $\mathbf{a}(t)$ is a set of random variables, $a_1(t), a_2(t), \ldots$, then Lax (1960) has derived a generalized Einstein relation of the form

$$\frac{d}{dt}\langle \mathbf{aa}\rangle = 2\langle \mathbf{D}(\mathbf{a}, t)\rangle + \langle \mathbf{aA}(\mathbf{a}, t)\rangle + \langle \mathbf{A}(\mathbf{a}, t)\mathbf{a}\rangle. \tag{4}$$

Here, $\mathbf{aa}$ is a dyadic abbreviation for a set of products $a_i a_j$, and $\mathbf{aA}$ is a corresponding abbreviation. The "drift vector," $\mathbf{A}(\mathbf{a},t)$, is defined by

$$\mathbf{A}(\mathbf{a}, t) \equiv \lim_{\Delta t \to 0} \frac{\langle \mathbf{a}(t + \Delta t) - \mathbf{a}(t)\rangle}{\Delta t} \text{ so that}$$

$$\frac{d\langle \mathbf{a}(t)\rangle}{dt} = \langle \mathbf{A}(\mathbf{a}, t)\rangle, \tag{5}$$

and the "diffusion matrix," $\mathbf{D}(\mathbf{a},t)$, is defined by

$$2!\mathbf{D}(\mathbf{a}) \equiv \lim_{\Delta t \to 0} \frac{\langle \Delta\mathbf{a}(t)\Delta\mathbf{a}(t)\rangle}{\Delta t}. \tag{6}$$

We understand the abbreviation

$$\Delta\mathbf{a}(t) = \mathbf{a}(t + \Delta t) - \mathbf{a}(t). \tag{7}$$

The 11 component

$$D_{11} = \frac{\langle [\Delta a_1(t)]^2\rangle}{2\Delta t} \tag{8}$$

reduces to the conventional diffusion constant when a_1 has the meaning of a position, x. If the diffusion coefficient D_{11}, labeled D_{xx} to remind us of its nature, is

a constant, the mean square displacement $\langle[\Delta x]^2\rangle$ grows linearly with the time interval, the conventional Brownian-motion result.

The evaluation of the averages in Eqs. (5) and (6) will make use of the nature of the noise process. For example, the system may be in interaction with a thermal reservoir. Or the noise may have shot-noise character, that is, it may arise because discrete particles are involved, and all the occupancies of states must be integers. We illustrate these mechanisms in Secs. 3 and 4.

The important point for the purposes of this section is that two major simplifications are possible. In the quasilinear case, there will be a set of associated stationary variables $\mathbf{a}_0$ and a set of small deviations

$$\boldsymbol{\alpha}(t) \equiv \mathbf{a}(\mathbf{t}) - \mathbf{a}_0. \tag{9}$$

The stationary points $\mathbf{a}_0$ are those at which the drift vectors vanish:

$$\mathbf{A}(\mathbf{a}_0) = 0. \tag{10}$$

Expanding around the steady operating point yields

$$\mathbf{A}(\mathbf{a}) \approx -\boldsymbol{\Lambda} \cdot \boldsymbol{\alpha}, \langle \mathbf{A}(\mathbf{a})\rangle \approx -\boldsymbol{\Lambda} \cdot \langle\boldsymbol{\alpha}\rangle,$$
$$\frac{d\langle\boldsymbol{\alpha}(t)\rangle}{dt} = -\boldsymbol{\Lambda} \cdot \langle\boldsymbol{\alpha}(t)\rangle, \tag{11}$$
$$\mathbf{D}(\mathbf{a}) \approx \mathbf{D}(\mathbf{a}_0) \equiv \mathbf{D}. \tag{12}$$

In that case, the generalized Einstein relation, Eq. (4), simplifies to

$$\frac{d\langle\boldsymbol{\alpha}\boldsymbol{\alpha}\rangle}{dt} = 2\mathbf{D} - \boldsymbol{\Lambda} \cdot \langle\boldsymbol{\alpha}\boldsymbol{\alpha}\rangle - \langle\boldsymbol{\alpha}\boldsymbol{\alpha}\rangle \cdot \boldsymbol{\Lambda}^{\dagger} \tag{13}$$

where $\boldsymbol{\Lambda}^{\dagger}$ is the transpose of $\boldsymbol{\Lambda}$. (In the complex case, the Hermitian adjoint would be needed, which is why we use the dagger symbol even though we are dealing with real variables here.)

The second major simplification is that we are dealing with fluctuations from a stationary state. In that case, a correlation such as $\langle\alpha_i(t)\alpha_j(t)\rangle$ is a function only of the time difference $t - t$. Hence, the time derivative vanishes, and we arrive at the stationary Einstein relation

$$2\mathbf{D} = \boldsymbol{\Lambda} \cdot \langle\boldsymbol{\alpha}\boldsymbol{\alpha}\rangle + \langle\boldsymbol{\alpha}\boldsymbol{\alpha}\rangle \cdot \boldsymbol{\Lambda}^{\dagger}. \tag{14}$$

For the case of a single variable, Eq. (14) reduces to the original Einstein relation between the diffusion constant and mobility; see Eq. (89).

In the presence of time reversal, the two terms on the right-hand side in Eq. (14) have been shown to be equal; see Onsager (1931) and Eq. (6.18) of Lax (1960). Thus the initial value $\langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle$ is given by

$$\langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle = \boldsymbol{\Lambda}^{-1} \cdot \mathbf{D}. \tag{15}$$

Equation (11) yields the conditional mean motion (transport)

$$\langle\boldsymbol{\alpha}(t)\rangle_{\boldsymbol{\alpha}(0)} = \exp(-\boldsymbol{\Lambda}t) \cdot \boldsymbol{\alpha}(0). \tag{16}$$

By calculating the autocorrelation in two steps – first an average conditional on the initial condition, and second an average over the initial condition,

$$\langle\boldsymbol{\alpha}(t)\boldsymbol{\alpha}(0)\rangle = \langle\langle\boldsymbol{\alpha}(t)\rangle_{\boldsymbol{\alpha}(0)}\boldsymbol{\alpha}(0)\rangle$$
$$= \exp(-\boldsymbol{\Lambda}t) \cdot \langle\boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle \tag{17}$$

– we have established the regression theorem, that the fluctuation contains the same time dependence as the transport. The Markoffian assumption was tacitly made, since only in the Markoffian case is the conditional average independent of any earlier history that could affect the second average.

Equation (16) is restricted to positive times, and so necessarily is Eq. (17).

However, stationarity permits us to write

$$\langle \boldsymbol{\alpha}(-t)\boldsymbol{\alpha}(0)\rangle = \langle \boldsymbol{\alpha}(0)\boldsymbol{\alpha}(t)\rangle = \langle \boldsymbol{\alpha}(0)\boldsymbol{\alpha}(0)\rangle \cdot \exp(-\boldsymbol{\Lambda}^{\dagger} t). \quad (18)$$

2 Spectral Measurements and Correlations

2.1 Introduction: An Approach to the Spectrum of Stochastic Processes

In this section we compare three definitions of noise: The standard engineering (SE) definition takes a Fourier transform over a finite time interval, squares it, divides by the time, and then takes the limit as the time approaches infinity. The second definition is the Fourier transform of the autocorrelation function. The equality between these two definitions is known as the Wiener (1930)–Khinchin (1934) theorem. The third procedure, introduced by Lax (1968a), passes the signal through a realizable filter of finite bandwidth, squares it, and averages over some large finite time. This simulates an actual measurement of the noise spectrum. As the bandwidth is allowed to approach zero, the result will (aside from a normalization factor) approach the ideal value of the two preceding definitions.

2.2 Standard Engineering Definition of Noise

The spectrum of noise $G_s(\alpha, \omega)$ in a single random variable $a(t)$ is a measure of the fluctuation energy in the frequency interval $[\omega, \omega + d\omega]$ associated with the fluctuating part $\alpha(t)$, where

$$\alpha(t) = a(t) - \langle a(t)\rangle; \ \langle \alpha(t)\rangle = 0. \quad (19)$$

The SE definition of noise, denoted by G_s, is chosen to obey the normalization

$$\int_0^{\infty} G_s(\alpha, \omega) df = \langle |\alpha(t)|^2\rangle, \quad (20)$$

because it is customary in engineering to emphasize positive frequencies $f = (\omega/2\pi) > 0$ only. For this reason, we adopt the definition

$$\frac{1}{2}G_s(\alpha, \omega) \equiv \frac{1}{2}G_s(\omega) \equiv (\alpha^2)_\omega = \lim_{T\to\infty} \frac{1}{2T}\left\langle \left| \int_{-T}^{T} \alpha(t) e^{-j\omega t} dt \right|^2 \right\rangle \quad (21)$$

and verify the normalization later.

The letter j is used to denote imaginary unit customarily in electrical engineering. The SE convention is that $\exp(j\omega t)$ describes positive frequencies and $R + j\omega L + 1/j\omega C$ is the impedance of a series circuit of a resistance R, an inductance L, and a capacity C. Because propagating waves are described by $\exp(ikx - i\omega t)$ in physical problems, we regard $\exp(-i\omega t)$ as describing positive frequencies so that the physics convention is equivalent to setting $j = -i$ consistently.

In this definition, Eq. (21), the interval on t is truncated to the region $-T \le t \le T$, its Fourier transform is taken, and the result is squared. Since such a measurement would attempt to filter out one component ω and square it, this definition is reasonable. Since the integral is not expected to converge, it must be normalized by T. What is not clear yet is why one divides by T rather than T^2. The angular brackets denote an ensemble average. It is curious that both the limit $T \to \infty$ and an ensemble average are taken. For an ergodic process, a time average – by definition – is equal to an ensemble average. One might guess

that the ensemble average could have been eliminated – that its use is primarily for convenience of computation. The above assumption is, however, wrong! Middleton (1960) shows that if the ensemble average is not performed, substantial fluctuations occur in the value of $G_s(\omega)$.

Presumably, this sensitivity occurs because we are asking for the noise at a precise frequency. Because of the Fourier relation between frequency and time, a measurement accurate to $\Delta\omega$ requires a time $t > 1/\Delta\omega$. Realistic noise measurements, to be discussed below, using filters of finite width, are presumably ergodic.

The above definition assumes that the noise is stationary. A more general definition of the noise at a frequency ω and time t is given by the Fourier transform

$$\frac{1}{2}G(\alpha, \omega, t) = \int_{-\infty}^{\infty} e^{j\omega\tau} R(\tau, t) d\tau \qquad (22)$$

of the Wigner (1932)–Moyal (1949) type of autocorrelation function,

$$R(\tau, t) = \left\langle \alpha\left(\frac{t+\tau}{2}\right)^* \alpha\left(\frac{t-\tau}{2}\right)\right\rangle, \quad (23)$$

guaranteed to yield a real $G(\alpha, \omega, t)$. For future use, we note that the inverse of Eq. (22) is

$$\langle \alpha(v)^* \alpha(u)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} e^{-j\omega(v-u)} \times G\left(\alpha, \omega, \frac{(u+v)}{2}\right) d\omega. \qquad (24)$$

The nonstationary case is discussed further in Lax (1968a). In the stationary case, to which we now restrict ourself in this section, the autocorrelation, Eq. (23), is invariant under a shift of time origin, hence to a change in t. Thus both $R(t, \tau)$ and $G(\alpha, \omega, t)$ are independent of t. Equation (22) can then be simplified to

$$G(\alpha, \omega) = 2\int_{-\infty}^{\infty} e^{j\omega t}\langle \alpha^*(t)\alpha(0)\rangle\, dt. \quad (25)$$

According to the SE definition, Eq. (21), the noise is manifestly real. Equation (25) can also be shown to be real by taking its complex conjugate and introducing $-t$ as a new variable of integration. This conclusion remains true even if α is a complex variable, or a non-Hermitian operator. In the classical case when time-reversal invariance holds, the ensemble average is an even function of time and

$$G(\alpha, \omega) = 4\int_0^{\infty} \cos\omega t \langle \alpha^*(t)\alpha(0)\rangle\, dt = 4\mathrm{Re}\int_0^{\infty} \exp(j\omega t)\langle \alpha^*(t)\alpha(0)\rangle\, dt. \qquad (26)$$

It is easy to verify from Eq. (25) that the normalization integral

$$\int_{-\infty}^{\infty} \frac{G(\alpha, \omega) d\omega}{4\pi} = \langle |\alpha(0)|^2\rangle \qquad (27)$$

is consistent with our original aim in Eq. (20).

If we have a classical variable V expressed as a linear combination of variables,

$$V(t) = \sum_m C_m \alpha_m(t), \qquad (28)$$

then the noise in V is given by

$$G(V, f) = \sum_{mn} C_m^* C_n G_{mn}(\boldsymbol{\alpha}, f), \qquad (29)$$

where the noise matrix

$$G_{mn}(\boldsymbol{\alpha}) = 4\mathrm{Re}\int_0^{\infty} \exp(j\omega t)\langle \alpha_m^*(t)\alpha_n(0)\rangle\, dt. \qquad (30)$$

If the variables $\boldsymbol{\alpha}$ have the decay matrix $\boldsymbol{\Lambda}$ of Eq. (11), the noise correlation G_{mn} is

given by

$$G_{mn}(\boldsymbol{\alpha}) = 4\mathrm{Re}\sum_s \quad [(-j\omega\mathbf{1} + \boldsymbol{\Lambda}^*)^{-1}]_{ms}\langle\alpha_s\alpha_n\rangle. \tag{31}$$

2.3 The Wiener–Khinchin Theorem

We shall prove the Wiener–Khinchin theorem by evaluating $G_s(\alpha, \omega)$ in terms of $G(\alpha, \omega)$:

$$G_s(\alpha, \omega) = \lim_{T\to\infty} \frac{1}{T} \times \left\langle \int_{-T}^{T} \alpha(v)^* \exp(j\omega v)dv \times \int_{-T}^{T} \alpha(u)\exp(-j\omega u)du \right\rangle. \tag{32}$$

In the stationary case (for which the theorem is valid), Eq. (24) can be written

$$\langle\alpha(v)^*\alpha(u)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} e^{j\omega(v-u)}G(\alpha, \omega)d\omega. \tag{33}$$

If Eq. (33) is inserted into Eq. (32), with ω replaced by ω', we have

$$\begin{aligned} G_s(\alpha, \omega) &= \lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} G(\alpha, \omega')d\omega' \times \frac{1}{4\pi}\left|\int_{-T}^{T}\exp[j(\omega-\omega')u]du\right|^2 \\ &= \lim_{T\to\infty}\int_{-\infty}^{\infty} G(\alpha, \omega')d\omega' \times \frac{\sin^2(\omega'-\omega)T}{\pi(\omega'-\omega)^2T} = G(\alpha, \omega). \end{aligned} \tag{34}$$

This completes our heuristic proof of the Wiener–Khinchin theorem. The last step moved the limiting procedure under the integral sign and used

$$\lim_{T\to\infty}\frac{\sin^2(\omega'-\omega)T}{\pi(\omega'-\omega)^2T} = \delta(\omega'-\omega). \tag{35}$$

The appropriateness of the limit, Eq. (35), is based on the facts that

1. the integral of the left-hand side, for any T, is 1;
2. the width of the function is of order $1/T$, and the maximum height, at $\omega' = \omega$, is of order T. This function becomes very tall and narrow. An integration against this function by any function $G(\omega')$ of bounded variation will be sensitive only to its value at the peak $\omega' = \omega$.

Note that Eq. (33) with $u = v = t$ leads to the normalization condition

$$\langle\alpha(t)^*\alpha(t)\rangle = \frac{1}{4\pi}\int_{-\infty}^{\infty} G(\alpha, \omega)d\omega. \tag{36}$$

This normalization is equivalent to the customary choice, Eq. (20), when $G(\alpha, \omega)$ is even in ω, but is generally valid even when it is not. It follows easily from time-reversal invariance that for classical variables evenness follows, but this is not true for quantum mechanical variables. (Our definitions apply to the quantum case if α^* is replaced in the quantum case by the Hermitian conjugate, $\alpha^\dagger$.)

2.4 Noise Measurements Using Filters

An actual measurement of noise at a frequency ω_0 passes the signal $\alpha(t)$ through a filter described in the time domain by

$$\alpha_{\mathrm{out}}(t) = \int_{-\infty}^{t} K(t-t')\alpha(t')\,dt', \tag{37}$$

where $K(t)$ is known as the indicial response of the filter, or its response to

a $\delta(t)$ input pulse. In order that the filter be realizable, hence causal, output can only appear after input, so that

$$K(t) = 0 \text{ for } t < 0. \tag{38}$$

The upper limit in Eq. (37) can thus be extended to infinity. In terms of Fourier components,

$$\alpha(\omega) \equiv \int_{-\infty}^{\infty} \exp(-j\omega t)\alpha(t)\, dt, \tag{39}$$

Eq. (37) yields the convolution-theorem result

$$\alpha_{\text{out}}(\omega) = k(\omega, \omega_0)\alpha(\omega), \tag{40}$$

where

$$k(\omega, \omega_0) \equiv \int_0^{\infty} \exp(-j\omega t) K(t)\, dt. \tag{41}$$

The ω_0 is arbitrarily introduced into our notation to be a reminder that we shall be dealing with a filter that emphasizes the frequency region near ω_0. Thus we expect the output spectrum to be $|k(\omega, \omega_0)|^2$ times the input spectrum:

$$(\alpha_{\text{out}}^2)_\omega = |k(\omega, \omega_0)|^2 (\alpha^2)_\omega. \tag{42}$$

However, this argument is heuristic, since the integral for α_ω does not converge in the usual sense.

What is actually measured is

$$G_m(\omega_0) = \frac{1}{2T} \int_{t0-T}^{t0+T} |\alpha_{\text{out}}(t)|^2\, dt, \tag{43}$$

the time average of the squared signal. For long enough T we expect ergodicity and can replace the time average by the ensemble average. Because of stationarity, this result simplifies to

$$\begin{aligned} G_m(\omega_0) = \langle |\alpha_{\text{out}}(t)|^2 \rangle &= \int_{-\infty}^{t} K(t-t')^*\, dt' \\ &\times \int_{-\infty}^{t} K(t-t'')\, dt'' \langle \alpha(t')^* \alpha(t'') \rangle \\ &= \int_0^{\infty} \int_0^{\infty} K(u)^* K(v) \\ &\times \langle \alpha(t-u)^* \alpha(t-v) \rangle du dv \\ &= \int_0^{\infty} K(u)^* du \int_0^{\infty} K(v) dv \langle \alpha(v)^* \alpha(u) \rangle. \end{aligned} \tag{44}$$

Order has been preserved in the above steps so that they remain valid for noncommuting operators. Using the Wiener–Khinchin theorem in reverse, Eq. (33), to eliminate the autocorrelation, we obtain

$$G_m(\omega_0) = \frac{1}{4\pi} \int_{-\infty}^{\infty} G(\alpha, \omega) d\omega |k(\omega, \omega_0)|^2. \tag{45}$$

The factor 4π arises because of the convention followed in Eq. (36).

2.4.1 A Realizable Filter

The simplest example of a realizable filter is to regard $\alpha(t)$ as a voltage placed across an R-L-C circuit, with the output $\alpha_{\text{out}}(t)$ obtained across the resistance. The differential equations describing this filter are

$$RI(t) = \frac{1}{C}\frac{dQ}{dt} + L\frac{dI}{dt} = \alpha(t), \tag{46}$$

$$\alpha_{\text{out}}(t) = I(t)R. \tag{47}$$

These equations result in the Fourier relation

$$\alpha_{\text{out}}(\omega) = k(\omega, \omega_0)\alpha(\omega) \tag{48}$$

with

$$k(\omega, \omega_0) = \frac{R}{R + j(\omega L - 1/\omega C)}, \tag{49}$$

where $\omega_0 = 1/\sqrt{LC}$ is the resonance frequency.

The measured spectrum $G_m(\omega_0)$ continues to be given by Eq. (45). In the limit when the $Q \equiv \omega_0 L/R$ of the oscillator

becomes large, there are two sharp resonances at $\pm\omega_0$, and we can approximate

$$|k(\omega,\omega_0)|^2 = [\delta(\omega-\omega_0) + \delta(\omega+\omega_0)]\frac{\pi R}{2L}, \tag{50}$$

where the coefficient $\pi R/(2L)$ was chosen to yield the correct integral

$$\int_{-\infty}^{\infty} R^2 d\omega R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2 = \frac{\pi R}{L}. \tag{51}$$

This integral was evaluated using formula 031.10 in Gröbner and Hofreiter (1950). Equations (45) and (50) combine to yield

$$G_m(\omega_0) = \frac{RG(\alpha,\omega_0)}{4L}, \tag{52}$$

a result sensitive only to the measurement frequency with a known renormalization factor $R/(4L)$.

3 Thermal Noise

3.1 Johnson Noise

Johnson (1928) measured the voltage noise in a variety of materials, as a function of resistance. He found the results shown in Fig. 1, namely, that $\langle V^2\rangle$ is proportional to the resistance R, independent of the material. See also Kittel (1958). He also found that the measured noise power in the frequency interval df is proportional to the temperature of the resistor from which the noise emanates:

$$\langle v^2\rangle = 4k\,df\,RT = G(v,f)\,df. \tag{53}$$

To within his experimental accuracy, k is found to agree with the Boltzmann constant. The theoretical support for this choice is based on the equipartition

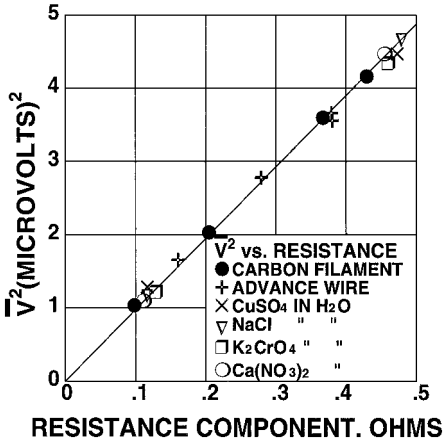


Fig. 1 The noise measured by Johnson (1928) versus resistance in six diverse materials

theorem and Nyquist's theorem discussed in Secs. 3.2 and 3.4.

Moullin (1938) gives a review of the experimental measurements of noise. He considers two resistors in parallel at two different temperatures and concludes that the noise in a frequency interval df is given by

$$\langle v^2\rangle = 4kdf\frac{R_1R_2}{(R_1+R_2)^2}[T_1R_2+T_2R_1] \equiv 4kdfR_eT_a, \tag{54}$$

where k is Boltzmann's constant and R_e is an effective resistance that is compared with experimental data in Fig. 2 by Williams (1937). Williams chooses T_a to be T_1, except in the limiting case of $R_1=\infty$, in which there is only one resistance, and he then takes T_a to be room temperature.

Moullin (1938, p. 31) generalizes this result to the case of an arbitrary number of impedances in parallel:

$$Z = \left[\sum_n \left(\frac{1}{Z_n}\right)\right]^{-1}, \tag{55}$$

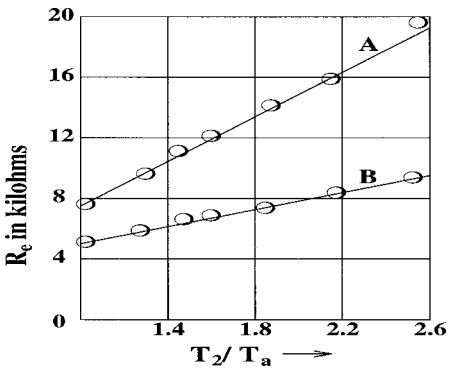


Fig. 2 Thermal noise for two resistors in parallel versus temperature obtained by Williams (1937) is plotted as an effective resistance defined by $R_e = \langle v^2 \rangle / [4kdfT_a]$ against T_2/T_a. Williams takes T_a to be T_1, except in the one-resistor case, for which T_1 is infinite, and he then chooses T_a to be room temperature. Both theory, in Eq. (54), and experiment are linear functions of the abscissa. Line B is the two-resistance case, and line A is the one-resistance case

$$\langle v^2 \rangle = 4kdf|Z|^2 \sum_n \frac{R_n T_n}{|Z_n|^2}. \qquad (56)$$

If we remember that admittance Y_n is related to the impedance $Z_n \equiv R_n + jX_n$ by

$$Y_n = \frac{1}{Z_n} = \frac{R_n - jX_n}{R_n^2 + X_n^2} \equiv g_n - js_n \qquad (57)$$

and write $V = ZI$, this result takes a simpler form in terms of current fluctuations,

$$\langle I^2 \rangle = 4kdf \sum_n g_n T_n. \qquad (58)$$

These alternative expressions are not surprising in terms of Thevenin's and Norton's theorems. The terms g_n and s_n are referred to as conductance and susceptance. The result, Eq. (54) or (58), used equilibrium theory to obtain results when true equilibrium (all temperatures equal) is absent.

3.2 Equipartition

In this section we establish a fundamental compatibility between Johnson noise and thermodynamic equilibrium by demonstrating that this noise source generates an energy $kT/2$ in inductances and capacitances to which it is connected.

Consider a series circuit of resistor R, inductance L, and capacitor C, with a Johnson noise voltage v in the resistor. Then the resulting current and charge fluctuations i and q are

$$i = \frac{v}{R + j(\omega L - 1/\omega C)}; \qquad q = \frac{i}{j\omega}. \qquad (59)$$

The fluctuation energy in the inductance is, from Eq. (53),

$$\begin{aligned} \frac{1}{2}L\langle i^2 \rangle &= \frac{1}{2}L \int \frac{G(v,f)\,df}{R^2 + (\omega L - 1/\omega C)^2} \\ &= 2kTRL \int_0^\infty \frac{df}{R^2 + (\omega L - 1/\omega C)^2} \\ &= \frac{1}{2}kT. \end{aligned} \qquad (60)$$

Similarly, the energy stored on the capacitance is

$$\begin{aligned} \frac{1}{2}\frac{1}{C}\langle q^2 \rangle &= \frac{2RkT}{C} \\ &\quad \times \int_0^\infty \frac{df}{\omega^2[R^2 + (\omega L - 1/\omega C)^2]} \\ &= \frac{1}{2}kT. \end{aligned} \qquad (61)$$

Thus, for both the inductance and the capacitor, the energy stored because of the noise in the resistor is precisely that expected by the equipartition theorem ($kT/2$ for each degree of freedom quadratic in the coordinate or the velocity.) Note that Eq. (59) is a relation between Fourier components, so that, for example, i and q

should be written i_ω and q_ω, whereas in Eqs. (60) and (61), we are really dealing with the time-dependent quantities $\langle i(t)^2\rangle$ and $\langle q(t)^2\rangle$, respectively.

A fundamental truth now emerges. Fluctuations must be associated with dissipation in order that $\langle q^2\rangle$ does not decay to zero, but maintains the appropriate thermal equilibrium energy.

3.3 Thermodynamic Derivation of Johnson Noise

In view of the compatibility with thermal equilibrium shown in the preceding section, it is not surprising that a simple thermodynamic argument can be used to demonstrate that the noise emanating from a resistor must be proportional to its resistance.

Consider two resistors in a series circuit shown in Fig. 3. The current i_1 through resistor 2 produced by the noise voltage v_1 in the first resistor is

$$i_1 = \frac{v_1}{(R_1 + R_2)}. \tag{62}$$

The power from v_1 into R_2 is given by

$$P_{2\leftarrow 1} = \left(\frac{v_1}{R_1 + R_2}\right)^2 R_2 = i_1^2 R_2. \tag{63}$$

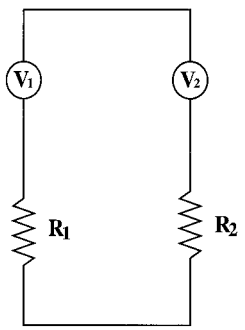


Fig. 3 The power transfer between two resistances in parallel, where v_j is the Johnson noise voltage in resistor j

Conversely, the power from 2 into 1 is given by

$$P_{1\leftarrow 2} = i_2^2 R_1 = \frac{v_2^2 R_1}{(R_1 + R_2)^2}. \tag{64}$$

If both resistors are at the same temperature, the second law of thermodynamics requires that there can be no steady net flow (in either direction). Equating Eqs. (63) and (64), we obtain

$$\frac{v_1^2}{R_1} = \frac{v_2^2}{R_2} \text{ or } \frac{G(v_1, f)}{R_1} = \frac{G(v_2, f)}{R_2} = W(f), \tag{65}$$

where $W(f)$ is a possibly frequency-dependent factor that is independent of the various resistances. In summary, the noise spectrum associated with an arbitrary resistance R is given by

$$G(v, f) = W(f)R. \tag{66}$$

If we connect a resistor R_1 to an impedance $R(f) + jX(f)$, the same equality requires

$$R(f)\left|\frac{v_1}{R_1 + R(f) + jX(f)}\right|^2 = \left|\frac{v_2}{R_1 + R(f) + jX(f)}\right|^2 R_1. \tag{67}$$

Thus we can conclude that the noise $G(v,f)$ associated with impedance $Z(f)$ is given by

$$G(v, f) = W(f)R(f) = W(f)\mathrm{Re}Z(f). \tag{68}$$

Thus the noise is proportional to $R(f) = \mathrm{Re}Z(f)$ even when the impedance $Z(f)$ is frequency dependent.

3.4 Nyquist's Theorem

A more complete derivation of the fluctuation–dissipation relation including a

determination of the universal coefficient $W(f)$ is provided by Nyquist (1927, 1928). Nyquist's procedure is to calculate the power dissipated in a load connected to the end of a transmission line in two different ways and compare the results. Equation (63) for the power from resistor R_1 into R_2 in the frequency interval df reduces with Eq. (66) to

$$P_{2\leftarrow 1} = \frac{W(f)R_1R_2}{(R_1+R_2)^2}\,df. \tag{69}$$

The maximum power is transferred when the impedance is matched, $R_1 = R_2$:

$$P_{2\leftarrow 1} = \frac{1}{4}W(f)\Delta f. \tag{70}$$

A transmission line can be terminated with its "characteristic impedance"

$$R_0 = \sqrt{\frac{L_1}{C}}, \tag{71}$$

where L_1 is the inductance per unit length of the line and C is its shunt capacitance per unit length. In this case, waves down the line are not reflected. The line acts as if it were infinite. Nyquist therefore chooses as his proof vehicle a transmission line terminated by R_0 at both ends. The line is assumed to have length L. The transmission line can be described in terms of harmonic-oscillator modes. If U is the energy density per mode, then the energy per mode is

$$UL = kT, \tag{72}$$

where we have made use of the equipartition theorem.

If the modes are described as plane waves $\exp(\pm ikx)$, then k takes the discrete values

$$k = \frac{2\pi n}{L}, \text{ whence } \Delta k = \frac{2\pi}{L} \tag{73}$$

is the mode spacing. With $\omega \equiv 2\pi f$ related to k by the group velocity, $v = \partial\omega/\partial k$, the number of modes propagating to the right in a given frequency interval is

$$\begin{aligned}\text{No. of modes} &= \frac{\text{modes}}{\delta k}\cdot\frac{\delta k}{\delta f}\Delta f \\ &= \frac{L}{2\pi}\frac{2\pi}{v}\Delta f = \frac{L}{v}\Delta f.\end{aligned} \tag{74}$$

Since each mode carries an energy U with the group velocity v, the power transmission down the line is

$$\begin{aligned}\text{flux} &= \left(\frac{\text{flux}}{\text{mode}}\right)(\text{No. of modes}) \\ &= (Uv)\left(\frac{L}{v}\right)\Delta f = UL\Delta f.\end{aligned} \tag{75}$$

Comparison with Eq. (70) yields

$$W(f) = 4UL. \tag{76}$$

In the limit of classical physics, Eq. (72) applies, and

$$W(f) = 4kT. \tag{77}$$

This result yields Nyquist's theorem

$$G(v,f)\Delta f = 4kTR\Delta f, \tag{78}$$

a result in agreement with Johnson's experimental results.

The beautifully simple Nyquist proof yields a result independent of frequency because all the harmonic-oscillator traveling modes have the same energy kT. The normalization must, of course, agree with that found in order to obtain the agreement with equipartition found in Sec. 3.2.

An apparent problem with Nyquist–Johnson noise is that the total voltage fluctuation

$$\langle v^2\rangle = \int_0^\infty G(v,f)df \tag{79}$$

diverges. Nyquist suggested that this problem could be removed if the classical energy, kT, associated with a harmonic oscillator were replaced by the quantum energy,

$$kT \to \hbar\omega\bar{n}(\omega) = \hbar\omega\left[\exp\left(\frac{\hbar\omega}{kT}\right) - 1\right]^{-1}, \tag{80}$$

which approaches kT at low frequencies and vanishes experimentally at high frequencies.

Of course, the actual energy associated with a harmonic oscillator,

$$E = \hbar\omega[\bar{n}(\omega) + \frac{1}{2}], \tag{81}$$

includes the zero-point energy. If the latter is retained, the divergence in the integrated energy reappears.

It is appropriate to argue, however, that the zero-point energy, while real, can never be absorbed and hence should be omitted in the Nyquist proof. This point, however, requires further discussion, as it relates to fundamental considerations of the effects of zero-point motion of the electromagnetic field in the area of quantum optics.

Callen and Welton (1951) considered a general class of systems (quantum mechanically) and established that Eq. (78), and its dual form

$$G(I, f)\Delta f = 4kTG\Delta f = 4kT\mathrm{Re}Y(\omega)\Delta f, \tag{82}$$

apply to all systems near equilibrium with the replacement of kT by Eq. (80) or (81) when necessary. The importance of the Callen–Welton work is the great generality of potential applications. The fact that all dissipative systems have corresponding noises associated with them is necessary in order that the second law of thermodynamics not be violated when such systems are connected.

3.5 Relation between Nyquist and Einstein

Consider a mechanical system with velocity v. Then the noise, Eq. (21), associated with v can be written

$$G(v, f) = \lim_{T\to\infty} \frac{2}{T}\left\langle\left|\int_{-T/2}^{T/2} v(t)e^{-j\omega t}dt\right|^2\right\rangle. \tag{83}$$

The zero-frequency noise is

$$G(v, 0) = \frac{2}{T}\langle[\Delta x]^2\rangle, \text{ where}$$
$$\Delta x = x\left(\frac{T}{2}\right) - x\left(-\frac{T}{2}\right). \tag{84}$$

But the usual diffusion constant, D, is defined by $\langle[\Delta x]^2\rangle = 2DT$, where T is the total time traveled. Thus the zero-frequency velocity noise is directly determined by the diffusion constant,

$$G(v, 0) = 4D. \tag{85}$$

Conversely, the fluctuation–dissipation theorem, Eq. (82), for the velocity (which is analogous to a current rather than a voltage, since the current is proportional to ev) is given by

$$G(v, f) = 4kT \text{ Re } Y(\omega), \tag{86}$$

where

$$Y(\omega) = \frac{v}{F} \tag{87}$$

is the admittance, or velocity per unit applied force. At zero frequency, we refer to v/F as the mechanical mobility B and v/E as the (electrical) mobility μ, so that

$$Y(0) = B = \frac{\mu}{e}. \tag{88}$$

The fluctuation–dissipation theorem at zero frequency now reads

$$D = kTB = \left(\frac{kT}{e}\right)\mu, \tag{89}$$

which is simply the Einstein (1905) relation between diffusion and mobility. For a verification of this relation for electrons and holes in semiconductors, see Transistor Teachers' Summer School (1953).

By relating the mechanical mobility to the viscosity determined by Stokes's law, Einstein was able to make a macroscopic determination of Avogadro's number in good agreement with current values. Although Einstein indicates that he had not seen the article by Brown (1828) describing microscopic observations on tiny pollen grains in water, in his second article Einstein (1906) refers to the work of Gouy (1888) with the quote that the "so-called Brownian motion is caused by the irregular thermal movements of the molecules of the liquid."

4 Shot Noise

4.1 The Poisson Process

There are two physical problems describable by the same random process. The first process is the radioactive decay of a collection of nuclei. The second is the production of photoelectrons by a *steady beam* of light on a photodetector. In both cases, we can let a discrete, positive, integer-valued variable $n(t)$ represent the number of counts emitted in the time interval between 0 and t. In both cases there is a constant probability per unit time ν such that νdt is the expected number of counts in $[t, t + dt]$ for small dt. We use the initial condition

$$n(0) = 0. \tag{90}$$

Then $n(t)$ will be the number of counts in the interval $[0,t]$. When we talk of $P(n,t)$ we can understand this to mean, $P(n, t|n = 0, 0)$, the conditional density distribution. Since the state $n(t) = n$ is supplied by transitions from the state $n - 1$ with production of photoelectrons at a rate νdt and is diminished by transitions from state n to $n + 1$, we have the rate equation

$$\frac{\partial P(n, t)}{\partial t} = \nu[P(n - 1, t) - P(n, t)] \quad \text{for } n > 0. \tag{91}$$

In the first term, n increases from $n - 1$ to n, and in the second, from n to $n + 1$. Since $n \geq 0$, we have no supply from the state $P(-1, t)$ so that

$$\frac{\partial P(0, t)}{\partial t} = -\nu P(0, t), \tag{92}$$

whose solution is

$$P(0, t) = P(0, 0)\exp(-\nu t) = \exp(-\nu t) \tag{93}$$

since $P(n, 0) = \delta_{n,0}$ at time $t = 0$, corresponding to the certainty that there are 0 particles at time $t = 0$.

The form, Eq. (93), of this solution suggests the transformation

$$P(n, t) = \exp(-\nu t)Q(n, t) \tag{94}$$

with the resultant equation

$$\frac{\partial Q(n, t)}{\partial t} = \nu Q(n - 1, t) \tag{95}$$

subject to the initial condition

$$P(n, 0) = Q(n, 0) = \delta_{n,0}. \tag{96}$$

Thus any $Q(n,t)$ may be readily obtained if $Q(n - 1)$ is known. But n, as described

by Eq. (91), can only increase. The result is the closed-form solution

$$P(n,t) = P(n,t|0,0) = \frac{(\nu t)^n}{n!}e^{-\nu t} \quad (97)$$

for $n \geq 0$ with a vanishing result for $n < 0$.

Distribution functions, such as Eq. (97), are also characterized by their moments μ'_j, and their central moments μ_j, defined by

$$\mu'_j = \langle n^j \rangle; \; \mu_j = \langle (n - \langle n \rangle)^j \rangle. \quad (98)$$

A third set of moments, introduced by statisticians because they indicate clearly the deviation of a random process from that of a Gaussian, are called Thiele (1903) semi-invariants or cumulants by Kendall and Stuart (1969). The cumulants or linked moments $\kappa_j \equiv \langle n^j \rangle_L$ are defined in terms of the characteristic function,

$$\begin{aligned}\langle \exp(ikn) \rangle &\equiv \sum_n \exp(ikn) P(n,t) \\ &= \exp\left[\sum_{j=1}^{\infty} \frac{(ik)^j}{j!}\kappa_j\right] \\ &= \exp\langle \exp(ikn) - 1 \rangle_L. \quad (99)\end{aligned}$$

By taking the logarithm of this equation, expanding in powers of k, and comparing coefficients, one can read off the cumulants. The first four are particularly simple:

$$\kappa_1 = \langle n \rangle; \; \kappa_2 = \langle n^2 \rangle_L = \langle (\Delta n)^2 \rangle; \quad (100)$$

$$\kappa_3 = \langle n^3 \rangle_L = \langle (\Delta n)^3 \rangle; \quad (101)$$

$$\langle n^4 \rangle_L = \langle (\Delta n)^4 \rangle - 3\langle (\Delta n)^2 \rangle^2; \quad (102)$$

where $\Delta n \equiv n - \langle n \rangle$ is the deviation from the mean. Cumulants beyond the second describe deviations from a Gaussian distribution.

The Poisson process is stationary. But no limit exists as $t \to \infty$, so that there is no time-independent $P(n)$. We can, however, evaluate the characteristic function of the conditional probability density:

$$\begin{aligned}\langle \exp(ikn) \rangle &\equiv \phi(k, t|n = 0, t = 0) \\ &= \sum_n e^{ikn} \frac{(\nu t)^n}{n!} e^{-\nu t} \\ &= \exp[\nu t(e^{ik} - 1)]. \quad (103)\end{aligned}$$

Comparison of Eq. (103) for the Poisson process with Eq. (99) shows that for this process all cumulants have the same value,

$$\kappa_j = \nu t. \quad (104)$$

4.2 Pure Shot Noise

The Poisson process just described is, in fact, the simplest example of shot noise. We can picture the actual number $n(t)$ as a staircase function that is flat except at a set of times, t_j, at which a jump of unity occurs. Mathematically, this is describable as the solution of the stochastic differential equation

$$\frac{dn}{dt} = \sum_j \delta(t - t_j) \equiv \hat{\nu}(t). \quad (105)$$

We shall first show that the noise source, $\hat{\nu}(t)$, in Eq. (105) is indeed white noise and demonstrate its relation to conventional shot noise.

The average pulse rate over the large time interval $[0,T]$ is

$$\begin{aligned}\nu(t) = \langle \hat{\nu}(t) \rangle &= \frac{1}{T}\sum_j \int_0^T \delta(t - t_j) dt_j \\ &= \frac{N}{T} = \nu, \quad (106)\end{aligned}$$

where N is the number of pulses in the interval $0 < t_j \leq T$. We call the ratio ν to conform with the Poisson process in

Sec. 3.1 and assume that this ratio is independent of t, and of the location of the time interval of length T.

The correlation in the fluctuation $\alpha(t) = \hat{\nu}(t) - \nu$ can be calculated in a similar way:

$$\langle \alpha(t)\alpha(t')\rangle = \langle \hat{\nu}(t)\hat{\nu}(t')\rangle - \langle \hat{\nu}(t)\rangle\langle \hat{\nu}(t')\rangle. \tag{107}$$

The first term involves a double sum

$$\langle \hat{\nu}(t)\hat{\nu}(t')\rangle = \sum_{i,j}\langle \delta(t-t_i)\delta(t'-t_j)\rangle, \tag{108}$$

but the diagonal ($j = i$) term involves the simpler average

$$\frac{1}{T}\sum_i \int_0^T \delta(t-t_i)\delta(t'-t_i)dt_i = \frac{N}{T}\delta(t-t'). \tag{109}$$

The terms with $j \neq i$ involve an uncorrelated average, which cancels against the last term in Eq. (107) in the limit as T goes to infinity, with the result

$$\langle \Delta\hat{\nu}(t)\Delta\hat{\nu}(t')\rangle = \nu\delta(t-t'). \tag{110}$$

The noise, by the Wiener–Khinchin formula, is

$$G(\nu,\omega) = 2\int_{-\infty}^{\infty}\langle \Delta\hat{\nu}(t)\Delta\hat{\nu}(t')\rangle dt = 2\nu, \tag{111}$$

a constant, or white noise.

If the pulses carry a charge e, the current is

$$\hat{I}(t) = e\hat{\nu}(t), \tag{112}$$

and we arrive at the traditional shot-noise formula

$$G(I,\omega) = 2e^2\nu = 2eI, \tag{113}$$

where $I = \langle \hat{I}\rangle$ is the average current.

4.3 Generalized Characteristic Functions

We must establish the equivalence of this pulse description in Eq. (105) with the Poisson distribution characterized by Eq. (91), by calculating the characteristic function by a direct method, and demonstrate its equality with that of the Poisson process given in Eq. (103). The characteristic function is defined by

$$\phi(k,t) = \langle \exp[ikn(t)]\rangle = \left\langle \exp\left[ik\int_0^t \hat{\nu}(s)\,ds\right]\right\rangle. \tag{114}$$

To save duplication later, we consider an appreciably more general problem. In this problem, the sequence of pulses that arrive can have a localized shape $f(t-t_j)$ that need not be a delta function. In the original problem considered by Campbell (1909), the discontinuity involved the charge on the electron, and so a factor q was added. The model was further generalized by Rice (1944, 1945, 1948) to permit the jumps to contain a random factor η. Thus Rice considered the process

$$\Theta(t) = q\sum \eta_j f(t-t_j), \tag{115}$$

where the η_j's are random jumps with a distribution independent of t_j, of t, and of j.

Our procedure for dealing with the same problem consists in writing

$$\Theta(t) = \int f(t-s)S(s)\,ds, \tag{116}$$

where the shot-noise function $S(s)$ is now given by

$$S(s) = q\sum_j \eta_j\delta(s-t_j). \tag{117}$$

Equation (116) describes $\Theta(t)$ as filtered shot noise. We can always remove the

filtering to obtain results appropriate to pure shot noise.

We can now relate the ordinary characteristic function of Θ to the generalized characterized function of the shot-noise function, $S(s)$,

$$\langle \exp ik\Theta(t)\rangle = \left\langle \exp i \int_{-\infty}^{\infty} \gamma(s) S(s) ds \right\rangle \tag{118}$$

by setting

$$\gamma(s) \equiv kqf(t-s). \tag{119}$$

The average in Eq. (118), for general $\gamma(s)$, was evaluated in two ways in Lax (1966b). The first made explicit use of Langevin techniques. The second, which follows Rice, is presented here. It makes use of the fact that the average can be factored:

$$\left\langle \exp i \int_{-\infty}^{\infty} \gamma(s) S(s) ds \right\rangle = \prod_{j=1}^{N} \langle \exp[i\eta_j \gamma(t_j)]\rangle. \tag{120}$$

Here, we have supposed that N pulses are distributed uniformly over a time interval T at the rate $\nu = N/T$. All N factors are independent of each other and have equal averages, so that after adding and subtracting 1 inside the brackets, the result is

$$\begin{aligned}
&\prod_{j=1}^{N} \int g(\eta_j) d\eta_j \left[1 + \frac{1}{T} \int_{-T/2}^{T/2} \{\exp[i\eta_j \times \gamma(t_j)] - 1\}\, dt_j \right] \\
&= \left[1 + \frac{1}{T} \int g(\eta) d\eta \int_{-T/2}^{T/2} \{\exp[i\eta \times \gamma(s)] - 1\}\, ds \right]^N \\
&= \exp\left[\nu \int g(\eta) d\eta \int_{-\infty}^{\infty} \{\exp[i\eta \times \gamma(s)] - 1\}\, ds \right],
\end{aligned} \tag{121}$$

where $g(\eta)$ is the normalized probability density for the random variable η. In the last step, we assumed that the integral over s converges and replaced it by its limit before taking a final limit in which N and T approach infinity simultaneously with the fixed ratio $N/T = \nu$.

4.4 Rice's Generalized Campbell Theorem

Campbell's two theorems are the first and second moments, $\langle\Theta\rangle$ and $\langle(\Delta\Theta)^2\rangle$, evaluated explicitly in Eqs. (126) and (127) below. Rice's generalized Campbell's theorem is obtained by setting $\gamma(s) = kqf(t-s)$ in Eq. (118). The result is not just the first two moments, but the complete characteristic function

$$\begin{aligned}
\langle \exp ik\Theta(t)\rangle &= \left\langle \exp\left[ikq \sum_j \eta_j f(t-t_j) \right] \right\rangle \\
&= \exp\left(\nu \int g(\eta) d\eta \int_{-\infty}^{\infty} \{\exp[ikq\eta \times f(t-s)] - 1\}\, ds \right).
\end{aligned} \tag{122}$$

Let us now specialize to the case of the original Campbell process by setting $g(\eta) = \delta(\eta - 1)$:

$$\langle \exp ik\Theta(t)\rangle = \exp\left[\nu \int_{-\infty}^{\infty} \{\exp[ikqf(t-s)] - 1\}\, ds \right]. \tag{123}$$

The probability density of the variable Θ of Eq. (115) may then be obtained by taking the inverse Fourier transform of

the characteristic function in Eq. (122):

$$P(\Theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-ik\Theta)dk \times \exp\left[\nu \int g(\eta)d\eta \int_{-\infty}^{\infty} \{\exp[ikq\eta \times f(s)] - 1\}\, ds\right]. \quad (124)$$

This form of generalized Campbell's theorem, like its antecedents, assumes that the t_j are randomly (and on the average uniformly) distributed in time. Moreover, there is assumed to be no correlation between successive pulse times. Lax and Phillips (1958) have found it convenient to exploit Eq. (124) in studying one-dimensional impurity bands.

The cumulants of the Rice process can be obtained from Eq. (122) as the coefficients of $k^r/r!$ in the exponent:

$$\kappa_r = \left\langle \left[q \sum_j \eta_j f(t - t_j) \right]^r \right\rangle = \nu q^r \int g(\eta)\eta^r d\eta \int_{-\infty}^{\infty} f(s)^r ds = \nu q^r \langle \eta^r \rangle \int_{-\infty}^{\infty} f(s)^r\, ds. \quad (125)$$

4.5 Campbell's Theorems

The choice $g(\eta) = \delta(\eta - 1)$ reduces the Rice process to the original Campbell process. Campbell's theorem itself includes only the cases $r = 1, 2$. These take the form

$$\langle \Theta \rangle = \nu q \int_{-\infty}^{\infty} f(s)\, ds, \quad (126)$$

$$\langle [\Delta\Theta]^2 \rangle = \nu q^2 \int_{-\infty}^{\infty} f(s)^2\, ds. \quad (127)$$

We close this section with an application of Campbell's theorem to the RC circuit shown in Fig. 4. A charge e from a vacuum tube generates a voltage pulse

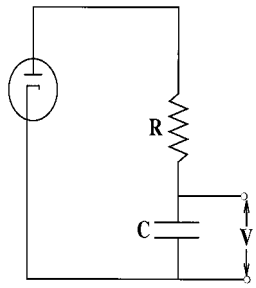


Fig. 4 The voltage fluctuation induced by shot noise into an RC circuit

$$ef(t) = \left(\frac{e}{C}\right) \exp\left(\frac{-t}{RC}\right), \quad (128)$$

across the resistor R. Campbell's theorems then yield

$$\langle V \rangle = \nu \left(\frac{e}{C}\right) \int_0^{\infty} e^{-t/RC} dt = \nu e R = IR \quad (129)$$

and

$$\langle (\Delta V)^2 \rangle = \nu \frac{e^2}{C^2} \int_0^{\infty} e^{-2t/(RC)} dt = \frac{IeR}{2C}. \quad (130)$$

The electronic charge can be determined by comparison of $\langle (\Delta V)^2 \rangle$ with $\langle V \rangle$:

$$e = \frac{2C\langle (\Delta V)^2 \rangle}{\langle V \rangle}. \quad (131)$$

4.6 Equivalence of Shot Noise to the Poisson Process

If we further specialize Eq. (123) by setting

$$qf(t - s) = 1 \quad \text{for } 0 \le s \le t, \quad (132)$$

and zero elsewhere, then Eq. (116) reduces to

$$\Theta(t) = \int_0^t \hat{\nu}(s) ds = n(t). \quad (133)$$

The characteristic function, Eq. (132), of Θ is then identical to that of n and is given by

$$\langle \exp ikn(t) \rangle = \exp\left[\nu \int_0^t [\exp(ik) - 1]\, ds \right], \tag{134}$$

which reduces immediately to Eq. (103) as desired. This completes our proof that the shot-noise process of Eq. (105) is equivalent to the Poisson process of Eq. (91).

4.7 Transit-Time Effects

A set of carriers, j, leave a source (e.g., a cathode) at a set of times t_j and move to a destination (e.g., a plate) with a velocity $v(t - t_j)$. All carriers travel the fixed distance L and take the same travel time T, such that

$$\int_0^T v(t)dt = L. \tag{135}$$

The current in the external circuit will be

$$I(t) = \frac{e}{L} \sum_j v(t - t_j). \tag{136}$$

This intuitive result can be justified by a detailed calculation using Maxwell's equations.

The current can then be written as a convolution

$$I(t) = \frac{1}{L} \int_{-\infty}^{\infty} v(t - s) S(s)\, ds, \tag{137}$$

where the shot-noise function is

$$S(s) = e\hat{v}(s) = e \sum_j \delta(s - t_j) \tag{138}$$

as in Eq. (112) or (117) with $\eta = 1$. This convolution translates into a product of Fourier transforms that are squared to get the desired noise result

$$G(I, \omega) = G(S, \omega) W(\omega) = 2eIW(\omega), \tag{139}$$

where we have used the pure shot-noise result, Eq. (113), and the window factor

$$W(\omega) = \left| \frac{1}{L} \int_0^T v(t) e^{(-j\omega t)} dt \right|^2 \tag{140}$$

accounts for the transit-time effects and reduces the noise from the maximum value of $W(0) = 1$ eventually to $W(\infty) = 0$, although not necessarily monotonically. The time dependence of the velocity will be determined by the nature of the forces acting on the electron.

4.8 Generation–Recombination Process

The generation–recombination process in semiconductors is a second example of a shot-noise process in which the diffusion constants can be calculated from first principles on the basis of an understanding of the physics of the process. It is also an example of what statisticians refer to as a birth and death process.

Let us define the random integer variable n as the occupancy of some state. We assume that the particles are generated at the rate $G(n)$ and disappear at the rate $R(n)$ (recombination). Then our mean equation of motion is

$$\frac{\partial \langle n \rangle}{\partial t} = \langle G(n) \rangle - \langle R(n) \rangle \equiv \langle A(n) \rangle, \tag{141}$$

where $\langle A(n) \rangle$ is our drift vector. Since the occupancy of state n is increased by generation from the state $n - 1$ and reduced by generation out of the state n, whereas it is increased by recombination out of state $n + 1$ and reduced by recombination out

of state n, the probability distribution function $P(n,t)$ obeys the following equation:

$$\frac{\partial P(n,t)}{\partial t} = G(n-1)P(n-1,t) - G(n)P(n,t) + R(n+1) \times P(n+1,t) - R(n)P(n,t). \quad (142)$$

The Poisson process is a special case of the present one with $R = 0$, $G = \nu$. If the derivative is written as a difference, and the symbol n is replaced by n', Eq. (142) can be written as a master equation

$$\frac{P(n', t+\Delta t) - P(n,t)}{\Delta t} = \sum_n \frac{P(n', t+\Delta t|n,t)}{\Delta t P(n,t)} \quad (143)$$

with the transition rate

$$\frac{P(n', t+\Delta t|n,t)}{\Delta t} = \sum_n \{G(n)\delta_{n,n'-1} + R(n)\delta_{n,n'+1} - [G(n) + R(n)]\delta_{n,n'}\}. \quad (144)$$

The rth diffusion coefficient D_r, defined by (see Secs. 7.1 and 7.2)

$$r!D_r = \frac{\langle (n'-n)^r \rangle}{\Delta t} = \sum_{n'} \frac{(n'-n)^r P(n', t+\Delta t|n,t)}{\Delta t} \quad (145)$$

takes the value

$$r!D_r = \sum_{n'} (n'-n)^r [G(n)\delta_{n,n'-1} + R(n)\delta_{n,n'+1}] = G(n) + (-1)^r R(n) \quad (146)$$

after omission of noncontributing terms from Eq. (144).

A simpler procedure is to use Taylor's theorem in the operator form

$$\exp\left(h\frac{\partial}{\partial n}\right) F(n) = F(n+h), \quad (147)$$

with h set to $+1$ or -1, to rewrite Eq. (142) in the operational form

$$\frac{\partial P}{\partial t} = \left[\exp\left(-\frac{\partial}{\partial n}\right) - 1\right] G(n)P(n) + \left[\exp\left(\frac{\partial}{\partial n}\right) - 1\right] R(n)P(n).$$

Expanding the Taylor series leads immediately to

$$\frac{\partial P}{\partial t} = \sum_{r=1}^{\infty} \left(-\frac{\partial}{\partial n}\right)^r \times \frac{[G(n) + (-1)^r R(n)P(n)]}{r!} P(n). \quad (148)$$

The rth-order diffusion constant derivative term read off from the coefficient of the rth derivative, in accord with Eq. (273), agrees with Eq. (146). [See Sec. 7.2 and Lax (1966a) for a detailed discussion of the generalized Fokker–Planck equation.] Thus all the even-numbered diffusion constants are proportional to the sum of the rate in plus the rate out, while all the odd-numbered diffusion constants are proportional to the difference, i.e., the rate in minus the rate out. In particular, the first and second moments, D_1 and D_2, obey

$$D_1 = A(a) = G - R, \quad (149)$$

$$2!D_2 = 2D = G + R = \text{rate in} + \text{rate out}. \quad (150)$$

These results are characteristic of shot noise. Thus, in the quasilinear approximation, the operating point n_0 and decay parameter $\boldsymbol{\Lambda}$ are determined by Eqs. (10)

and (11):

$$A(n_0) = G(n_0) - R(n_0) = 0, \tag{151}$$

$$\Lambda \equiv -\frac{\partial A}{\partial n}\bigg|_{n=n_0} = R'(n_0) - G'(n_0), \tag{152}$$

and the diffusion coefficient is

$$D = \frac{1}{2}[G(n_0) + R(n_0)] = G(n_0). \tag{153}$$

Then the Einstein relation, Eq. (15), in the case of a single variable yields the mean square fluctuation from the average value in terms of the diffusion constant D and the decay constant Λ:

$$\langle(\Delta n)^2\rangle = \frac{D}{\Lambda} = \frac{G(n_0)}{R'(n_0) - G'(n_0)}. \tag{154}$$

We now want to calculate the autocorrelation function, $\langle\Delta n(t)\Delta n(0)\rangle$. Since the transport equation and its solution are

$$\frac{\partial\langle\Delta n(t)\rangle}{\partial t} = \langle A\rangle \approx -\Lambda\langle\Delta n\rangle,$$

$$\langle\Delta n(t)\rangle = e^{-\Lambda t}\Delta n(0), \tag{155}$$

the regression theorem and stationarity determine the correlations, via Eq. (18), to be

$$\langle\Delta n(t)\Delta n(0)\rangle = e^{-\Lambda|t|}\langle[\Delta n(0)]^2\rangle. \tag{156}$$

The noise in n is then given by Eq. (26),

$$G(n,\omega) = \frac{4\Lambda}{\Lambda^2 + \omega^2}\langle(\Delta n)^2\rangle. \tag{157}$$

5 Resistance Modulation Fluctuations

5.1 Conductivity Fluctuations

Information about carrier-number fluctuations can be obtained by injecting a constant current and measuring the voltage fluctuations induced by the conductivity modulation caused by carrier-concentration fluctuations. The admittance can be written

$$Y = A(pe\mu_p + ne\mu_n) = \frac{(Pe\mu_p + Ne\mu_n)}{L}, \tag{158}$$

where μ_p and μ_n are the hole and electron mobilities, p and n are the hole and electron concentrations, and P and N are the total hole and electron numbers over the volume AL between the electrodes, of area A and separation L. Thus the fractional voltage fluctuations are given by

$$\frac{\Delta V}{V} = -\frac{\Delta Y}{Y} = -\frac{\Delta P\mu_p + \Delta N\mu_n}{P\mu_p + N\mu_n}. \tag{159}$$

If only electrons and holes are present (and not traps), charge neutrality will be enforced up to the (very high) dielectric relaxation frequency, so that to an excellent approximation

$$\Delta N(t) = \Delta P(t). \tag{160}$$

Thus the voltage autocorrelation is given by

$$\langle\Delta V(t)\Delta V(0)\rangle = V^2\left[\frac{(1+b)}{(P+Nb)}\right]^2 \times \langle\Delta P(t)\Delta P(0)\rangle, \tag{161}$$

where $b = \mu_n/\mu_p$. It follows that the voltage noise is given by

$$G(V,\omega) = V^2\left(\frac{1+b}{P+Nb}\right)^2\langle(\Delta P)^2\rangle \times \int_0^\infty 4\cos\omega t dt\Phi(t), \tag{162}$$

and the total voltage fluctuation may be obtained by replacing the integral by unity.

The "after-effect function" $\Phi(t)$ is defined by

$$\langle \Delta P(t) \Delta P(0) \rangle = \langle (\Delta P)^2 \rangle \Phi(t). \quad (163)$$

The total noise, which only involves $\Phi(0) = 1$, is consistent with the normalization condition, Eq. (27), in the noise spectrum. The after-effect function will be calculated in later sections.

5.2 A Thermodynamic Treatment of Total Carrier Fluctuations

We first consider a set of free electrons in the nondegenerate case when Boltzmann (rather than Fermi) statistics are applicable. Then Shockley (1950) has shown that the total number of electrons obeys the statistical mechanical relation

$$N = N_c \exp\left[\frac{(\mu - E_c)}{kT}\right], \quad (164)$$

where E_c is the energy at the bottom of the conduction band and N_c is a temperature-dependent effective density of states. The conventional symbol μ is used to represent the electron Fermi level or chemical potential. Since it will not appear later, there will be no confusion with the mobility. The thermodynamic formula, Eq. (3), then yields

$$\langle (\Delta N)^2 \rangle = kT \left(\frac{\partial N}{\partial \mu}\right)_T = N. \quad (165)$$

This result is not entirely surprising, since the total number of carriers is an integral

$$N = \int_V \sum_j \delta(\mathbf{r} - \mathbf{r}_j) d^3 r; \quad (166)$$

this is the three-dimensional analog of Eqs. (105) and (133). Indeed, Eq. (165) can be derived directly from Eq. (166) using only the assumption that the $\mathbf{r}_j$ are uniformly distributed in space.

A less obvious case is that of a set of N_t traps interacting with a reservoir of chemical potential μ_t. We assume that the trap occupancy is sufficiently high that Fermi statistics are necessary. In that case, the number of filled traps by use of Fermi–Dirac statistics is

$$\hat{N} = \frac{N_t}{1 + \exp[(E_t - \mu_t)/kT]}. \quad (167)$$

In that case, the thermodynamic formula, Eq. (165), becomes

$$\langle (\Delta \hat{N})^2 \rangle = kT \left(\frac{\partial \hat{N}}{\partial \mu_t}\right)_T = \hat{N}\left[1 - \frac{\hat{N}}{N_t}\right]. \quad (168)$$

It can be seen that the fluctuations are reduced by a factor equal to the fraction of empty states. The reason for this result is made clear in the next section, in which a kinetic approach is used for the same problem. If both N and $\hat{N}$ are allowed to vary simultaneously, the simplest distribution consistent with these second moments is

$$W \propto \exp\left[-\frac{(\Delta N)^2}{2\langle (\Delta N)^2 \rangle} - \frac{(\Delta \hat{N})^2}{2\langle (\Delta \hat{N})^2 \rangle}\right]. \quad (169)$$

The term in $\Delta N \Delta \hat{N}$ vanishes because N does not depend on μ_t and $\hat{N}$ does not depend on μ. Within the quasilinear approximation, it is appropriate to ignore higher cumulants than the second and stop at the Gaussian approximation.

Suppose, now, that the electrons in traps do not have an independent reservoir, but are obtained from the free carriers. Then we must impose the conservation condition

$$\Delta \hat{N} = -\Delta N. \quad (170)$$

If this constraint is inserted into Eq. (169), we obtain a Gaussian in a single variable with the second moment

$$\langle(\Delta N)^2\rangle = \langle(\Delta\hat{N})^2\rangle = \frac{\langle(\Delta N)^2\rangle\langle(\Delta\hat{N})^2\rangle}{\langle(\Delta N)^2\rangle + \langle(\Delta\hat{N})^2\rangle} = \frac{N\hat{N}(N_t - \hat{N})}{NN_t + \hat{N}(N_t - \hat{N})}. \quad (171)$$

In the next section, we show that this result can be derived by the Einstein relation.

The situation for holes is similar to that for electrons. If the holes have their own reservoir, then the typical Poisson process prevails:

$$\langle(\Delta P)^2\rangle = P. \quad (172)$$

If holes, traps, and electrons are all present and coupled to each other, then charge neutrality imposes the constraint

$$\Delta N + \Delta\hat{N} = \Delta P. \quad (173)$$

In the presence of compensating centers, N_{co}, there is also a neutrality condition for the steady state:

$$N + \hat{N} = P + N_t - N_{co}. \quad (174)$$

The influence of these constraints on the fluctuations is developed in the next section.

5.3 Einstein Derivation of Carrier Fluctuations with Traps

As in the previous section, we consider traps interacting with free electrons. This problem is a generation–recombination problem with a generation rate proportional to the number of carriers in the traps and a recombination rate proportional to the product of the number of free carriers and the number of empty traps,

$$g(N) = \gamma\hat{N}, \quad r(N) = \rho N(N_t - \hat{N}). \quad (175)$$

We regard g and r as functions only of N since $\hat{N}$ is a function of N given in Eq. (174). The diffusion coefficient in the steady state is then

$$D(N) = r(N) \quad (176)$$

in view of Eqs. (151) and (153); and the decay coefficient, in accord with Eq. (152), is

$$\Lambda = r'(N) - g'(N) = \rho[(N_t - \hat{N}) + N] - (-\gamma), \quad (177)$$

where the derivatives were taken using $\Delta\hat{N} = -\Delta N$. The electron carrier fluctuation $\langle(\Delta N)^2\rangle = D/\Lambda$ is in agreement with Eq. (171).

We next consider carrier fluctuations in a p-type material in the presence of electrons and traps. We will regard N and $\hat{N}$ as the determining variables, with P obtained from the neutrality constraint, Eq. (174). Our discussion will follow Lax and Mengert (1960) with the replacement of C/Ω in that article by ρ, and of g in that article by γ, to conform with our notation in Eq. (175). Our nonlinear transport equation is taken to be

$$\frac{d\langle N\rangle}{dt} = \langle G - R + \gamma\hat{N} - \rho(N_t - \hat{N})N\rangle,$$
$$\frac{d\langle\hat{N}\rangle}{dt} = \langle\rho(N_t - \hat{N})N - \gamma\hat{N}\rangle. \quad (178)$$

The terms in G and R refer to generation and recombination of electron–hole pairs, and the other terms refer to generation and recombination from the traps, as in Eq. (175).

The underlying transition rate per unit time for this model is

$$\begin{aligned} w(N', \hat{N}'; N, \hat{N}) &= [G\delta(N', N+1) \\ &+ R\delta(N', N-1)]\delta(\hat{N}', \hat{N}) \\ &+ \rho\delta(N', N+1)\delta(\hat{N}', \hat{N}-1) \\ &+ \gamma(N_t - N)N\delta(N', N-1) \\ &\times \delta(\hat{N}', \hat{N}+1). \end{aligned} \tag{179}$$

The first moments of the transition rate are obtained by summing over the primed variables:

$$\begin{aligned} &\sum (N' - N) w(N', \hat{N}'; N, \hat{N}) = \\ &\quad G - R + \rho\hat{N} - \nu N, \\ &\sum (\hat{N}', -\hat{N}) w(N', \hat{N}'; N, \hat{N}) = \\ &\quad \nu N - \rho\hat{N}, \end{aligned} \tag{180}$$

where we use the abbreviation

$$\nu = \gamma(N_t - \hat{N}). \tag{181}$$

The results agree with our phenomenological transport equations, Eq. (178). If we linearize these equations, we obtain the two-by-two system

$$\frac{d}{dt}\begin{bmatrix} \langle \Delta N \rangle \\ \langle \Delta N \rangle \end{bmatrix} = -\mathbf{\Lambda}\begin{bmatrix} \langle \Delta N \rangle \\ \langle \Delta N \rangle \end{bmatrix} \tag{182}$$

with

$$\mathbf{\Lambda} = \begin{bmatrix} \nu + r & -g' \\ -\nu & g' \end{bmatrix}, \tag{183}$$

where the parameters are now evaluated under the steady-state conditions

$$\gamma\hat{N} = \rho(N_t - \hat{N})N = \nu N, \quad R = G. \tag{184}$$

The symbol r now stands for the experimental electron–hole recombination rate

$$r = \frac{\partial(R - G)}{\partial N}, \tag{185}$$

and

$$g' = \gamma + \rho N. \tag{186}$$

The diffusion constants may be obtained from the second moments of the transition rate:

$$\begin{aligned} D_{11} &= \sum \frac{(N' - N)^2 w(N', \hat{N}'; N, \hat{N})}{2} \\ D_{22} &= \sum \frac{(\hat{N}' - \hat{N})^2 w(N', \hat{N}'; N, \hat{N})}{2} \\ D_{12} &= \sum \frac{(N' - N)(\hat{N}' - \hat{N})\, w(N', \hat{N}'; N, \hat{N})}{2}, \end{aligned} \tag{187}$$

so that

$$D_{11} = \frac{[(G + R) + (\gamma\hat{N} + \nu N)]}{2}, \tag{188}$$

where the evaluation of D_{11} was done at the steady state using the conditions

$$G = R, \quad \nu N = \gamma\hat{N}. \tag{189}$$

The complete matrix at the steady state is thus found to be

$$\mathbf{D} = \begin{vmatrix} R + \gamma\hat{N} & -\gamma\hat{N} \\ -\gamma\hat{N} & \gamma\hat{N} \end{bmatrix}. \tag{190}$$

The population fluctuations that result from Eq. (15) are then

$$\mathbf{\Lambda}^{-1}\mathbf{D} = \begin{bmatrix} R/r & 0 \\ (\nu R/r - \gamma\hat{N})/g' & \hat{N}\gamma/g' \end{bmatrix}. \tag{191}$$

There is a coupling between the free and bound electron populations. However, if the hole (majority) population P is much larger than N or $\hat{N}$, it would provide uncoupled reservoirs for these two populations, and we would expect the 21 element of the matrix to disappear. If we write the recombination as proportional to the product of the number of holes and electrons,

$$R = \alpha(N + \Delta N)(P + \Delta P), \quad G = NP, \tag{192}$$

then

$$R - G = \alpha[(P + N)\Delta N + N\Delta\hat{N}] \quad (193)$$

after use of charge neutrality, Eq. (173). Thus Eq. (185) yields

$$r = \alpha(P + N) \approx \alpha P = \frac{R}{N}. \quad (194)$$

In view of Eq. (189), the 21 elements vanish. With the help of Eq. (181), we find

$$\frac{\gamma}{g'} = 1 - \left(\frac{\hat{N}}{N_t}\right), \quad (195)$$

so that

$$\mathbf{\Lambda}^{-1}\mathbf{D} = \begin{bmatrix} N & 0 \\ 0 & \hat{N}' \end{bmatrix}, \quad (196)$$

where in this connection we have introduced the shorthand notation

$$\hat{N}' = \hat{N}\left[1 - \left(\frac{\hat{N}}{N_t}\right)\right]. \quad (197)$$

5.4 The Spectrum of Resistance Modulation Fluctuations

We return to the problem of the last section with a large number of majority carriers, P, a smaller number of trapped carriers, $\hat{N}$, and a still smaller number of free electrons, N. Because of the neutrality condition, Eq. (173), the voltage fluctuation of Eq. (159) can be written

$$\Delta V = C_1\Delta N + C_2\Delta\hat{N}, \quad (198)$$

where

$$C_2 = \frac{V}{(P + bN)}, \quad C_1 = (1 + b)C_2 \quad (199)$$

and $b = \mu_n/\mu_p$.

In view of the total population fluctuations given in Eqs. (196) and (197), the total voltage noise is given by

$$\int_0^\infty G(V, f)df = \langle(\Delta V)^2\rangle,$$
$$= |C_2|^2[\hat{N}' + (1 + b)^2 N]. \quad (200)$$

The spectrum of voltage fluctuations is given by Eqs. (29) and (31):

$$G(V, f) = 4C_2^2\mathrm{Re}\{(i\omega 1 + \mathbf{\Lambda})_{22}^{-1}\hat{N}' + (1 + b) \times [(i\omega 1 + \mathbf{\Lambda})_{12}^{-1}\hat{N}' + (i\omega 1 + \mathbf{\Lambda})_{21}^{-1}N] + (1 + b)^2(i\omega 1 + \mathbf{\Lambda})_{11}^{-1}N\}. \quad (201)$$

Here we have omitted the star, since $\mathbf{\Lambda}$ is real, and replaced $-j$ by i to avoid minus signs. The elements of $\mathbf{\Lambda}$ were given in Eq. (183). An exact evaluation by Lax and Mengert (1960) yields a ratio of a quadratic to a quartic in ω. Lax and Mengert re-expressed their results by a partial-fraction analysis into the more understandable form

$$G(V, f) = \frac{4V^2}{(P + Nb)^2} \times \left[\frac{A_S\lambda_S}{\omega^2 + \lambda_S^2} + \frac{A_F\lambda_F}{\omega^2 + \lambda_F^2}\right] \quad (202)$$

of a sum of two Lorentzians whose integrated contributions are A_S and A_F associated with the slow and fast modes with eigenvalues λ_S and λ_F. The trace and determinant of $\mathbf{\Lambda}$, Eq. (183), yield the sum and product relations

$$\lambda_F = \nu + r + g' - \lambda_S, \quad \lambda_S = \frac{rg'}{\lambda_F}. \quad (203)$$

Typical parameters from the experiments of Hornbeck and Haynes (1955) and Haynes and Hornbeck (1955), as revised by Lax and Mengert (1960), indicate that the trapping rate $\nu = 3 \times 10^7$/s for free electrons greatly exceeds the effective release rate $g' = 3$/s from traps. These would be the basic decay rates for N and

$\hat{N}$ respectively. It is then appropriate to assume that the slow rate $\lambda_S \ll \lambda_F$. If we iterate starting with $\lambda_S = 0$, we obtain a first approximation

$$\lambda_F \approx \nu + r + g', \quad \lambda_S \approx \frac{g'r}{(\nu + r + g')} \tag{204}$$

Since $\lambda_F \approx \nu = 3 \times 10^7$, and $\lambda_S \approx g'r/\nu \approx 3 \times 10^{-3}$, Eq. (204) has a relative accuracy of $\lambda_S/\lambda_F \approx 10^{-10}$. Lax and Mengert (1960) use this large decay ratio to simplify the formulas for A_S and A_F to

$$A_S \approx \hat{N}' + (1 + b)N, \quad A_F \approx (b^2 + b)N. \tag{205}$$

Perhaps the surprise in these numbers is that the slow eigenvalue λ_S of the combined system is 1000 times smaller than the emission rate from traps, which represents the decay of $\hat{N}$. The reason this is so is that the electrons rapidly ("adiabatically") respond to the bound trap density and generate captures that almost cancel the release rate. To verify this, we note that the adiabatic procedure in lowest order, as described in Lax (1967), is to set the time derivative of the rapidly changing quantity, $d\langle \Delta N \rangle / dt = 0$. This yields

$$\Delta N = \left[\frac{g'}{(\nu + r)} \right] \Delta \hat{N}. \tag{206}$$

When this is inserted into the second equation (for $\Delta \hat{N}$), the effective slow equation takes the form

$$d\langle \Delta \hat{N} \rangle = \lambda_S \langle \Delta \hat{N} \rangle \tag{207}$$

with the reduced value

$$\lambda_S \approx \frac{g'r}{(\nu + r)} \tag{208}$$

obtained after a near cancellation between the two terms.

6 Concentration Fluctuations in Semiconductors

6.1 General Theory of Concentration Fluctuations

The thermodynamic discussion of occupancy fluctuations in Sec. 5.2 can be generalized by noting that the average occupancy of a state of energy E is given by

$$\langle n \rangle = \frac{1}{\{\exp[\beta(E - \mu) + \varepsilon\}]}, \tag{209}$$

where for Fermi, Boltzmann, and Bose particles

$$\varepsilon = \begin{cases} 1 & \text{for Fermi particles,} \\ 0 & \text{for Boltzmann particles,} \\ -1 & \text{for Bose particles.} \end{cases} \tag{210}$$

For application to a particular state a, replace n by $n(a)$ and E by $E(a)$. The fluctuation in occupancy of that state is given by the thermodynamic formula, Eq. (3), or the first part of Eq. (165), to be

$$\begin{aligned} \langle \Delta n(a) \Delta n(c) \rangle &= n(a)' \delta(a, c) \\ &\equiv n(a)[1 + \varepsilon n(a)] \delta(a, c), \end{aligned} \tag{211}$$

which includes all three statistical cases, Fermi, Boltzmann and Bose, with the three choices of ε above. This result is true in equilibrium.

We have also established the truth of Eq. (211) for the nonequilibrium steady state in Lax (1960) by explicitly constructing a model in which there are transition probabilities for the transfer of particles between states, by determining the diffusion coefficients for this model, and then by solving the Einstein relation, Eq. (14), for the second moments. Since the details

of the analysis are not particularly illuminating, they will not be repeated here except to note that if there is a constraint, such as the total number of particles in all the states being fixed, the simple formula, Eq. (211) above, is replaced by

$$\langle \Delta n(a) \Delta n(c) \rangle = \frac{n(a)\delta(a,c) - n(a)'n(c)'}{\sum_b n(b)'}. \tag{212}$$

This result clearly obeys the constraint

$$\left\langle \left[\sum_a \Delta n(a) \right] \Delta n(c) \right\rangle = 0. \tag{213}$$

Another point of interest is that the diffusion matrix in Lax (1960), Sec. 12, was found to have the diagonal elements

$$2D_{bb} = \text{total jump rate out of } b + \text{total jump rate into } b, \tag{214}$$

and the nondiagonal elements are

$$2D_{ab} = -[\text{transition rate from } a \text{ to } b + \text{rate from } b \text{ to } a]. \tag{215}$$

We therefore regard such diffusion constants as characteristic of shot noise because of particle-number quantization.

6.1.1 Application to Semiconductors with Electrons, Holes, and Traps

Equation (212) can be readily applied to the case in which $n(1) = N =$ the number of conduction electrons, $n(2) = \hat{N} =$ the number of trapped electrons, and $n(3) =$ number of electrons in the valence band $= N_v - P$, where N_v is the number of valence-band states and $P =$ the number of holes. Thus

$$n(1)' = N\left[1 - \frac{N}{N_c}\right] \approx N, \tag{216}$$

$$n(2)' = \hat{N}\left[1 - \frac{\hat{N}}{N_t}\right] = \hat{N}', \tag{217}$$

$$n(3)' = (N_v - P)\left(\frac{P}{N_v}\right) \approx P. \tag{218}$$

We have assumed nondegeneracy for the holes and the free electrons, but not for the trapped electrons. Since $\Delta n(3) = -P$, we can write the second moments in the form

$$\langle (\Delta N)^2 \rangle = \frac{N - N^2}{(P + N + \hat{N}')}, \tag{219}$$

$$\langle (\Delta \hat{N})^2 \rangle = \frac{\hat{N}' - (\hat{N}')^2}{(P + N + \hat{N})}, \tag{220}$$

$$\langle (\Delta P)^2 \rangle = \frac{P - P^2}{(P + N + \hat{N})}, \tag{221}$$

$$\langle \Delta N \Delta \hat{N} \rangle = -\frac{N\hat{N}}{(P + N + \hat{N})}, \tag{222}$$

$$\langle \Delta P \Delta N \rangle = \frac{PN}{(P + N + \hat{N})}, \tag{223}$$

$$\langle \Delta P \Delta \hat{N} \rangle = \frac{P\hat{N}}{(P + N + \hat{N})}. \tag{224}$$

These results are consistent with the constraint $\Delta P = \Delta N + \Delta \hat{N}$ of charge neutrality. If the number of traps is zero, they reduce to $\Delta N = \Delta P$ and

$$\langle (\Delta N)^2 \rangle = \langle (\Delta P)^2 \rangle = \frac{NP}{(N + P)}. \tag{225}$$

When $P \gg N$ and $P \gg \hat{N}$, these second moments reduce to those in Eq. (196), for the case of electrons and traps in the presence of majority holes.

6.2 The Influence of Drift and Diffusion on Resistance Modulation Noise

To concentrate on the influence of drift and diffusion on density fluctuations and

modulation noise, let us return to the trap-free case discussed in Sec. 5.1. The spectrum of voltage noise, already given in Eq. (162), can be rewritten as a product

$$G(V, \omega) = \langle(\Delta V)^2\rangle g(\omega) \tag{226}$$

of the total noise

$$\langle(\Delta V)^2\rangle = V^2 \left(\frac{1+b}{P+Nb}\right)^2 \langle(\Delta P)^2\rangle \tag{227}$$

and the normalized spectrum

$$g(\omega) = \int_0^\infty 4\cos\omega t dt \Phi(t), \tag{228}$$

where

$$1 = \int_0^\infty g(\omega) df = \int_{-\infty}^\infty g(\omega)\frac{d\omega}{4\pi}. \tag{229}$$

Note that this normalization is four times that used for $g(\omega)$ in Lax and Mengert (1960). For simplicity, we confine ourselves to a one-dimensional geometry, as was done by Hill and van Vliet (1958), and calculate the total hole fluctuation as

$$\Delta P(t) = \int_0^L \Delta p(x, t)\, dx. \tag{230}$$

We can now apply our techniques to continuous-parameter systems by replacing the sum

$$\alpha_i(t) = \sum_i [\exp(-\mathbf{\Lambda} t)]_{ij}\alpha_j(0) \tag{231}$$

by the integral

$$\alpha_x(t) = \int [\exp(-\mathbf{\Lambda} t)]_{xx'}\, dx' \alpha_{x'}(0). \tag{232}$$

Introducing a more convenient notation for the Green's function

$$[\exp(-\mathbf{\Lambda} t)]_{xx'} = K(x, x', t), \tag{233}$$

we can write

$$\Delta p(x, t) = \int K(x, x'', t)\, dx'' \Delta p(x'', 0), \tag{234}$$

so that the correlation at two times is, as usual, related to the pair correlation at the initial time:

$$\langle\Delta p(x, t)\Delta p(x', 0) = \int K(x, x'', t)\, dx'' \times \langle\Delta p(x'', 0)\Delta p(x', 0)\rangle. \tag{235}$$

It is customary to treat fluctuations at the same time at two places as uncorrelated. This is clearly the case, for independent carriers. It is less obvious when Coulomb attractions (say between electrons and holes) are included. It was shown, however, in Appendix C of Lax and Mengert (1960) that a delta-function correlation is valid, as long as we are dealing with distances greater than the screening radius. Thus we can take

$$\langle\Delta p(x'', 0)\Delta p(x', 0)\rangle = \langle(\Delta P)^2\rangle L^{-1}\delta(x'' - x'), \tag{236}$$

where the coefficient of the delta function is chosen so that the fluctuation in the total number of carriers $\langle(\Delta P)^2\rangle$ is given correctly by Eq. (225). Here L is the distance between the electrodes.

The definition, Eq. (163), of $\Phi(t)$ yields the expression

$$\Phi(t) = \frac{1}{L}\int_0^L dx \int_0^L dx' K(x, x', t). \tag{237}$$

If the Green's function is defined, appropriately as in Lax (1960), to vanish for $t < 0$, it will obey an equation of the form

$$\left(\frac{\partial}{\partial t} + \mathbf{\Lambda}\right) K(x, x', t) = \delta(x - x')\delta(t), \tag{238}$$

where, in the continuous-variable case,

$$\mathbf{\Lambda} = r + v\frac{\partial}{\partial x} - D\frac{\partial^2}{\partial x^2} \tag{239}$$

is an operator. Here, v and D are the ambipolar drift velocity and diffusion

constants found by Van Roosbroeck (1953) to describe the coupled motion of electrons and holes while maintaining charge neutrality:

$$v = e\mu_a E, \quad \mu_a = \frac{(N-P)\mu_n\mu_p}{(N\mu_n + P\mu_p)}; \quad (240)$$

$$D_a = \frac{(N+P)D_pD_n}{(ND_n + PD_p)}, \quad (241)$$

where the individual diffusion constants and mobilities are related by the Einstein relation.

Equation (238) for the Green's function can be solved by a Fourier-transform method:

$$K(x, x', t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \exp[ik(x-x') - \lambda(k)t]\, dk, \quad (242)$$

where

$$\lambda(k) = r + ivk + Dk^2. \quad (243)$$

Here $\lambda(k)$ are the eigenvalues of the $\mathbf{\Lambda}$ operator,

$$\mathbf{\Lambda}\exp(ikx) = \lambda(k)\exp(ikx). \quad (244)$$

With Eq. (242) for K, the after-effect function can be calculated from Eq. (237):

$$\Phi(t) = \frac{L}{\pi}\int_{-\infty}^{\infty} dk \frac{\sin^2(z)}{z^2}\exp[-\lambda(k)t], \quad (245)$$

where $z = kL/2$. Thus the spectrum, Eq. (228), is

$$g(\omega) = \frac{4}{\pi}\int_{-\infty}^{\infty} dz \frac{\sin^2 z}{z^2}\frac{1}{i\omega + \lambda(z)}, \quad (246)$$

where λ has been re-expressed as a function of z,

$$\lambda(z) = r + 2i\left(\frac{v}{L}\right)z + 4\left(\frac{D}{L^2}\right)z^2. \quad (247)$$

Lax and Mengert (1960) provide an exact evaluation of this integral. However, the resulting expressions are complicated. It is therefore worth while to treat some limiting cases. For example, if there is no diffusion, then

$$K(x, x', t) = \exp(-rt)\delta(x - x' - vt), \quad (248)$$

and the after-effect function is given by

$$\Phi(t) = \exp(-rt)\left[1 - \left(\frac{t}{T_a}\right)\right], \quad (249)$$

where $T_a = L/v$ is the transit time and the spectrum is governed by a windowing factor W:

$$g(\omega) = 4T_a W, \quad (250)$$

with the window factor given by

$$W = \int_0^{T_a} \cos\omega t e^{-rt}\left[1 - \frac{t}{T_a}\right] d\left(\frac{t}{T_a}\right). \quad (251)$$

Indeed, the current noise, in this special case, can be written in the form given by Hill and van Vliet (1958),

$$G(I, \omega) \equiv \left(\frac{I^2}{V^2}\right) G(I, \omega) = 2eI_{\text{eq}}, \quad (252)$$

which emphasizes the similarity to shot noise. The equivalent current is defined by

$$I_{\text{eq}} = 2I\frac{b + 2 + b^{-1}}{|P/N - N/P|}W. \quad (253)$$

The window factor still takes a complicated form,

$$W = \frac{(\tau/T_a)^2}{1+(\omega\tau)^2}\left\{\frac{T_a}{\tau} - \frac{1-(\omega\tau)^2}{1+(\omega\tau)^2} + \frac{\exp(-T_a/\tau)}{1+(\omega\tau)^2}\left\{[1-(\omega\tau)^2] \times \cos(\omega T_a) - 2\omega\tau\sin(\omega T_a)\right\}\right\}. \quad (254)$$

Here $\tau = 1/r$. Even this result is complicated to understand. If we take the limiting case when recombination is unimportant over the transit time, the result simplifies to

$$W = \frac{1}{2}\frac{\sin^2(\omega T_a/2)}{(\omega T_a/2)^2}, \tag{255}$$

a windowing factor similar to that found associated with the effect of transit time on shot noise. See Sec. 4.2.

In the opposite limit, in which diffusion is retained but drift is neglected, the exact result for the spectrum is given by

$$g(\omega) = 4\mathrm{Re}J, \tag{256}$$

where

$$J = (r + i\omega)^{-1} - \left[\frac{D}{(r + i\omega)^3}\right]^{1/2} \frac{[1 - \exp(-\Gamma L)]}{L} \tag{257}$$

and

$$\Gamma = \left[\frac{(r + i\omega)}{D}\right]^{1/2} \tag{258}$$

is the reciprocal of the diffusion length. The exponential term represents an interference term between the two boundaries that is usually negligible since they are separated by substantially more than a diffusion length. A simple approximate form over intermediate frequencies is

$$\frac{g(\omega)}{4} \approx \frac{r}{(\omega^2 + r^2)} + \left(\frac{1}{L}\right)\left(\frac{D}{2\omega^3}\right)^{1/2}. \tag{259}$$

In summary, in addition to the first term, which represents the volume noise easily computed just by using the total carrier, $P(t)$, effects, we get an inverse frequency to the three-halves power that arises from diffusion across the boundary at the electrodes.

7 Langevin Processes

7.1 Simplicity of Langevin Processes

Langevin treated noise by adding a random noise source to the linear equations describing the transport. Langevin methods, at least for a linear or quasilinear system, have the simplicity of the circuit equation of electrical engineering. The noise source may arise from thermal reservoirs as in Johnson noise, or shot noise from the discreteness of particles. But once the noise is represented as a voltage source with known moments, the physical nature of the source is no longer important. We have the analog for noise sources of current sources and Norton's theorem. The sources can be thought of as a black box, with an impedance in series with a voltage source, or an admittance in parallel with a current source. And the original nature of the sources will not enter into the solution of problem.

For the quasilinear case, we can write our set of Langevin equations in the form

$$\frac{d\boldsymbol{\alpha}}{dt} + \boldsymbol{\Lambda} \cdot \boldsymbol{\alpha} = \mathbf{F}(t), \tag{260}$$

where $\boldsymbol{\alpha} = \mathbf{a} - \langle \mathbf{a} \rangle$ is a multicomponent object, as is the force $\mathbf{F}(t)$. Equation (260) can be regarded as the definition of $\mathbf{F}(t)$. What condition must be imposed on $\mathbf{F}(t)$ in order that the resulting process $\boldsymbol{\alpha}(t)$ be a Markoffian process? (A Markoffian process is analogous to a student who only remembers the last thing he was told.) It was shown in Sec. 1 of Lax (1960) that a necessary and sufficient condition for $\boldsymbol{\alpha}(t)$ to be a Markoffian set of variables is for the Langevin forces $\mathbf{F}(t)$ to be delta correlated,

$$\langle \mathbf{F}(t)\mathbf{F}(u) \rangle = 2\mathbf{D}\delta(t - u). \tag{261}$$

The second moments in Eq. (261) are sufficient to calculate all second-order correlation functions of the α's for linear processes. Since these second-order correlations are insensitive to the higher moments of the noise sources, it is then permissible to regard the forces as Gaussian random variables. Then all linked moments beyond the second have been set equal to zero.

Shot-noise sources, however, do have higher linked moments than the second, and, at the linear level, we can consider a complete set of linked moments for any linear Markoffian process:

$$\langle F(t_1)F(t_2)\dots F(t_n)\rangle^L = n!\mathbf{D}_n\delta(t_1-t_2)\dots\delta(t_{n-1}-t_n). \quad (262)$$

The delta-function nature of these correlations guarantees the absence of memory required by a Markoffian process.

Indeed, if we choose

$$n!D_n = \frac{\langle(\Delta\alpha)^n\rangle}{\Delta t} = \nu\int \eta^n g(\eta)d\eta, \quad (263)$$

then the Langevin process, Eq. (260), reduces to the generalized one-dimensional shot-noise process of Sec. 4.2, associated with

$$\frac{da}{dt} = \sum_k \eta_k\delta(t-t_k), \quad (264)$$

where $g(\eta)$ is the probability of jumps in a of size η.

7.2 Relation to the Fokker–Planck Equation

To provide a bridge between the Langevin and Fokker–Planck points of view, we must show how the nth-order diffusion constants just introduced, by Eq. (262), into a Langevin description determine the motion of a general function $M(a)$ of the random process $a(t)$.

A general random process (not necessarily a Markoff one) necessarily obeys the relation on conditional probabilities

$$P(a', t+\Delta t|a_0, t_0) = \int P(a', t+\Delta t|a, t; a_0, t_0)daP(a, t|a_0, t_0). \quad (265)$$

The transition probability $P(a', t+\Delta t|a, t; a_0, t_0)$ describes the probability of arriving at a' at $t+\Delta t$ if one starts at a at time t, remembering that one started the entire process at a_0 at time t_0. This last bit of information may be omitted if the process is Markoffian. In that case Eq. (265) reduces to the Chapman–Kolmogoroff condition.

The nth-order conditional diffusion constant is defined by the relation

$$\begin{aligned} D_n(a, t|a_0t_0) &= \frac{1}{n!}\lim_{\Delta t\to 0}\int (a'-a)^n \times P(a', t+\Delta t|a, t; a_0, t_0)\frac{da'}{\Delta t} \\ &= \frac{1}{n!}\lim_{\Delta t\to 0}\frac{1}{\Delta t} \times \langle[a(t+\Delta t)-a(t)]^n\rangle|_{\substack{a(t)=a\\ a(t_0)=a_0}}. \end{aligned} \quad (266)$$

For a Markoffian process the dependence on a_0 can be omitted everywhere. Equation (265) can then be replaced by

$$P(a', t+\Delta t) = \int P(a', t+\Delta t|a, t)daP(a, t). \quad (267)$$

The average motion of an arbitrary function $M(a)$ of the random process $a(t)$ may be obtained by integrating $M(a')$ against $P(a', t+\Delta t)$. On the right-hand side of the equation, we shall replace $M(a')$ by its Taylor expansion

$$M(a') = M(a) + \sum_{n=1}^{\infty}(a'-a)^n\frac{M^{(n)}(a)}{n!} \quad (268)$$

The integrals over $(a' - a)^n$ give rise to the diffusion coefficients, so that we obtain

$$\langle M(a)\rangle_{t+\Delta t} = \langle M(a)\rangle_t + \Delta t \sum_{n=1}^{\infty} \int D_n(a, t) \times M^{(n)}(a) P(a, t)\, da. \quad (269)$$

Thus

$$\frac{d}{dt}\langle M(a)\rangle = \sum_{n=1}^{\infty} \left\langle D_n(a, t) \frac{\partial^n (M(a)}{\partial a^n} \right\rangle, \quad (270)$$

where for any function $g(a)$ the average means

$$\langle g(a)\rangle \equiv \int g(a) P(a, t)\, da. \quad (271)$$

To obtain the equation of motion for $P(a,t)$, we write Eq. (270) in the explicit from

$$\int \left[\sum_{n=1}^{\infty} D_n(a, t) P(a, t) \frac{\partial^n M(a)}{\partial a^n} - M(a) \frac{\partial P}{\partial t}(a, t) \right] da = 0. \quad (272)$$

After an integration by parts, we obtain

$$\int \left\{ \sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n [D_n(a, t) P(a, t)] - \frac{\partial P(a, t)}{\partial t} \right\} M(a) da = 0. \quad (273)$$

Since this equation is to be valid for any choice of $M(a)$, the coefficient of $M(a)$ in the above equations must vanish, yielding the generalized Fokker–Planck equation:

$$\frac{\partial P(a, t)}{\partial t} = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n \times [D_n(a, t) P(a, t)]. \quad (274)$$

The ordinary Fokker–Planck equation is the special case in which the series terminates at $n = 2$.

Since Eq. (267) was true for all processes, including non-Markoffian, the existence of a generalized Fokker–Planck equation does not guarantee that the associated process is Markoffian. If we had done all our averages retaining the initial condition at t_0, we would have obtained the equation of motion

$$\frac{d}{dt}\langle f(a)\rangle_{a(t_0)=a_0} = \times \sum_{n=1}^{\infty} \left\langle D_n(at|a_0 t_0) \frac{\partial^n f(a)}{\partial a^n} \right\rangle_{a(t_0)=a_0}, \quad (275)$$

and the Fokker–Planck equation is

$$\frac{\partial P(at|a_0 t_0)}{\partial t} = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\partial}{\partial a}\right)^n \times [D_n(a, t|a_0 t_0) P(a, t|a_0 t_0)]. \quad (276)$$

In the Markoffian case, Eq. (276) agrees with Eq. (274). Thus if (and only if) the process is Markoffian, $P(at|a_0 t_0)$ obeys the same equation of motion as $P(a,t)$. Then, we can calculate $P(at|a_0 t_0)$ by solving Eq. (274) subject to the initial condition

$$P(a, t) = \delta(a - a_0) \text{ at } t = t_0. \quad (277)$$

7.3 An Exactly Solvable Gaussian Example

We start with the simple example

$$\frac{dx}{dt} = \mu + \sigma f(t), \quad (278)$$

where μ and σ are constants and $f(t)$ is a Gaussian random process with mean zero and known autocorrelation

$$\langle f(t) f(u)\rangle = 2R(t - u). \quad (279)$$

Then the simpler variable $y = x - \mu t$, which obeys

$$\frac{dy}{dt} = \sigma f(t), \quad (280)$$

represents an integral over Gaussian variables of mean zero; hence it too is Gaussian with mean zero and variance

$$H(t) = \langle [y(t)]^2 \rangle = 2\sigma^2 \int_0^t \int_0^t drdsR(r-s). \tag{281}$$

We have used $x(0) = y(0) = 0$ as an initial condition. Thus the Gaussian probability distribution for y is given by

$$P(y,t) = \frac{1}{[2\pi H(t)]^{1/2}} \exp\left[-\frac{1}{2}\frac{y^2}{H(t)}\right]. \tag{282}$$

The special case in which

$$H(t) = Ct^{2H}, \tag{283}$$

in which C is an arbitrary constant and H is a constant between zero and one, describes "fractional Brownian motion." Ordinary Brownian motion is the special case $H = 1/2$.

A more interesting generalization occurs when one sets

$$\frac{S}{S(0)} = \exp(x) = \exp(y + \mu t). \tag{284}$$

Using $dx = dS/S$, Eq. (278) becomes

$$\frac{dS}{dt} = \mu S(t) + \sigma S(t) f(t). \tag{285}$$

The relations

$$y = \ln\left[\frac{S}{S(0)}\exp(-\mu t)\right], \quad dy = \frac{dS}{S}, \tag{286}$$

can be substituted into Eq. (282) to obtain the distribution for S:

$$\begin{aligned}\hat{P}(S,t|S(0),0)dS &= P(y,t|0,0)dy \\ &= \frac{1}{[2\pi H(t)]^{1/2}} \\ &\times \exp\left[-\frac{\ln^2[Se^{-\mu t}/S(0)]}{2H(t)}\right]\frac{dS}{S}.\end{aligned} \tag{287}$$

In the Brownian-motion limit, the noise is white, as represented by a delta correlation, $R(t-u) = \delta(t-u)$. Equation (281) specializes to $H(t) = 2Dt$ with $D = \sigma^2$. Equation (287) then reduces to

$$\begin{aligned}\hat{P}(S,t|S_0,0) &= \frac{S(0)}{S}\frac{1}{(4\pi Dt)^{1/2}} \\ &\times \exp\left[-\frac{\ln^2[Se^{-\mu t}/S(0)]}{4Dt}\right].\end{aligned} \tag{288}$$

The Fokker–Planck equation obeyed by $P(x,t)$ follows from Eqs. (278) and (276) in the Brownian-motion limit in which $D_n = 0$ for $n > 2$,

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\mu P] + \frac{\partial^2}{\partial x^2}[\sigma^2 P]; \tag{289}$$

it contains the constant diffusion term $D = \sigma^2$, and the constant drift term $A = \mu$. One can obtain the equation for

$$\hat{P}(S,t) = \frac{P(x,t)S(0)}{S}, \tag{290}$$

by starting from the previous equation for $P(x,t)$, introducing the relation, Eq. (284), of x to S. The result after some labor is

$$\frac{\partial \hat{P}}{\partial t} = -\frac{\partial}{\partial S}[(\mu + \sigma^2)S\hat{P}] + \frac{\partial^2}{\partial S^2}[\sigma^2 S^2 \hat{P}]. \tag{291}$$

The diffusion term [in the notation of Eq. (8)] $D_{SS} = \sigma^2 S^2$ is easy to understand. We can write

$$D_{SS} \equiv \frac{(\Delta S)^2}{2\Delta t} = \left(\frac{\partial S}{\partial x}\right)^2 \frac{(\Delta x)^2}{2\Delta t}. \tag{292}$$

Thus the diffusion constants are related by

$$D_{SS} = S^2 D_{xx} = S^2 \sigma^2. \tag{293}$$

One might think that

$$\frac{\Delta S}{\Delta t} = \frac{\partial S}{\partial x}\frac{\Delta x}{\Delta t}. \tag{294}$$

However, this choice omits the second term in the drift vector. There is a correction because the diffusion coefficient D depends on S:

$$\frac{\Delta S}{\Delta t} = \frac{\partial S}{\partial x}\frac{\Delta x}{\Delta t} + \frac{1}{2}\frac{\partial^2 S}{\partial x^2}\frac{(\Delta x)^2}{\Delta t} \tag{295}$$

or

$$A_S = SA_x + \frac{1}{2}S(2D) = S(\mu + \sigma^2) \tag{296}$$

in agreement with Eq. (291).

7.4 Stochastic Integrals: The Ito–Stratonovich Controversy

Because mathematicians concentrate on Brownian motion, which is rather singular in behavior, it is not clear how to define the integral of a product of a random variable and a random force. In particular, how should we convert the differential equation

$$\frac{da}{dt} = B(a(t)) + \sigma(a)f(t), \tag{297}$$

where $f(t)$ is a standard white-noise source with

$$\langle f(t)f(u)\rangle = 2\delta(t-u), \tag{298}$$

to an integral? The Riemann sum of integral calculus would be

$$a(t) = a(0) + \sum_j [B(a(\bar{t}_j)) + \sigma(a(\bar{t}_j))f(\bar{t}_j)] \times (t_{j+1} - t_j), \tag{299}$$

where

$$t_j \le \bar{t}_j \le t_{j+1}. \tag{300}$$

The Riemann integral exists if the sum approaches a limit independent of the placement of $\bar{t}_j$ in the interval in Eq. (300). But the conditions of boundedness required for the existence of the Riemann integral (see Jeffries and Jeffreys, 1950) are violated. Thus the value of the sum approaches different limits, according to how the point $\bar{t}_j$ is chosen. [This is not true for the B term, but only the one involving the white-noise source $f(t)$.] Ito (1951), Ito and McKean (1965), and Doob (1953) avoid the difficulty by evaluating σ at the beginning of the interval. They also convert the integral to Stieltjes form by introducing

$$w(t) \equiv \int_0^t f(s)\,ds; \quad \langle [w(t)]^2\rangle = 2t. \tag{301}$$

Some authors omit the factor of 2 in Eqs. (298) and (301). The result is

$$\int^t \sigma(a(s))f(s)ds = \sum \sigma(a(t_j)) \times [w(t_{j+1}) - w(t_j)]. \tag{302}$$

However, even the latter does not converge to a unique integral, and the evaluation at the beginning is an arbitrary choice. The effect of this choice, since $f(s)$ is independent of $a(t)$ for $t < s$, is that the average of the second term in Eq. (297) vanishes, so that the Ito drift vector that follows from Eq. (297) and the integration rule, Eq. (302), is

$$\text{Ito}: A(a) = B(a) + \langle \sigma(t)f(t)\rangle = B(a). \tag{303}$$

Stratonovich (1963) makes the fortuitous choice of using

$$\overline{a(t_j)} = \frac{1}{2}[a(t_j) + a(t_{j+1})], \tag{304}$$

the average of the values at the two end points of each interval. In our notation, Stratonovich's choice is equivalent to equating

$$A(a) = B(a) + \langle \sigma(a) f(t) \rangle = B(a) + \sigma(a) \frac{\partial \sigma(a)}{\partial a}, \quad (305)$$

a result in agreement with the outcome, Eq. (311), below, of the iterative procedure we advocate. It is intuitively clear that an average of the end points is better than using either one. But is the average value always the best choice?

Our viewpoint, expressed in Lax (1966b), Sec. 3, is that mathematicians have concentrated too exclusively on the Brownian-motion white-noise process, which has delta correlation functions. Real processes can have a sharp correlation time of finite width. Thus their spectrum is flat, but not up to infinite frequency. Thus, for real processes, the Riemann sums do converge, and no ambiguity exists. After the integration is performed, the correlation time can be allowed to go to zero, that is, one can then approach the white-noise limit. Thus the ambiguity is removed by approaching the integration limit and the white-noise limit in the correct order.

To obtain the drift vector $A(a)$, the differential equation, Eq. (297), is rewritten as an integral equation:

$$a(t+\Delta t) - a(t) = \int_t^{t+\Delta t} B(a(s)) ds + \int_t^{t+\Delta t} \sigma(a(s)) f(s)\, ds. \quad (306)$$

This equation can be solved by iteration. The lowest approximation is obtained by replacing $a(s)$ by $a(t)$ under the integral. This first approximation,

$$(\Delta a)_1 = B(a(t))\Delta t + \sigma(a(t)) \int_t^{t+\Delta t} f(s)\, ds, \quad (307)$$

is equivalent to the Ito choice. The first term is already of order Δt and need not be improved. In the second term, let us insert the first approximation,

$$a(s) = a(t) + B(a(t))(s-t) + \sigma(a(t)) \int_t^s f(u)\, du, \quad (308)$$

into Eq. (306) to get

$$(\Delta a)_2 = B(a(t))\Delta t + \sigma(a(t)) \int_t^{t+\Delta t} f(s) ds + \sigma(a(t)) \frac{\partial \sigma}{\partial a(t)} \int_t^{t+\Delta t} ds \int_t^s du f(s) f(u). \quad (309)$$

We have retained only terms of order Δt, or f^2, but not $\Delta t f$ or higher. If we restrict ourselves to Gaussian processes, their linked moments of order higher than two vanish. Thus the process terminates. Indeed,

$$\langle (\Delta a)^n \rangle_L = 0 \text{ for } n > 2. \quad (310)$$

For $n = 2$ it is sufficient to use the first approximation,

$$\langle (\Delta a)^2 \rangle_L = \sigma^2 \int_t^{t+\Delta t} ds \int_t^{t+\Delta t} du \langle f(s) f(u) \rangle = 2\sigma^2 \Delta t = 2! D_2 \Delta t. \quad (311)$$

In the last step, we have specialized to the Brownian, white-noise limit. For $n = 1$ it is necessary and sufficient to use the second

approximation of Eq. (309) to get

$$\langle \Delta a \rangle = B(a(t))\Delta t + \sigma(a(t))\frac{\partial \sigma}{\partial a(t)} \times \int_t^{t+\Delta t} ds \int_t^s du 2\delta(s-u),$$

so that

$$A(a)\Delta t = B(a)\Delta t + \sigma \frac{\partial \sigma}{\partial a}\Delta t. \qquad (312)$$

The factor of 2 disappears because only half the area of the delta function contributes at the boundary point s. This result is clear if the correlation function $R(s-u)$ is any sharp symmetric function. In summary, our Langevin process is equivalent to an ordinary (no derivatives higher than second) Fokker–Planck process with coefficients

$$D(a) = [\sigma(a)]^2, \qquad (313)$$

$$A(a) = B(a) + \sigma \frac{\partial \sigma}{\partial a} = B(a) + \frac{1}{2}\frac{\partial D}{\partial a}. \qquad (314)$$

These results agree completely with those found in the previous section for our exactly solvable example.

We note that the Fokker–Planck can be written in two completely equivalent forms:

$$\frac{\partial P(a)}{\partial t} = -\frac{\partial}{\partial a}[A(a)P(a)] + \frac{1}{2}\frac{\partial^2}{\partial a^2}[D(a)P(a)] = -\frac{\partial}{\partial a}[B(a)P(a)] + \frac{\partial}{\partial a}\left[\sigma(a)\frac{\partial}{\partial a}(\sigma(a)P(a))\right]. \qquad (315)$$

The Ito choice can be compensated for by choosing the correct Fokker–Planck equation, but his procedure is sufficiently counterintuitive that the wrong choice is often made. In applying Ito's procedure to the financial world Hull (1992) almost always makes the correct choice. Hull uses the equation

$$dx = \frac{dS}{S} = \mu dt + \sigma\, dw, \qquad (316)$$

where w is the Wiener process defined in Eq. (299), which is simply Eq. (278) written in the notation preferred by mathematicians. Since σ is a constant in Hull's application, the second term on the right-hand side of Eq. (314) vanishes, and there is no distinction between the Ito and Stratonovich viewpoints. Hull (1992) then states correctly that $\Delta S/S$ is normally distributed with mean "$\mu\Delta t$ and standard deviation $\sigma\sqrt{\Delta t}$. In other words,

$$\frac{\Delta S}{S} \approx \phi(\mu\Delta t, \sigma\sqrt{\Delta t}), \qquad (317)$$

where $\phi(m, s)$ denotes a normal distribution with mean m and standard deviation s." Equation (316) cannot be simply rewritten (Hull, 1992) as an Ito equation in the form

$$dS = \mu S dt + \sigma S dw \qquad (318)$$

because Ito's variables do not obey the usual rules of calculus. Equation (318) is correct, as a Stratonovich equation, but the second term does not average to zero but modifies the drift term as in Eqs. (305) and (309). To be an Ito equation, the first term should contain A, the modified drift term, and not B, the unmodified drift term. Equation (291) shows that the correct Ito equation is

$$dS = (\mu + \sigma^2)S dt + \sigma S\, dw. \qquad (319)$$

This result could also have been obtained using Ito's lemma, which describes how the drift vector changes under a transformation of variables. If, however, one makes the natural error of regarding Eq. (318) as a valid Ito equation and then makes the inverse transformation

$x = \ln S/S(0)$ using Ito's lemma (now carefully), one obtains (Hull, 1992)

$$dx = [\mu - \sigma^2]dt + \sigma\, dw, \tag{320}$$

which disagrees with the natural starting point, Eq. (316), as does the associated distribution after a finite time interval $\Delta t = T - t$, which takes the form

$$\ln S_T - \ln S \approx \phi(\mu - \sigma^2 \Delta t, \sigma\sqrt{\Delta t}) \tag{321}$$

that disagrees with Hull's earlier intuitively correct result, Eq. (317). The Ito procedure can be done correctly by using Eq. (319) instead of (318), but its counterintuitive nature can be misleading even to experts.

For the record, we note that the behavior of diffusion constants and drift vectors under a transformation from one set of variables a_j to a new set of variables a'_i was derived in Lax (1966b) from Langevin considerations that are a generalization of the above arguments to the multivariable case. The diffusion constants transform simply as

$$D'_{ij} = \frac{\partial a'_i}{\partial a_k}\frac{\partial a'_j}{\partial a_l} D_{kl}. \tag{322}$$

The B drift vectors also transform simply:

$$B'_i = \frac{\partial a'_i}{\partial t} + \frac{\partial a'_i}{\partial a_k} B_k, \tag{323}$$

where the second term is the natural transformation of a vector, and the first term enters only if the transformation is time dependent. Under a nonlinear transformation, however, there is a change in the contribution from the nonconstancy of the diffusion coefficient as shown by the last term in Eqs. (309)–(314). The A drift vectors, which enter the Fokker–Planck equation, possess therefore the more complicated transformation

$$A'_i = \frac{\partial a'_i}{\partial t} + \frac{\partial a'_i}{\partial a_k} A_k + \frac{\partial^2 a'_i}{\partial a_m \partial a_n} D_{mn}, \tag{324}$$

which can be understood simply from the relation

$$\Delta a'_i = \frac{\partial a'_i}{\partial t}\Delta t + \frac{\partial a'_i}{\partial a_k}\Delta a_k + \frac{\partial^2 a'_i}{\partial a_m \partial a_n}\frac{\Delta a_m \Delta a_n}{2}. \tag{325}$$

In summary, there are two kinds of Langevin equations, those whose random term need not average to zero [as used in Eq. (297)] and those used by Ito in which the average of the random term vanishes by Ito's definition of the stochastic integral. Both can be used correctly, but Ito's choice requires more care because the usually permissible way in which we handle equations is no longer valid. An able analysis of this situation, with references to earlier discussion, is presented by Van Kampen (1992).

Hull's (1992) book is an extremely well-written text on *Options, Futures and Other Derivative Securatives*. The Ito–Stratonovich controversy applies to physics, chemistry, and other fields. We have analyzed Hull's treatment in detail because his use of the Black–Scholes (1973) work involving an application of the Langevin Eq. (316) to the pricing of options coincides with the one example, in Sec. (7.3), for which we have an exact solution.

Although the Ito choice can be dangerous, as shown above, we trust that by now the Ito choice is used consistently in practice. However, there may be more serious problems, since real options may have statistics based on more wildly fluctuating processes than Brownian motion,

such as Lévy and fractal processes. This has recently been emphasized by Peters (1994) and Bouchard and Sornette (1994). For an excellent review of Brownian and fractal walks. Ghashghaie et al. (1996) have also found a parallelism between prices in foreign exchange markets and turbulent fluid flow. Thus the conventional Brownian-motion approach will be invalid in such markets.

8 Further Contributions to Stochastic Processes

8.1 Overview

In the following sections we review current contributions. We start with random walks, since the approach remains linear, and the applications are usually stationary. Then we discuss contributions in which one or more of our assumptions of (1) stationarity, (2) linearity, and (3) white noise are eliminated.

8.2 Random-Walk Problems

Many problems can be mapped into a random-walk problem on a lattice, or actually is a problem on a lattice. These problems are discussed first since the master equation for the probability density is linear, and simple techniques based on discrete periodicity can be used.

Recent contributions have been made to diffusion and reaction kinetics by considering one or more walkers on a lattice. Discrete-time random walks have been generalized to continuous-time walks. For such walks the distribution $\Psi(t)$ of time intervals to a hop from a given site are not exponential. If the first or second moment of $\Psi(t)$ is not finite, one generates a "Lévy" process. Other choices lead to stretched-exponential processes. Analytic solutions are often possible by taking advantage of the regularity of the lattice, as has been shown by Montroll, Weiss, Scher, Shlesinger, and others. An extension was also made to walks on fractals and to disordered lattices. Scher and Lax (1973) showed that a disordered lattice could be replaced by a continuous-time random walk that yields a conductivity varying approximately as $\omega^{0.9}$, in agreement with the experimental conductivity found in semiconductor impurity bands.

8.3 Linear with Time-Dependent Decay

The case of homogeneous noise with linear (time-dependent) damping [namely, $\mathbf{\Lambda}(t)$, but $\mathbf{D}(t) = \mathbf{D}$] is only mentioned here since an analytic evaluation was found for the generalized characterized functional

$$M[q] = \left\langle \exp\left[i \int_0^t q(u) a(u)\, du \right] \right\rangle, \quad (326)$$

where $q(u)$ is an arbitrary function of u, in Eq. (2.16) of Lax (1966b). Many desired multiple-time averages can be deduced from this generalized characterized function. For example, the multiple-time moments are given by a functional derivative

$$m(t_1, t_2, \ldots t_n) = i^{-n} \left. \frac{\delta^n M[q]}{\delta q(t_1) \ldots \delta q(t_n)} \right|_{q=0} . \quad (327)$$

as discussed by Hänggi (1989).

Lax (1966a) introduced a more general functional involving an arbitrary nonlinear function of $q(u)$, and Lax and Zwanziger (1973) evaluated it for quadratic functions

to obtain the photocount distribution in lasers near threshold.

8.4 The Nonlinear (Fokker–Planck) Case: Reaction-Rate Theory

If we eliminate the linearity assumption, including the case of weak noise and a quasilinearity approximation, many possibilities arise. If the variables are continuous and the noise is white and Gaussian, the problem is properly described by a Fokker–Planck equation.

Many applications have been made of Fokker–Planck theory, often with approximations that are special to a particular application. One example is to the theory of self-sustained oscillators discussed in Sec. 8.5. Applications have been made to systems subject to bifurcation, and to quantum mechanical systems. An application of long standing is to reaction-rate theory, which is described briefly here, and to stochastic resonance, which is described briefly in Sec. 8.5. Many of these applications are reviewed in the three-volume series by Moss and McClintock (1989). An excellent review of reaction-rate theory, which we cannot summarize adequately here, is the 90-page review by Hänggi et al. (1990).

The modern surge of effort in reaction-rate theory stems from the widely read article by Kramers (1940), since the author made a simple Fokker–Plank model for the escape of a particle from a metastable state. However, Landauer (1989), in a review, points out many examples of the rediscovery of previously published ideas. In particular, he remarks on an earlier article by Pontryagin et al. (1933) (PAV) that contained the essential ideas in Kramers's article.

Kramers's model, in the notation of Jung (1993), is

$$\frac{dx}{dt} = \frac{p}{m} \tag{328}$$

$$\frac{dp}{dt} = -\gamma p - V'(x) + \sqrt{m\gamma k_{\mathrm{B}} T}\xi(t) + E\sin\theta;$$

$$\theta \equiv \Omega t + \text{const.}, \tag{329}$$

where $E\sin\theta$ is the periodic term needed in the next section on stochastic resonance, and the Brownian motion is induced by the noise with second moment

$$\langle \xi(t)\xi(u)\rangle = 2\delta(t-u). \tag{330}$$

(The periodic term is not included in the articles by Kramers or PAV.)

Kramers considered both weak and strong damping cases. We review only the strong-damping case here because Kramers's extensive mathematical treatment was reviewed by Hänggi et al. (1990) and the main ideas, contained in the strong-damping case, are easier to explain because they lead to a closed-form solution.

In the strong-damping case, we assume that γ is faster than any other decay mechanisms, or reciprocal times. Then d/dt can be neglected compared to γ. Thus the term in dp/dt can be omitted. It is thus possible to solve Eq. (329) for p and substitute the result into Eq. (328). Discussions of the adiabatic approximation are given by Lax (1967) and Van Kampen (1985). The resulting equation for x is

$$\frac{dx}{dt} = A(x) + \sqrt{D}\xi(t), \tag{331}$$

where

$$A(x) = -\frac{V'(x)}{m\gamma} + \left(\frac{E}{m\omega}\right)\sin\theta, \tag{332}$$

$$D = \frac{k_B T}{m\gamma}. \tag{333}$$

The Fokker–Planck equation associated with Eqs. (332) and (333) is

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[A(x,t)P(x,t)] + \frac{\partial^2}{\partial x^2}[D(x)P(x,t)]. \tag{334}$$

In the PAV article, there is no time-dependent force, and $A(x)$ and $D(x)$ are regarded as known arbitrary functions of x [with $D(x)$ written as $b(x)/2$].

The Fokker–Planck equation can be regarded as an equation of continuity:

$$\frac{\partial P(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0, \tag{335}$$

where

$$J(x,t) = A(x,t)P(x,t) - \frac{\partial D(x)P(x,t)}{\partial x}. \tag{336}$$

If the periodic force is absent, there will be a stationary solution if $A(x)$ is a restoring force, namely, negative for large positive x and positive for large negative x. Such a stationary solution obeys

$$\frac{dJ(x)}{dx} = 0;$$

$$J(x) = AP(x) - \frac{d[D(x)P(x)]}{dx}, \tag{337}$$

and the current must be a constant,

$$J(x) = C_1. \tag{338}$$

Although Lax (1966b), Sec. 4, has encountered a case in which a current flow is present, we are normally concerned with stationary states that carry no current. If we set $C_1 = 0$, $P(x)$ obeys a simple first-order equation whose solution will be denoted $P_0(x)$:

$$P_0(x) = \frac{C}{D(x)} \exp[-U(x)]\, U(x) \equiv -\int_{x_{st}}^{x} \frac{A(y)}{D(y)}\, dy. \tag{339}$$

Here it is customary to choose x_{st} to be the stationary point, which obeys

$$A'(x_{st}) = 0. \tag{340}$$

This is the point, Eq. (10), about which a quasilinear expansion would be made. PAV then introduce a probability $\phi(x,t)$ that a particle starting at $t = 0$ in the interval $a < x < b$ will reach a boundary in the time interval [0,t]. Instead of a direct attack on this problem, they introduce the modern approach of invariant embedding made popular by Bellman (1968) and Casti and Kalaba (1973) by writing a recursion relation relating $\phi(x, t+\tau)$ to $\phi(x,t)$ for small τ:

$$\phi(x, t+\tau) = \int_a^b \phi(y,t)\, dy\, P(y, t+\tau | x, t), \tag{341}$$

which assumes that a limit in which $\tau \to 0$ will be taken and that "the probability of the random point leaving the interval ab during the short time τ is very small." PAV then expand $\phi(y,t)$ about $\phi(x,t)$. The results involve the incomplete moments

$$\mu_n = \int_a^b (y-x)^n\, dy. \tag{342}$$

Because the conditional probability is so narrow in y (of order $\sqrt{D\tau}$), the limits of integration can be extended to full interval $[-\infty, \infty]$ for almost all x. Then

$$\mu_0 \to 1; \quad \mu_1 \to A(x)\tau; \quad \mu_2 \to D(x)\tau; \tag{343}$$

with higher moments vanishing faster than linearly in τ. The result is that ϕ

obeys the equation adjoint to Eq. (334) for $P(x,t)$:

$$\frac{\partial\phi(x,t)}{\partial t} = \frac{\partial\phi}{\partial x}A(x) + \frac{\partial^2\phi}{\partial x^2}D(x). \quad (344)$$

Equation (344) is no easier to solve analytically than the original Fokker–Planck equation (334). However, $dt\partial\phi/\partial t$ is the probability of a crossing in the interval $[t, t+dt]$, so that

$$M(x) = \int_0^\infty t\frac{\partial\phi}{\partial t}dt \quad (345)$$

is the mean time to a crossing, starting at x. By taking the time derivative of Eq. (344), multiplying by t, and integrating over t, PAV obtain

$$\int_0^\infty t\frac{\partial^2\phi}{\partial t^2}dt = A(x)\frac{dM(x)}{dx} + D(x)\frac{d^2M(x)}{dx^2}. \quad (346)$$

Since $\phi(x,0) = 0$ and $\phi(x,\infty) = 1$, integration by parts shows that the left-hand side of Eq. (346) is -1. Thus $dM(x)/dx$ obeys a first-order ordinary differential equation. The integral of this equation subject to the boundary conditions

$$M'(a) = 0, \quad M(b) = 0, \quad (347)$$

appropriate to the case of a being a reflecting boundary and b being an absorbing boundary, is

$$M'(x) = -\exp[U(x)]\int_a^x \exp[-U(z)]\frac{dz}{D(z)}. \quad (348)$$

As pointed out by Stratonovich (1989), the final integral can be written in terms of the steady-state distribution itself:

$$M(x) = \int_x^b \frac{d\gamma}{D(\gamma)P_0(\gamma)}\int_a^\gamma dz P_0(z). \quad (349)$$

More explicit formulas are available only after making some approximations. Stratonovich (1989) simplifies the mean lifetime to

$$\begin{aligned}\bar{T} &= 2M(x_{st}) \\ &= 2\pi\left[\frac{D(b)}{D(x_{st})}\right]^{1/2}[|A'(x_{st})|A'(b)]^{-1/2} \\ &\quad\times\exp[U(b) - U(x_{st})] = 2T_{\mathrm{K}}, \quad (350)\end{aligned}$$

where T_K is referred to as the Kramers time. The factor of 2 was inserted on the assumption that after reaching the point b there is a 50% chance of escaping and a 50% chance of returning to the point x_{st} of stable equilibrium. If we revert to the original definitions of A and D in Eqs. (332) and (333), the Kramers time is given by

$$T_{\mathrm{K}} = \pi\frac{\gamma}{\omega_{st}\omega_b}\exp\left\{\frac{V(b) - V(x_{st})}{k_{\mathrm{B}}T}\right\}, \quad (351)$$

where $(\omega_j)^2 = |V''(x_j)|/m$ are the frequencies associated with quadratic approximations to the local potential at the points in question.

The results display the expected Arrhenius law associated with the activation energy. The mean rate of transition vanishes with γ demonstrating explicitly that the noise associated with the damping induces the transitions.

Although we have demonstrated the result for the one-dimensional case, the original authors all considered x to be a one-dimensional coordinate along the direction of the reaction embedded in a multidimensional space. The qualitative ideas remain the same, but the results are, of course, sensitive to details such as symmetry that are not present in the one-dimensional case. Contributions by Brinkman (1956), Landauer and Swanson (1961), and Langer (1968) are summarized in detail by Hänggi et al. (1990). A recent review of the current status of Kramers's

reaction-rate theory is given by Talkner and Hänggi (1995).

8.5 Stochastic Resonance

The subject of stochastic resonance is concerned with systems such as those discussed by Kramers, Pontryagin et al., Stratonovich, and others, to which has been added an applied periodic force. Such a term was already included in Eq. (329). Since this field is more complex than the original reaction-rate theory, there are more limits and approximations to consider. The aim of this section is to indicate the different regions for which approximate solutions have been found and to explain the nature of stochastic resonance, namely, the way in which noise-induced transitions are enhanced by a periodic force, and conversely the way in which noise can enhance the response to a periodic force. For a detailed examination of stochastic resonance, the reader is referred to the review of this subject by Jung (1993) and to the NATO Conference Proceedings (Bulsara et al., 1993).

The simplest case is the application of a periodic force to an Uhlenbeck–Ornstein (1930) system. Since this system is linear, an analytical solution is possible, both of the Fokker–Planck equations and of the Langevin equations themselves. Since the linear response to Gaussian forces is Gaussian, the full time-dependent distribution function is a Gaussian in all its variables. The form is the same as if there were no force applied, except that each variable has a mean shifted by the force. The shift is, in fact, simply calculable by ignoring the noise. The Gaussian distribution function is readily expressible in terms of the second moments (taken relative to the mean). And these are unchanged from those in the original Uhlenbeck–Ornstein article. For details see Jung (1993).

The next simplest case is that in which the frequency of the driving force is low compared to all the other rates in the problem. Then the adiabatic procedure can be applied just as when no periodic forces were present. We discuss one such example later.

The case of a high-frequency force can only be handled (by perturbation theory) if the force is weak. Since the force is periodic, the solution for the Fokker–Planck probability distribution takes the Floquet form. Perturbation theory has been applied to obtain correlation functions and/or the Green's function for the Fokker–Planck equation in terms of the corresponding solutions in the absence of a force by Presilla et al. (1989), Fox (1989), and Gang et al. (1990). For the escape problem, an extensive set of numerical results are plotted in Jung (1993). For large driving frequencies, the enhancement of escape rates, according to Jung, Eq. (8.43), decrease inversely with the square of the driving frequency. Resonance activation appears to occur. But the results are largely numerical, except for one result quoted by Jung from an unpublished thesis by Linkowitz (1989).

A simple qualitative explanation for the fairly broad stochastic resonance is that it occurs when the potential is modulated at a period T that is synchronized with the Kramers reaction time T_K. In a double-well system, the potential modulation must complete one period in a time T just large enough for a Kramers transition from one well to the other and back. See Fig. 5.

Thus we require

$$T = 2T_K,$$

where T_K is the Kramers time given in Eq. (350). This resonance not only appears

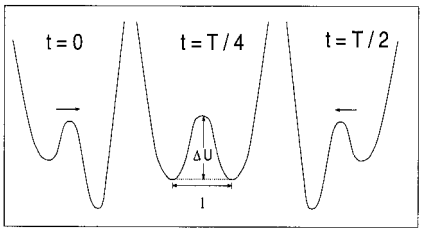


Fig. 5 Bistable potential with external modulation, shown at three times [after Jung (1993), but recomputed]

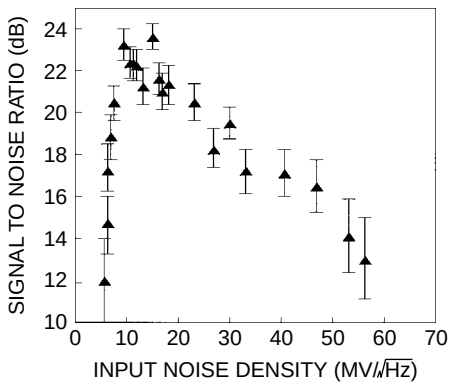


Fig. 6 Signal-to-noise ratio of one mode in a bistable laser. Taken from Jung (1993), based on experimental results of Roy and Mandel (1980) and Lett and Mandel (1985)

in the escape rate but also appears as an enhancement in the output spectrum at the driving frequency. The signal-to-noise ratio in a bistable ring laser displays this smooth resonance in Fig. 6

8.6 Self-Sustained Oscillators

The quasilinear approximation to a nonlinear system can break down because of an instability in some coordinate. Lax (1968a) and Hempstead and Lax (1967) have argued that all autonomous self-sustained oscillators possess one degree of freedom that is neutrally unstable. These are oscillators described by a set of ordinary differential equations containing no coefficients that depend explicitly on the time. For this reason, if $a_j(t)$ is a solution (in the absence of noise), then so is $a_j(t+\tau)$ for arbitrary τ. This instability against a shift of the time origin is equivalent to an instability in a phaselike variable.

The onset of oscillation is like a phase transition characterized by a long time constant. Thus most internal degrees are relatively fast and can be adiabatically eliminated, leaving one complex (field) amplitude to describe the behavior near threshold.

A self-sustained oscillator operates by connecting an absorptive system, such as a laser cavity containing absorbing materials and windows for energy to escape, to a gain system such as a set of pumped atoms. In equivalent-circuit terms, the absorptive system has a positive resistance, and the pumped system has a negative resistance. The oscillator will stabilize at a level of operation at which these two (nonlinear) resistances cancel. The frequency of operation will be that at which the reactances cancel. Lax and Louisell (1967) found an equation for a laser of the form

$$\frac{d\beta}{dt} = -(1 - i\alpha)R(|\beta|^2)\beta + f(t), \quad (352)$$

for the complex electromagnetic field amplitude β after the rotating-wave approximation has been made.

The parameter α, which first appeared as the detuning parameter in Lax (1965) also appears in Lax (1967) as a coupling parameter between amplitude and phase fluctuations. [The parameter α of Lax (1965) is identical to $\tan\beta$ of Lax (1967). Thus in both cases, the linewidth is enhanced by a factor $1+\alpha^2$.] It is

this coupling between intensity and phase fluctuations that produces the strong enhancement of linewidth in semiconductor lasers found by Fleming and Mooradian (1981) and explained by Henry (1986). The function R describes the dependence of the net resistance on the intensity $|\beta|^2$. At the operating point, $|\beta|^2 = p$, this resistance must vanish. For any well-designed self-sustained oscillator, it was shown in Appendix B of Hempstead and Lax (1967) that it is adequate to expand the resistance to the linear term $|\beta|^2 - p$. The relative importance of the first neglected term is shown to be of the order of the ratio of the noise in the oscillator to the signal. For a well-designed oscillator, this ratio is small. In Appendix A of Hempstead and Lax, the ratio is shown to be one over the number of photons in the cavity. Once this expansion is made, it is possible to rescale both amplitude and time to obtain

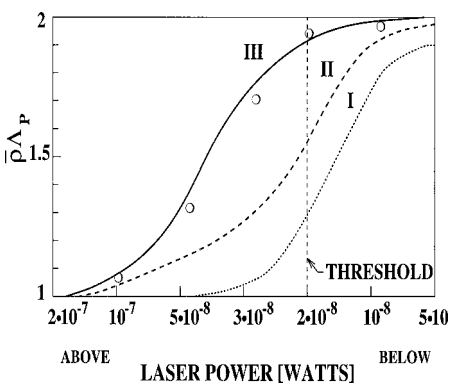


Fig. 7 Spectrum for phase and amplitude fluctuations of a laser is Lorentzian to a good approximation. The dimensionless half-width Λ_p times the dimensionless mean power $\bar{\rho} \equiv \langle|\beta|^2\rangle$ is plotted against laser power. The experimental results are those of Gerhardt et al. (1972). The theoretical curves I and II are due to Grossman and Richter (1971), and curve III is due to Risken and Vollmer (1967) and Hempstead and Lax (1967)

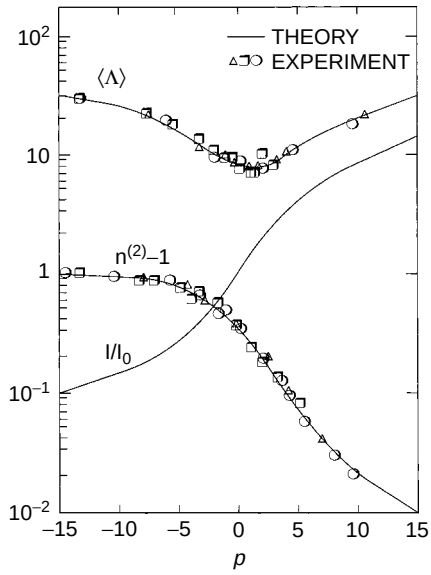


Fig. 8 Experimental verification of the Van der Pol model for lasers operating near threshold. The theoretical curves are from Hempstead and Lax (1967). The experimental points △, □, and ○ are from Gamo et al. (1968), Davidson and Mandel (1967), and Arecchi et al. (1967a, 1967b), respectively. Lower-case p is the pump parameter of the model, $\langle\Lambda\rangle$ is the effective linewidth of intensity fluctuations, and $n^{(2)}$ is the normalized second factorial moment $\langle n(n-1)\rangle/\langle n\rangle^2$ of the photocount distribution $p(n,T)$ as $T \to 0$. In this limit, $n^{(2)} - 1 = \langle(\Delta\rho)^2\rangle/\langle\rho\rangle^2$, and I/I_0 is the light intensity normalized to threshold value ($p = 0$); i.e., $I/I_0 = \langle\rho\rangle/\langle\rho\rangle_{th} = \langle\rho\rangle/1.128$. Here ρ is an abbreviation for $|\beta|^2$

a universal equation that we have called the rotating-wave van der Pol with noise:

$$\frac{d\beta}{dt} = (1 - i\alpha)(p - |\beta|^2)\beta + h(t). \quad (353)$$

The rescaling in amplitude was chosen to make the coefficient of $|\beta|^2$ unity. The scaling in time was chosen to yield the simple correlation function for the c-number noise source:

$$\langle h(t)h(u)\rangle = 4\delta(t - u), \quad (354)$$

ignoring, here, the coupling between reactive and resistive effects, by setting the parameter α equal to zero.

Away from threshold, Lax (1967) carried out a quasilinear treatment of self-sustained oscillators. Near threshold, fluctuations of the order of 100% are possible, and a numerical solution of the associated Fokker–Planck equation becomes necessary. Risken and Vollmer (1967) and Hempstead and Lax (1967) carried out such calculations to obtain the linewidth associated with field-amplitude fluctuations and the linewidth associated with intensity fluctuations. Comparisons of these theoretical predictions with subsequent experimental results are shown in Fig. 7 and Fig. 8, respectively.

Glossary

Brownian Motion: The limit of a random-walk process in which the individual steps become negligible. The position is continuous but not differentiable. The mean square displacement grows linearly with time. The particle density obeys the diffusion equation.

Characteristic Function: The Fourier transform of the probability density.

Chapman–Kolmogoroff Condition: A statement that conditional probabilities must obey if the process is to be Markoffian.

Fokker–Planck Process: A stochastic process that generalizes the Brownian process by having drift and diffusion coefficients that can be position dependent.

Fokker–Planck Equation: The partial differential equation, involving a single time derivative and up to two space derivatives, obeyed by a Fokker–Planck process. The independent (position) variables can be replaced by other variables – for example, spin angles in the description of spin diffusion.

Generalized Fokker–Planck Equation: A Fokker–Planck equation that may contain derivatives to all orders in the independent (position) variables.

Langevin Equation: An equation for a random variable in which the randomness is generated by a random (''force'') term added usually on the right-hand side. The noise source is typically delta correlated in time (white noise) but nondelta correlation (colored noise) is permitted.

Markoff Process: A random process in which the probability of future events depends on present information and no prior information.

Noise Spectrum: The power as a function of frequency obtained by passing the signal through a narrow filter, squaring the output, and averaging it.

Shot Noise: The input noise assumed to be associated with a series of Dirac delta-function pulses arriving at random at a uniform average rate. The term is also used to describe the response a physical system to such input white noise.

Wiener–Khinchin Theorem: Equality between the noise spectrum and the Fourier transform of the autocorrelation function.

List of Works Cited

Arecchi, F. T., Giglio, M., Sona, A. (1967a), ''Dynamics of the laser radiation at threshold,'' *Phys. Lett.* **A25**, 341–342.

Arecchi, F. T., Rodari, G. S., Sona, A. (1967b), ''Statistics of the laser radiation at threshold,'' *Phys. Lett.* **A25**, 59–60.

Bellman, R. (1968), *Some Vistas of Modern Mathematics; Dynamic Programming, Invariant Imbedding, and the Mathematical Biosciences*, Univ. of Kentucky Press.

Black, F., Scholes, M. (1973), ''The pricing of options and corporate liabilities,'' *J. Polit. Econ.* **81**, 637–654.

Bouchard, , Jean-Philippe, , Sornette, , Didier, (1994), ''The Black-Scholes option pricing problem in mathematical finance: generalization and extensions for a large class of stochastic processes,'' *J. Phys. I (Paris)* **4**, 863–881.

Brinkman, H. C. (1956), ''Brownian motion in a field of force and the diffusion theory of chemical reactions,'' *Physica* **22**, 149–155.

Brown, R. (1828), *Philos. Mag.* **4**, 161; *Ann. Phys. Chem.* **14**, 294.

Bulsara, A., Hänggi, P., Marchesoni, F., Moss, F., Shlesinger, M. (Eds.) (1993), Proceedings of the NATO Advanced Research Workshop, *J. Stat. Phys.* **70**, 1–512.

Callen, H. B., Welton, T. A. (1951), ''Irreversibility and generalized noise,'' *Phys. Rev.* **83**, 34–40.

Callen, Herbert B. (1985), *Thermodynamics*, New York: Wiley.

Campbell, N. (1909), ''The Study of discontinuous phenomena,'' *Proc. Camb. Philos. Soc.* **15**, 117–136.

Casti, J. L., Kalaba, R. E. (1973), *Imbedding Methods in Applied Mathematics*, Reading, MA: Addison-Wesley.

Davidson, F., Mandel, L. (1967), ''Correlation measurements of laser beam fluctuations near threshold,'' *Phys. Lett.* **A25**, 700–701.

Deutsch, , Ralph, (1962), *Nonlinear Transformations of Random Processes*, Englewood Cliffs, NJ: Prentice Hall.

Doob, J. L. (1953), *Stochastic Processes*, New York: Wiley, Chap. VI, Eqs. (3.1) and (3.4), and Chap. IX, Eq. (2.6).

Einstein, A. (1906), ''On the theory of the brownian movement,'' *Ann. Phys. (Leipzig)* **19**, 371–381.

Einstein, , Albert, (1905), ''On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat,'' *Ann. Phys. (Leipzig)* **17**, 549–560. See the English translation of this and four later, related articles in Einstein, A. (1956), *Investigations on the Theory of the Brownian Movement*, New York: Dover.

Fleming, M. W., Mooradian, A. (1981), ''Fundamental line broadening of single-mode (GaAl)As diode lasers,'' *Appl. Phys. Lett.* **38**, 511–513.

Fox, R. F. (1989), ''Stochastic resonance in a double well,'' *Phys. Rev. A* **39**, 4148–4153.

Gamo, H., Grace, R. E., Walter, T. J. (1968), ''Statistical analysis of intensity fluctuations in single mode laser radiation near the oscillation threshold,'' *IEEE J. Quantum Electron.* **QE-4**, 344.

Gang, H., Nicolis, G., Nicolis, C. (1990), ''Periodically forced Fokker–Planck equation and stochastic resonance,'' *Phys. Rev. A* **42**, 2030–2041.

Gerhardt, H., Welling, H., Güttner, A. (1972), ''Measurements of the laser line width due to quantum phase and quantum amplitude noise above and below threshold,'' *Z. Phys.* **253**, 113–126.

Ghashghaie, S., Breymann, W., Peinke, J., Talkner, P., Dodge, Y. (1996), ''Turbulent cascades in foreign exchange markets,'' *Nature* **381**, 767–770.

Gouy, M. (1888), *J. Phys. (Paris)* **7**, 561.

Gröbner, W., Hofreiter, N. (1950), *Integraltafel, Zweiter Teil, Bestimmte Integrale*, Wien: Springer.

Grossman, S., Richter, P. H. (1971), ''Laser threshold and nonlinear Landau fluctuation theory of phase transitions,'' *Z. Phys.* **242**, 458–475.

Hänggi, P., Talkner, P., Borkovec, M. (1990), ''Reaction-rate theory: fifty years after Kramers,'' *Rev. Mod. Phys.* **62**, 251–341.

Hänggi, , Peter, (1989), ''Colored noise in continuous dynamical systems: a functional calculus approach,'' in: F. Moss, P. V. E. McClintock (Eds.), Noise in Dynamical Nonlinear Dynamical Systems, Vol. 1, *Theory of continuous Fokker-Planck Systems*, Cambridge, U.K.: Cambridge Univ. Press.

Haken, H., Weidlich, W. (1967), ''A theorem on the calculation of multi-time-correlation functions by the single-time density matrix,'' *Z. Phys.* **205**, 96–102.

Haynes, J. R., Hornbeck, J. A. (1955), ''Trapping of minority carriers in Si.II.n type,'' *Phys. Rev.* **100**, 606–615.

Hempstead, R., Lax, M. (1967), ''Classical noise VI. Noise in self-sustained oscillators near threshold,'' *Phys. Rev.* **161**, 350–366.

Henry, C. H. (1986), ''Phase noise in semiconductor lasers,'' *IEEE J. Lightwave Technol.* **LT-4**, 298–311.

Hill, J. E., van Vliet, K. M. (1958), ''Ambipolar transport of carrier density fluctuations in germanium,'' *Physica* **24**, 709–720.

Hornbeck, J. A., Haynes, J. R. (1955), ''Trapping of minority carriers in Si.I.p type,'' *Phys. Rev.* **97**, 311–321.

Hull, , John, (1992), *Options, Futures and Other Derivative Securities*, Englewood Cliffs: Prentice Hill.

Ito, K. (1951), ''On stochastic differential equations,'' *Am. Math. Soc. Memoirs* No. 4, 1–51.

Ito, K., McKean, H. P. (1965), *Diffusion Processes and Their Sample Paths*, Grundlehren der mathematischen Wissenschaften, Vol. 125, Berlin: Springer.

Jeffries, , Harold, , Jeffreys, Bertha Swirles (1950), *Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge University Press, Sec 1.10.

Johnson, J. B. (1928), ''Thermal agitation of electricity in conductors,'' *Phys. Rev.* **32**, 97–109.

Jung, , Peter, (1993), ''Periodically driven stochastic systems,'' *Phys. Rept.* **234**, 175–295.

Kendall, M. G., Stuart, A. (1969), The Advanced Theory of Statistics, Vol. I, 3rd ed., New York: Hafner.

Khinchin, A. (1934), ''Korrelationstheorie der stationären stochastischen Prozesse,'' *Math. Ann.* **109**, 604–615.

Kittel, C. (1958), *Elementary Statistical Physics*, New York: Wiley.

Kramers, H. A. (1940), ''Brownian motion in a field of force and the diffusion model of chemical reactions,'' *Physica* **7**, 284–304.

Landauer, R. (1989), ''Noise-activated escape from metastable states: an historical review,'' in: F. Moss, P. V. E. McClintock (Eds.) Noise in Nonlinear Dynamical Systems, Vol. 1, *Theory of Continuous Fokker–Planck Systems*, Cambridge, U.K.: Cambridge Univ. Press.

Landauer, R., Swanson, J. A. (1961), *Phys. Rev.* **121**, 1668–1674.

Langer, J. (1968), ''Theory of nucleation rates,'' *Phys. Rev. Lett.* **21**, 973–976.

Lax, M., Zwanziger, M. (1973), ''Exact photocount statistics: lasers near threshold,'' *Phys. Rev. A* **7**, 750–771.

Lax, M. (1960), ''Fluctuations from the nonequilibrium steady state,'' *Rev. Mod. Phys.* **32**, 25–64.

Lax, M. (1963), ''Formal theory of quantum fluctuations from a driven state,'' *Phys. Rev.* **129**, 2342–2348.

Lax, M. (1965), ''Quantum noise V: phase noise in a homogeneously broadened maser,'' in: P. L. Kelley, B. Lax, P. E. Tannenwald (Eds.), *Physics of Quantum Electronics*, New York: McGraw-Hill.

Lax, M. (1966a), ''Classical noise III: nonlinear markoff processes,'' *Rev. Mod. Phys.* **38**, 359–379.

Lax, M. (1966b), ''Classical noise IV: Langevin methods,'' *Rev. Mod. Phys.* **38**, 541–566.

Lax, M. (1967), ''Classical noise V: noise in self-sustained oscillators,'' *Phys. Rev.* **160**, 290–307.

Lax, M. (1968a), ''Fluctuations and coherence phenomena in classical and quantum physics,'' in: M. Chretien, E. P. Gross, S. Deser (Eds.), *Statistical Physics*, 1966 Brandeis Summer Lecture Series, Vol. 2, New York: Gordon and Breach Science Publishers, pp. 270–478.

Lax, M. (1968b), ''Quantum noise XI. multitime correspondence between quantum and classical stochastic processes,'' *Phys. Rev.* **172**, 350–361.

Lax, M. (1991), ''The theory of laser noise,'' 1990 Conference on Laser Science and Optics Applications – Boston, MA, *Proc. SPIE* **1376**, 2–20.

Lax, M., Louisell, W. H. (1967), ''Quantum noise IX: quantum Fokker–Planck solution for laser noise,'' *J. Quantum Electronics* **QE-3**, 47–58.

Lax, M., Mengert, P. (1960), ''Influence of trapping, diffusion and recombination on carrier concentration fluctuations,'' *J. Phys. Chem. Solids* **14**, 248–267.

Lax, M., Phillips, J. C. (1958), ''One dimensional impurity bands,'' *Phys. Rev.* **110**, 41–49.

Lett, P., Mandel, L. (1985), ''Investigation of time-dependent correlation properties of the bidirectional dye ring laser,'' *J. Opt. Soc. Am. B* **2**, 1615–1625.

Linkowitz, S. (1989), Ph.D. Thesis.

Middleton, D. (1960), *Statistical Communications Theory*, New York: McGraw-Hill.

Moss, F., McClintock, P. V. E. (1989), Noise in Nonlinear Dynamical Systems, Vol. I, Theory of continuous Fokker–Planck Systems; Vol. II, Theory of Noise Induced Processes in Special Applications; Vol. III, *Experiments and Simulations*, Cambridge, U.K.: Cambridge Univ. Press.

Moullin, E. B. (1938), *Spontaneous Fluctuations of Voltage*, Oxford: Clarendon Press.
Moyal, J. E. (1949), "Quantum mechanics as a statistical theory," *Proc. Camb. Philos. Soc.* **45**, 99–124.
Nyquist, H. (1927), *Phys. Rev.* **29**, 614.
Nyquist, H. (1928), "Thermal agitation of electric charge in conductors," *Phys. Rev.* **32**, 110–113.
Onsager, L. (1931), "Reciprocal relations in irreversible processes," Part 1, *Phys. Rev.* **37**, 405–426; Part 2, **38**, 2265–2679.
Peters, , Edgar, E. (1994), *Fractal Market Analysis*, New York: Wiley.
Pontryagin, L., Andronov, A., Vitt, A. (1933), "On the statistical treatments of dynamical systems," *Zh. Eksp. Teor. Fiz.* **3**, 165–180. An English translation appears on pp. 329–348 of F. Moss, P. V. E. McClintock (Eds.) (1989), Noise in Nonlinear Dynamical Systems, Vol. 3, *Experiments and Simulations*, Cambridge, U.K.: Cambridge Univ. Press.
Presilla, C., Marchesoni, F., Gammaitoni, L. (1989), "Random Time-Modulated Bistable Systems: Non-stationary Statistical Properties," *Phys. Rev. A* **40**, 2105–2113.
Rice, S. O. (1944), "Mathematical analysis of random noise, part I," *Bell Syst. Technol. J.* **23**, 282–332.
Rice, S. O. (1945), "Mathematical analysis of random noise, part II," *Bell Syst. Technol. J.* **24**, 46–156.
Rice, S. O. (1948), "Mathematical analysis of random noise, part III," *Noise Through Nonlinear Devices* **27**, 109–157.
Risken, H., Vollmer, H. D. (1967), "Correlation function of the amplitude and of the intensity fluctuation near threshold," *Z. Phys.* **191**, 301–312.
Roy, R., Mandel, L. (1980), "Optical bistability and first order phase transition in a ring dye laser," *Opt. Commun.* **34**, 133–136.
Scher, H., Lax, M. (1973), "Stochastic transport in a disordered solid: I and II," *Phys. Rev. B* **7**, 4491–4502, 4502–4591.
Shockley, W. (1950), *Electrons and Holes in Semiconductors*, Princeton, NJ: Van Nostrand.
Stratonovich, R. L. (1963), Topics in the Theory of Random Noise, Vol. I, New York: Gordon and Breach. See also Vol. **II** (1967).
Stratonovich, R. L. (1989), "Some Markoff methods in the theory of stochastic processes in nonlinear dynamical systems," in: F. Moss, P. V. E. McClintock (Eds.), Noise in Nonlinear Dynamical Systems, Vol. I, Cambridge, U.K.: Cambridge Univ. Press.
Talkner, P., Hänggi, P. (1995), *New Trends in Kramers' Reaction Rate Theory*, Dordrecht/Boston/London: Kluwer Academic Publishers.
Thiele, T. N. (1903), "Theory of observations," *Ann. Math. Stat.* **2**, 165.
Transistor Teachers' Summer School (1953), "Experimental verification of the relation between diffusion constant and mobility of electrons and holes," *Phys. Rev.* **88**, 1368–1369.
Uhlenbeck, G. E., Ornstein, L. S. (1930), "On the theory of the Brownian motion," *Phys. Rev.* **36**, 823–841.
Van Kampen, N. (1985), "Elimination of fast variables," *Phys. Rep.* **124**, 69–160.
Van Kampen, N. (1992), *Stochastic Processes in Physics and Chemistry*, Amsterdam: North Holland.
Van Roosbroeck, W. (1953), "Transport of added current carriers in a homogeneous semiconductors," *Phys. Rev.* **91**, 282–289.
Wiener, N. (1930), "Generalized harmonic analysis," *Acta Math.* **55**, 117–258.
Wigner, E. (1932), "Quantum corrections for thermodynamic equilibrium," *Phys. Rev.* **40**, 749–760.
Williams, E. C. (1937), "Thermal fluctuations in complex networks," *J. Electrical Eng.* **81**, 751.

Further Reading

Bell, D. A. (1960), *Electrical Noise*, Princeton: Van Nostrand.
Burgess, R. E. (1965), *Fluctuation Phenomena in Solids*, New York: Academic.
Chandrasekhar, S. (1943), "Stochastic problems in physics and astronomy," *Rev. Mod. Phys.* **15**, 1–89.
Feller, , William, (1971), An Introduction to Probability Theory and Its Applications, Vol. I, New York: Wiley.
Gillespie, Daniel T. (1996), "The mathematics of Brownian motion and Johnson noise," *Am. J. Phys.* **64**, 225–240.
Hamming, Richard W. (1991), *The Art of Probability for Scientists and Engineers*, Reading, MA: Addison Wesley.
Lawson, J. L., Uhlenbeck, G. E. (1950), *Threshold Signals*, M.I.T. Radiation Lab Series, Vol. 24, New York: McGraw Hill, Dover.

Lax, M. (1964), "Quantum relaxation, the shape of lattice absorption and inelastic neutron scattering lines," *J. Phys. Chem. Solids* **25**, 467–503.

Lax, M. (1966), "Quantum Noise IV: Quantum Theory of Noise Sources," *Phys. Rev.* **145**, 110–129.

Lax, M. (1974), *Symmetry Principles in Solid State and Molecular Physics*, New York: Wiley.

Louisell, W. H. (1973), *Quantum Statistical Properties of Radiation*, New York: Wiley, Appendix A.

Lukacs, E. (1960), *Characteristic Functions*, London: Griffith.

MacDonald, D. K. C. (1962), *Noise and Fluctuations*, New York: Wiley.

Middleton, , David, (1960), *Introduction to Statistical Communication Theory*, New York: McGraw-Hill.

Moullin, E. B. (1938), *Spontaneous Fluctuations of Voltage*, Oxford: Clarendon Press.

Robinson, F. N. H. (1962), *Noise in Electrical Circuits*, Oxford: Clarendon.

Robinson, F. N. H. (1974), *Noise and Fluctuations in Electronic Devices and Circuits*, Oxford: Clarendon Press.

Valley, George E., Jr., Wallman, , Henry, (1948), *Vacuum Tube Amplifiers*, M.I.T. Radiation Lab Series, Vol. 13, New York: McGraw-Hill.

Wang, Ming Chen, Uhlenbeck, G. E. (1945), "On the theory of Brownian motion II," *Rev. Modern Phys.* **17**, 323–342.

Symmetries and Conservation Laws

Gino Segrè
Department of Physics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Introduction *565*
1 Symmetries of Rigid Bodies *566*
2 Symmetries of Crystals *567*
3 Symmetries and the Lorentz Group *568*
4 Conservation Laws and Noether's Theorem *570*
5 Discrete Symmetries: T, C and P *572*
6 Internal Symmetries *573*
7 Broken Symmetries and Goldstone Bosons *575*
8 Gauge Symmetries *578*
9 The Higgs Mechanism *581*
10 Experimental Limits on Symmetry Breaking *582*
Glossary *584*
List of Works Cited *585*
Further Reading *586*

Introduction

The topic of symmetry is a vast one. Books have been written about each one of the topics in the table of contents, and so our treatment perforce covers only some of the highlights.

Simple notions of symmetry are present in many considerations, but the central role of symmetry was not realized until the 20th century. In fact Wigner repeatedly argued that progress in physics was largely based on the ability to separate a physical problem into analyses of the laws of nature independently of the initial conditions. The latter are arbitrary and complicated while the former are general. Symmetries aid in formulating

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

the regularities in physical laws without reference to the specific dynamics. To quote Gross (1995) in his tribute to Wigner, ''Wigner argued that invariance principles provide a structure and coherence to the laws of nature just as the laws of nature provide a structure and coherence to a set of events''.

In this article we attempt to provide a birds-eye view of the application of symmetry considerations to problems in atomic and molecular physics, condensed matter physics, particle physics, and field theory. The article is weighted to the last because conservation laws appear naturally as a consequence of continuous symmetries of the Lagrangian and are seen most clearly in this context. The brief description of symmetry in other phenomena is intended as an attempt at completeness.

The central mathematical tool in a discussion of symmetry is group theory. This was also realized by Wigner (1959) soon after the development of quantum mechanics. Much of the advance of physics has been based on the employment of techniques it provides – point groups, Lie groups, or more exotic creations.

1 Symmetries of Rigid Bodies

The symmetry transformations of a finite system such as an atom or a molecule consist of the set of transformations that leave the body unchanged. All such transformations are combinations of rotations and reflections. Rotations are through definite angles about specified axes and reflections are through a plane.

The usefulness of such an analysis is considerable. Not only does the symmetry restrict the form of the allowed interactions, but it also provides a simple method for finding selection rules of transition-matrix elements.

One must specify the order in which one performs operations since, in general, two operations do not commute: that is, performing a given rotation and then a reflection generally gives a different result from that of performing these operations in the reverse order.

A body is said to have an axis of symmetry of nth order if it is left unaltered by a rotation of $2\pi/n$ about a given axis. We denote this operation symbolically by R_n. Repeating this operation n times gives us back the identity transformation, which we denote by E, and so we write

$$R_n{}^n = E. \tag{1}$$

Symmetry transformations of finite bodies must leave at leave at least one point invariant. Symmetry groups having this property are known as *point groups*. Landau and Lifshitz (1981) give the description of these groups. They are classified as C_n, S_{2n}, C_{nh}, C_{nv}, D_n, D_{nh}, D_{nd}, the tetrahedron group, the octahedron group and the icosahedron group. The simplest example, C_n, has already been described. C_{nh} and C_{nv} are obtained by adding to the axes in C_n planes of symmetry that respectively are perpendicular to or pass through a symmetry axis. D_n has two symmetry axes, one of nth order and a second one perpendicular to the original axis. D_{nh} and D_{nd} are defined by adding planes of symmetry in addition to the axes of symmetry. S_{2n} is a group with $2n$ elements defined by symmetry transformations about axes defined by combinations of rotations and reflections. Two continuous groups are also possible symmetry groups, namely the full rotation group and the group of rotations about a fixed axis. It is clear that a spherically symmetric body's appearance is unchanged

by an arbitrary rotation and a cylindrically symmetric one's by a rotation about the axis of symmetry.

The analysis of the matrix elements of operators between atomic or molecular states is enormously simplified through the classification of the states by their transformation properties under the rotation group – e.g., for atoms the *S, P, D,* etc. states. To pick a more sophisticated example consider the ammonia molecule, NH_3, which one may picture in terms of the plane defined by the location of the three hydrogen atoms with the nitrogen atom free to oscillate through the plane. The group of transformations that leave the molecule in its equilibrium state invariant is C_{3v}. The analysis proceeds by studying the group properties of C_{3v} on the space of the coordinates of the four atoms. For an excellent pedagogical review of this case see Lax's book (Lax, 1974).

2 Symmetries of Crystals

The symmetries of atoms and molecules are described by rotations and reflections. At least one point in the body must be left unchanged by the combination of transformations; e.g., a single body cannot be transformed into itself by successive rotations about two nonintersecting axes.

For infinite crystals, there is another type of symmetry transformation, namely translations. This is a vast subject, of which we can do no more than present a few highlights. For good introductions to the subject see the text by Ashcroft and Mermin (1976).

As we already said, symmetry operations are combinations of rotations, reflections, and translations. Rotations about a given axis combined with a translation perpendicular to the axis do not give a new symmetry; they are simply equivalent to the original rotation, but now about an axis parallel to the first one. Similarly reflections through a plane combined with displacements perpendicular to the plane do not give new symmetries.

On the other hand, if we combine a rotation about an axis with a displacement along the same axis we do obtain a new type of symmetry. The axis is known as a screw axis. In an n-fold screw, the system repeats itself after a rotation by $2\pi/n$ and a displacement by some distance d.

Similarly reflection through a plane combined with a displacement parallel to the plane leads to a so-called glide reflection plane symmetry, in which the lattice has symmetry through the simultaneous operations of reflection and translation.

The structure of a crystalline lattice is described most easily making use of the concept of the Bravais lattice, which specifies the periodic ordering of the units of the lattice. These units may be atoms, molecules or other entities. All we are interested in now is their location. The Bravais lattice is defined as the set of all position vectors $\boldsymbol{R}$ such that

$$\boldsymbol{R} = n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3 \qquad (2)$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are three vectors not all lying in the same plane and n_1, n_2, n_3 are allowed to range through all the values of the integers. The vectors $\boldsymbol{a}_i$ are called primitive vectors and are said to span the lattice. In general there are many inequivalent choices of primitive vectors for any given Bravais lattice.

The primitive unit cell is defined as that volume of space which, when translated by the allowed vectors of the Bravais lattice, fills all of space without leaving any voids. For a cubic lattice, e.g., it is simply a cube.

Once again there is no unique choice of the primitive cell.

The symmetries of a crystal are usually classified by the simpler set of symmetries of the underlying Bravais lattice. The set of operations that take the Bravais lattice into itself is known as the symmetry group or space group of the Bravais lattice. The space group can be shown to consist of

1. translations through Bravais lattice vectors,
2. operations, i.e., rotations and reflections, that leave a particular point of the lattice fixed, and
3. combinations of the first two.

The subset of the symmetry group that leaves a particular point in the Bravais lattice fixed is known as the point group of the Bravais lattice. There are only seven distinct point groups and hence seven crystal systems. Relaxing the restriction to point groups, one finds 14 distinct symmetry groups and hence 14 independent Bravais lattices. The symmetry group of a crystal structure is still larger, depending on both the symmetry of the object and the symmetry of the underlying Bravais lattice. For crystals of arbitrary shape based on arbitrary Bravais lattices, there are 230 possible space groups. For a discussion of these we refer the reader to the text by Ashcroft and Mermin (1976).

A fundamental concept to which one is led in the analysis of periodic structures is that of the reciprocal lattice. In particular, it is the key tool for systematizing discussions of diffraction scattering. Consider a plane wave $e^{i\boldsymbol{K}\cdot\boldsymbol{R}}$. In general the plane wave will not have the symmetries of the Bravais lattice; i.e., translation through a vector $\boldsymbol{R}$ will not leave the plane wave unchanged; but there is a set of vectors $\boldsymbol{K}$ for which the wave is unchanged by translations through $\boldsymbol{R}$. They satisfy

$$e^{i\boldsymbol{K}\cdot\boldsymbol{R}} = 1. \tag{3}$$

The vectors $\boldsymbol{K}$ form what is called the reciprocal lattice. It is straightforward to show that the reciprocal lattice is itself a Bravais lattice. The primitive cell of the reciprocal lattice is known as the first Brillouin zone. The great usefulness of the reciprocal lattice lies in the fact that, because of the above equation, when rays or neutrons are scattered from a crystal, the scattering is coherent if the momentum transfer is $h\boldsymbol{K}$, with $\boldsymbol{K}$ a reciprocal lattice vector and h equal to Planck's constant. The pattern of bright spots due to the coherent scattering allows us to reconstruct the crystal structure.

An interesting development in crystal symmetry of the past decade is the discovery of *quasicrystals*, an intermediate between crystal structures and glassy structures. The former, as we have already seen, have

1. long-range translational order and
2. long-range order under rotations and reflections.

A glassy structure, by contrast, has none of the above. Quasicrystals have long-range translational and rotational order. However, the translational order is of a subtler kind: it is quasiperiodic rather than periodic. Quasicrystals also may have rotational symmetries that are not allowed for ordinary crystals. Their group-theoretical analysis leads to some interesting new properties (Steinhardt and Ostlund, 1987).

3 Symmetries and the Lorentz Group

We now turn to the main subject of this review, namely symmetries and conservation laws. In quantum mechanics an

operator A that has no explicit time dependence is said to be conserved if its commutator with the Hamiltonian of the system, H, vanishes:

$$[H, A] = 0. \tag{4}$$

A state is represented by a normalized vector Ψ in Hilbert space, where Hilbert space is a complex vector space (Weinberg, 1995). Observables are represented by Hermitian operators A, linear mappings $\Psi \to A\Psi$ of the Hilbert space into itself. The probability of a measurement yielding the result that a state Ψ is in one of a set of states represented by mutually orthogonal vectors Ψ_n is given by

$$P = (\Psi, \Psi_n)^2. \tag{5}$$

A symmetry operation is one that does not change the results of possible experiments: i.e., if Ψ and Ψ' are two states related to one another by a symmetry transformation, then

$$(\Psi, \Psi_n)^2 = (\Psi', \Psi'_n)^2. \tag{6}$$

What we are saying is that if an observer O makes a measurement and obtains a result an observer O' related to O by a symmetry transformation will obtain the same result.

In Wigner's (1959) pioneering work in the 1930s, he showed that to satisfy the above equation Ψ had to be related to Ψ' by either a unitary or an antiunitary transformation U. Both take the form

$$U^\dagger = U^{-1} \tag{7}$$

Unitary transformations can be discrete or continuous. We treat discrete symmetry transformations in Sec. 5 and here concentrate on continuous ones. The transformation $U = 1$ is of course just the identity. Continuous transformations that differ infinitesimally from the identity are represented by

$$U = 1 + i\varepsilon t \tag{8}$$

where ε is an infinitesimal real parameter and t is an operator. The fact that U is unitary means that t is Hermitian and therefore a physical observable.

Symmetry transformations forming a group are of particular interest. A group has multiplication and an inverse defined for each element. A set of symmetry operations forming a group corresponds to a set of unitary transformations $U(T)$ acting on the vectors Ψ of the Hilbert space.

Of particular interest in physics are the connected Lie groups. They are continuous groups; any element of the group can be connected to the identity by a path that lies within the group space. This allows us to describe the group elements by infinitesimal transformations, because any finite transformation can be built up by an infinite sequence of infinitesimal transformations. We therefore write

$$U(T(\theta)) = 1 + i\theta_a t_a + \cdots \tag{9}$$

The t_a are the generators of infinitesimal transformations, repeated indices are assumed to be summed over, and we take $a = 1, 2, \ldots, n$. The condition of group multiplication requires the t_a to satisfy a set of consistency conditions, which can be expressed as commutation relations:

$$[t_a, t_b] = iC_{abc}t_c \tag{10}$$

with C_{abc} a set of numbers, the so-called structure constants of the group. This set of commutation relations form what is called a Lie algebra and characterize the group. If all the commutators vanish, or equivalently all the structure constants are zero, we

have what is called an Abelian Lie group. In this case the finite transformations can be trivially built up out of infinitesimal ones and we obtain

$$U(T(\theta)) = \exp(i\theta_a t_a). \tag{11}$$

In general, Lie groups are non-Abelian.

The most important symmetry operations are the rotations, the translations, and the Lorentz boosts. To these we append the discrete reflection transformations. These operations form the so called Poincaré group or inhomogeneous Lorentz group (the homogeneous Lorentz group is the latter minus translations). The coordinate transformation corresponding to such an operation is

$$x'_\nu = \Lambda^\mu_\nu x_\mu + a_\nu. \tag{12}$$

In the above a_ν is the four-vector corresponding to translations and Λ^μ_ν the tensor corresponding to rotations and boosts. The equivalence of inertial frames dictates that the latter must satisfy

$$\eta_{\mu\tau}\Lambda^\mu_\sigma \Lambda^\tau_\rho = \eta_{\sigma\rho} \tag{13}$$

where $\eta_{\mu\nu}$ is the diagonal matrix

$$\eta_{11} = \eta_{22} = \eta_{33} = 1, \ \eta_{00} = -1. \tag{14}$$

For the identity transformation, we simply have Λ^ν_μ equal to the Kronecker delta. For infinitesimal transformations we have

$$\Lambda^\mu_\nu = \delta^\mu_\nu + \omega^\mu_\nu \tag{15}$$

and

$$U(\omega, \varepsilon) = 1 + \frac{1}{2} i\omega_{\mu\nu} J^{\mu\nu} - i\varepsilon_\mu P^\mu. \tag{16}$$

The four components of P_μ, the generator of infinitesimal translations, are respectively P_0, the Hamiltonian or energy of the system, and P_i, the three-vector momentum (the latin indices take on values 1, 2, 3 as opposed to greek indices which take values 0, 1, 2, 3). $J_{\mu\nu}$ is an antisymmetric tensor with six independent components: J_{23}, J_{31}, J_{12} are the generators of infinitesimal rotations, labelled as $J_i = \varepsilon_{ijk} J_{jk}$. The J_{i0} are the generators of Lorentz boosts, K_i.

The commutators, which specify the algebra of the group, are

$$[J_i, J_j] = -[K_i, K_j] = i\varepsilon_{ijk} J_j, \tag{17}$$

$$[J_i, K_j] = i\varepsilon_{ijk} K_k, \tag{18}$$

$$[J_i, P_j] = i\varepsilon_{ijk} P_k, \tag{19}$$

$$[K_i, P_j] = iH\delta_{ij}, \tag{20}$$

$$[K_i, H] = iP_i, \tag{21}$$

$$[J_i, H] = [P_i, H] = [H, H] = 0. \tag{22}$$

The above imply that $\boldsymbol{J}$, $\boldsymbol{K}$, $\boldsymbol{P}$, whose Cartesian components are J_i, K_i, P_i, are all vectors under ordinary spatial rotations and that J_i, P_i, H are all conserved quantities, i.e., constants in time. To rephrase these results in a more intuitive manner, note that if H is invariant under a continuous symmetry U, we have $UHU^{-1} = H$. This implies that the generator of infinitesimal transformations commutes with the Hamiltonian. The last of the array of equations above simply restates then that the Hamiltonian is invariant under rotations and space-time translations.

4 Conservation Laws and Noether's Theorem

Classical mechanics can be formulated by the principle of least action. This states that for all paths that go from a fixed q_1 at time t_1 to a fixed q_2 at time t_2, the physical trajectory corresponds to a stationary value of the action

$$I = \int_{t_1}^{t_2} \mathrm{d}t \, L(q(t), q(t)). \tag{23}$$

where $L(q, \dot{q})$ is the Lagrangian of the system. The stationarity of the action implies in turn Lagrange's equations

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}} = 0. \tag{24}$$

For the simplest cases L is the difference between the kinetic and the potential energy of the system and is unchanged by the addition of a total time derivative.

An alternative formulation of mechanics makes use of the Hamiltonian, which is simply related to the Lagrangian by

$$H = p\dot{q} - L, \tag{25}$$

with p the momentum conjugate to q, defined as

$$p = \frac{\partial L}{\partial \dot{q}}. \tag{26}$$

Hamilton's equations can be shown to follow from the principle of least action (Itzykson and Zuber, 1980). For a system described by N independent coordinates and hence $2N$ independent variables, we extend the principle straightforwardly and obtain N independent Lagrange's equations.

Let us now consider continuous infinitesimal changes of coordinates $q_i \to q_i + \varepsilon \Delta q_i$ where ε is an infinitesimal parameter and Δq_i is a deformation of q_i. Such transformations are symmetry operations if they leave the equations of motion unchanged. This is ensured if L is invariant or, more generally, only changes by a total time derivative (in both cases the action is unchanged):

$$\begin{aligned}\Delta L &= \frac{\partial L}{\partial q_i}\Delta q_i + \frac{\partial L}{\partial q_i}\Delta q_i + \frac{\mathrm{d}C}{\mathrm{d}t} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial L}{\partial q_i}\Delta q_i + C\right] \\ &\quad + \Delta q_i\left[\frac{\partial L}{\partial q_i} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial q_i}\right].\end{aligned} \tag{27}$$

Repeated indices (in this case $i = 1, 2, \ldots, N$) are assumed to be summed over. The second term in the above equation vanishes because of Lagrange's equations. We see that a continuous symmetry of L, which implies $\Delta L = 0$, leads to the existence of a conserved quantity, namely a quantity whose time derivative is zero,

$$Q = \frac{\partial L}{\partial q_i}\Delta q_i + C = p_i \Delta q_i + C. \tag{28}$$

This is Noether's theorem.

For systems with an infinite number of degrees of freedom, we introduce the concept of a field $\phi(x)$ where x is a continuous variable denoting time and spatial locations. Again this is readily generalizable to N independent fields. We introduce a Lagrangian density $\mathscr{L}$ and the action is now

$$I = \int \mathrm{d}^4x \mathscr{L}(x). \tag{29}$$

The Lagrange equations for the fields take the form

$$\frac{\partial \mathscr{L}}{\partial \phi(x)} - \partial_\mu \frac{\partial \mathscr{L}}{\partial[\partial_\mu(\phi(x))]} = 0. \tag{30}$$

The least-action principle once again leads to conservation laws. Consider, for instance, the effect of translation of the coordinates by $x \to x + a$. The change in the Lagrange density is

$$\mathscr{L}[x + a] = \mathscr{L}[\phi(x + a), \partial_\mu \phi(x + a)]. \tag{31}$$

For an infinitesimal displacement, the variation of the action is given by

$$\begin{aligned}\delta I = \int \mathrm{d}^4x \Bigg[&\partial_\nu \mathscr{L} \\ &-\partial_\mu\left[\frac{\partial \mathscr{L}}{\partial(\partial_\mu \phi)}\partial_\nu \phi(x)\right]\Bigg]\delta a^\nu = 0.\end{aligned} \tag{32}$$

From the vanishing of δI for arbitrary δa^μ we deduce the existence of a tensor $T^{\mu\nu}$ satisfying the conservation law

$$\partial_\mu T^{\mu\nu} = 0, \tag{33}$$

where the tensor, known as the energy-momentum tensor, is defined as

$$T^{\mu\nu} = \frac{\partial \mathscr{L}}{\partial(\partial_\mu \phi)} \partial_\nu \phi - \eta_{\mu\nu} \mathscr{L}. \tag{34}$$

The $\eta_{\mu\nu}\mathscr{L}$ term in the above equation is the analogue of the $\mathrm{d}C/\mathrm{d}t$ term defined earlier. From the form of the Lagrangian density, it usually follows that the energy-momentum tensor is explicitly symmetric in μ and ν. If this is not the case, one defines a modified energy-momentum tensor, which has the desired symmetry. Since $T^{\mu\nu}$ is conserved, we can also define four conserved charges

$$P^\nu = \int \mathrm{d}^3x T^{0\nu}, \tag{35}$$

which are time independent if the fields vanish sufficiently fast at spatial infinity so that

$$\int \mathrm{d}^3x \partial_i T^{i\nu} = 0. \tag{36}$$

As before latin indices $i = 1, 2, 3$ refer only to spatial components. This implies that

$$\partial_0 P^\nu = \int \mathrm{d}^3x \partial_0 T^{0\nu} = \int \mathrm{d}^3x \partial_\mu T^{\mu\nu} = 0. \tag{37}$$

The conserved quantities P^ν we have just defined are of course the four components of a four-vector and are simply the energy and momentum of the system as described in the previous section. They are also the generators of infinitesimal time and space translations.

So far we have discussed space-time symmetries. As we shall see shortly, Noether's theorem is also useful for internal symmetries.

5 Discrete Symmetries: *T*, *C* and *P*

Discrete symmetries play an important role in field theory. Space reflection corresponds to the improper Lorentz transformations, which reverse the signs of coordinates leaving time unchanged – i.e., under space reflection

$$\boldsymbol{x}' = -\boldsymbol{x}, \ t' = t; \tag{38}$$

and, under time reflection, $t' = -t$ while spatial coordinates are unchanged. Under a parity transformation scalars and axial vectors are invariant while vectors change sign. A pseudoscalar, which may be thought of as the invariant product of a vector and an axial vector, is of course an invariant under rotations, but changes sign under parity transformations.

Charge conjugation C relates particles with equal masses and spins, but opposite charges, i.e., it relates particles to antiparticles. As an example, it changes electrons into positrons, protons into antiprotons, and π^+ into π^-. Some particles are their own antiparticles. The photon is one such example: charge-conjugation invariance of the theory implies that the photon changes under C into minus itself since electrons change into positrons; these couple to photons as electrons do, but with the opposite sign.

In a second-quantized formalism the Lagrangian is written in terms of fields, which are themselves linear combinations of creation and annihilation operators. For a charged scalar field $a(k)$ and $b(k)$ are the annihilation operators for quanta of opposite charges, so that, e.g., $a(k)$ annihilates quanta of charge $+1$ while $b^\dagger(k)$ creates quanta of charge -1. In both cases the charge of the system is increased by the action of the operator. One of the types of particles associated

with the quanta is called particle and the other antiparticle. As mentioned, some particles, such as the photon, may be their own antiparticles.

There exists a unitary operator $\mathscr{C}$ which takes the fields for particles into those for antiparticles. This is the field-theoretical version of charge conjugation. For instance, the photon field $A_\mu(x)$ transforms into minus itself, i.e.,

$$\mathscr{C} A_\mu(x) \mathscr{C}^\dagger = -A_\mu(x). \tag{39}$$

For the charged scalar field

$$\mathscr{C} a(k) \mathscr{C}^\dagger = b(k). \tag{40}$$

In a classic series of experiments in the 1950s, following the suggestion of Lee and Yang (1956), conservation of parity was found to be maximally violated. It was believed at the time that *CP*, the combined operation of parity and charge reflection, was a good symmetry, but this was disproved by the observation in 1964 of the long-lived kaon decaying into two pions (Christenson et al., 1964).

It is still believed, however, that though *T, C* and *P* are individually not conserved, invariance under the combined operation of *TCP* holds. The validity of this belief, derived from the existence of local relativistic quantum field theory (Streater and Wightman, 1968), ensures the equality of masses and lifetimes of particles and antiparticles. We discuss tests of these invariances in the last section of the article.

6
Internal Symmetries

We have so far considered symmetries associated with Lorentz transformations. There are also many internal symmetries of systems, generally associated with new quantum numbers. An example is isospin, the symmetry that relates protons to neutrons. By analogy with an ordinary spin-$\frac{1}{2}$ field for which the $S_3 = \pm\frac{1}{2}$ components are related by spatial rotations, we imagine the proton and neutron to be an isospin $I_3 = \pm\frac{1}{2}$ doublet.

The internal symmetries we consider in this section are global symmetries as distinguished from local symmetries; in the former the transformations are independent of space-time, i.e., all protons in a nucleus are rotated by the same amount when we apply an isospin rotation to the state of a nucleus. For local or gauge symmetries, on the other hand, the transformation itself is a function of space-time, i.e., protons in different locations would be rotated by different amounts.

There is another important difference between global and local symmetries: the former are generally broken; in fact many theorists believe there are no exact global symmetries. On the other hand local or gauge symmetries are believed to be conserved in the absence of spontaneous symmetry breaking (this occurs when the Lagrangian is symmetric, but the ground state, or vacuum, of the system is not symmetric). Local symmetries are discussed in Sec. 8.

Taking for granted that the symmetries (from now on in this section we refer to global symmetries) are broken, they may nevertheless be very useful to study, particularly if the breaking is small. Isospin symmetry is a good example; the strong interactions that bind nucleons in a nucleus are believed to be approximately symmetric under an isospin rotation. Electromagnetism obviously distinguishes between protons and neutrons since the former have electric charge and the latter do not. On the other hand electromagnetism usually leads to only

small perturbations in nuclei and, after all, the proton and neutron masses only differ by one part in two thousand; isospin is a useful symmetry for nuclei.

In the early 1960s the concept of isospin was extended by Gell-Mann and Neeman (Cheng and Li, 1984) to unitary symmetry, or SU(3), with spectacular success. In SU(3) strangeness, an additional quantum number of hadrons that is conserved in the strong interactions was unified with isospin. The baryons, the pseudoscalar mesons, and vector mesons were found to lie in octet representations, and the baryon resonances in a decuplet representation; and a variety of predictions regarding masses, mixings and decays were found to be true with remarkable precision. Once again symmetry proved a powerful tool.

For the above reasons it is useful to begin by treating these symmetries as if they were exact and look at Noether's theorem (see Sec. 4). We also saw in Sec. 3 how continuous Lie groups were defined. Let us begin to apply some of these notions.

Consider a set of N real scalar fields $\phi_1 \cdots \phi_N$, which we will label as ϕ_i with i running from 1 to N. We assume that the fields belong to a representation of a group G; that is, the fields may be thought of belonging to an N-dimensional vector space acted on by the rotations in G. Calling the Lagrangian density $\mathscr{L}$ as before and assuming that it is a function of ϕ_i and its derivatives, we consider variations of the fields

$$\delta\phi_i(x) = i\varepsilon_a(t_a)_{ij}\phi_j(x) \tag{41}$$

where ε_a is taken to be an infinitesimal quantity assumed independent of space and time. As before, repeated indices are summed over.

If the Lagrangian density is invariant under the rotation in question we have, using Lagrange's equations,

$$\delta\mathscr{L} = \partial_\mu \frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}\delta\phi_i + \frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}\partial_\mu(\delta\phi_i), \tag{42}$$

which, on substituting the value for $\delta\phi_i$, implies

$$\delta\mathscr{L} = \varepsilon_a\partial_\mu\left[\frac{\delta\mathscr{L}}{\delta(\partial_\mu\phi_i)}i(t_a)_{ij}\phi_j\right]. \tag{43}$$

Since the Lagrangian is invariant and the above vanishes for any ε_a, we see that the quantity in the square brackets is a conserved current

$$J_{\mu,a} = -i\frac{\delta\mathscr{L}}{\delta(\partial^\mu\phi_i)}(t_a)_{ij}\phi_j. \tag{44}$$

The charges, defined by

$$Q_a = \int \mathrm{d}^3x J_{0,a}(x), \tag{45}$$

can also be written, using the definition of conjugate momentum,

$$Q_a = -i\int \mathrm{d}^3x \pi_i(t_a)_{ij}\phi_j. \tag{46}$$

Using the equal-time commutation relations between fields and their conjugate momenta, one can easily prove that the charges satisfy in turn the same commutation relations as the generators of the group G, *i.e.*,

$$[Q_a(t), Q_b(t)] = iC_{abc}Q_c(t). \tag{47}$$

The charges are all evaluated at the same time since the commutation relations of the fields and their conjugate momenta are specified at equal times. These relations are called, for obvious reasons, the algebra of charges.

Interestingly enough, since the commutation relations of fields and conjugate momenta hold in any case, these relations hold even if the symmetry is broken. This

observation was the starting point for the major development of current algebra in the 1960s (Adler and Dashen, 1968).

This subject is a broad one in itself, but at its roots there is the important identification of the currents associated with the generators of the symmetry with the currents observed in the weak interactions. By this we mean the following: the form of the weak interactions, originally proposed by Fermi in 1934, with suitable modifications, was seen to be an adequate phenomenological description of the weak interactions. It states that the effective Hamiltonian is

$$H_{weak} = G_F J_\mu J^\mu. \tag{48}$$

where $J_\mu(x)$ is a charged current of the $V - A$ form, namely with equal vector and axial vector parts (Commins and Bucksbaum, 1983).

The initial step was to identify the vector part of this current with the current of the isospin-raising generator; this has many important physical consequences. It not only dictates that the weak current must connect any two particles that lie in an isospin multiplet and differ by one unit of charge (e.g., π^+, π^0), but also specifies the relative normalization of the matrix elements.

This hypothesis, the so called conserved vector-current hypothesis, was confirmed experimentally and extended to an interpretation of the strangeness-changing currents as the currents in the generators of the strangeness operators in the algebra of SU(3). It also became natural then to look for a group-theoretical or symmetry counterpart of the axial currents in the weak interactions.

The algebra of SU(2), with charges Q_a, was extended to an algebra of SU(2) × SU(2) by the inclusion of pseudoscalar charges Q_a^5 with commutation relations

$$[Q_a^5, Q_b^5] = i\varepsilon_{abc} Q_c \tag{49}$$

$$[Q_a, Q_b^5] = i\varepsilon_{abc} Q_c^5, \tag{50}$$

$$[Q_a, Q_b] = i\varepsilon_{abc} Q_c. \tag{51}$$

This of course can then be generalized to SU(3) × SU(3). This is the historical path. The modern point of view relies heavily on these notions, but incorporates in a central way the notion of spontaneous symmetry breaking, which we turn to next.

7
Broken Symmetries and Goldstone Bosons

Symmetries of the Lagrangian may be broken explicitly by the introduction of non-invariant terms. More subtly, the Lagrangian may be invariant, but not the physical vacuum or ground state. We distinguish the physical vacuum from what we call the bare vacuum: the former is the true ground state of the system while the latter is the state with no particles or excitations. The physical or true vacuum and the bare vacuum may not coincide, as is the case in, e.g., superconductivity, where the true vacuum contains a superposition of Cooper pairs.

If a continuous symmetry of the Lagrangian is not a symmetry of the physical vacuum, we say that the symmetry is spontaneously broken. To refer back to the case of superconductivity, the Lagrangian has a U(I) symmetry associated with gauge transformations of electromagnetism. The vacuum, on the other hand, does not possess this symmetry since it contains electron pairs.

In general the spontaneous breaking of a continuous symmetry leads to the existence of massless scalar bosons, usually

called Nambu–Goldstone bosons or simply Goldstone bosons (Weinberg, 1995). To see how such a phenomenon arises we can begin by observing that a global rotation of the vacuum (of course here and in what follows we mean the physical vacuum or ground state) costs no energy, or, in other words, vacua differing only by a global rotation are degenerate in energy. Vacua differing by a local rotation differ in energy by the shear energy associated with the deformation of the vacuum. The energy of the shear goes to zero as the wave number goes to zero. In a quantum field theory, there exist states, connected to the physical vacuum by these local rotations, that are degenerate in energy and are populated by quanta whose energy goes to zero as their momentum goes to zero. They are therefore zero-mass bosons, referred to usually as Goldstone bosons. An important caveat, which we discuss in more detail later, is that these massless quanta are not observable in all cases. The number of types or species of such bosons is related to the degree of reduction of the symmetry, i.e., how much less symmetry the vacuum has than the Lagrangian. An explicit field-theoretical example may be helpful.

Consider a theory of scalar fields invariant under the three-dimensional rotation group O(3), now, however, an internal symmetry group. Let the scalars be in the vector, or three-dimensional, representation of the group. The scalars are described by fields $\phi_a(x)$ with $a = 1, 2, 3$, and an effective potential

$$V(\phi) = -\mu^2 \frac{\phi^2}{2} + \lambda \frac{\phi^4}{4}, \tag{52}$$

with μ a mass and λ a dimensionless coupling constant. As shorthand we have used ϕ^4 equal to the square of $\phi^2 = \phi_a \phi_a$.

The fields ϕ cannot have the usual interpretation as quantum fields since the mass term is negative and hence unphysical. Corresponding to this, the minimum of the potential is not at $\phi^2 = 0$, but rather at $\phi^2 = \mu^2/\lambda = v^2$. Normally in quantum field theory one expands about the minimum of the potential in fields that are described by excitations about the minimum. In this case the minimum of the potential corresponds to a value of ϕ with modulus equal to v, and so the quantum excitations are about this point and not about $\phi = 0$. We say then that in the true vacuum ϕ takes on a nonzero value and the expansion in quantum excitations needs to be made about this value. Pick then a direction in O(3), say the third direction, and let

$$\tilde{\phi}_a = \phi_a - (0, 0, v). \tag{53}$$

The vacuum expectation value (vev) of the field $\tilde{\phi}_a$ is zero while the vev of ϕ_a is

$$\langle 0|\phi_a|0\rangle = (0, 0, v). \tag{54}$$

The fields $\tilde{\phi}_a$ have zero vev and hence may be described in terms of the usual creation and annihilation operators, i.e., they have a standard quantum interpretation. Let us now examine the potential given above as a function of the physical fields. We find, after some algebra, that

$$V(\tilde{\phi}_a) = \mu^2 \tilde{\phi}_3^2 + 2v\tilde{\phi}_3 \tilde{\phi}^2 + \frac{\lambda \tilde{\phi}^4}{4}, \tag{55}$$

where we have dropped an overall constant and $\tilde{\phi}^2 = \tilde{\phi}_a \tilde{\phi}_a$.

Looking at the quantum fields, $\tilde{\phi}_a$, we see that the first two components are massless while the third has a positive mass. Let us recapitulate what we have found, suitably phrased so that we may generalize our conclusions. We started with a Lagrangian which was explicitly O(3) invariant, but the

vacuum was not invariant, as we saw when the field ϕ_3 was shown to have a nonzero vev. This breaking selected a preferred direction in O(3) space which we took to be the third direction, but which was chosen arbitrarily. In terms of canonical or physical fields, we found one massive field, ϕ_3, and two massless scalar fields, namely the other two components of $\tilde{\phi}_a$.

The symmetry of the Lagrangian, which was O(3), was reduced to rotations about the third axis, so that the symmetry associated with rotations about the 1 and 2 axes was lost. At the same time we saw the appearance of two massless scalar fields. We studied a particular example, but the situation is general. The massless scalars, also known as Goldstone bosons or Goldstone modes, are a general feature of spontaneously broken symmetries. They appear in particle physics, but they also appear in a variety of condensed matter physics problems, e.g., as spin waves in antiferromagnets (Fradkin, 1991).

SU(2) × SU(2) and even SU(3) × SU(3) appear to be good symmetries of the strong interactions. Why is this so? In a simple model of a free proton and neutron, the breaking of vector SU(2) is proportional to the proton-neutron mass difference:

$$[Q_a, H] = i \int \mathrm{d}^3x \partial_0 J_a^0 = \int \mathrm{d}^3x \partial_\mu J_a^\mu \sim (M_p - M_n); \quad (56)$$

but the breaking of the axial SU(2) is proportional to the sum of the neutron and proton masses and hence is apparently not small (Cheng and Li, 1984); however, the symmetry may be spontaneously broken. This requires a triplet of massless pseudoscalars (pseudoscalars rather than scalars since the symmetry is an axial symmetry); they are the Goldstone bosons of the theory. In this case we identify them with the pions. A theory with nucleons and pions can have SU(2) axial symmetry if either nucleons or pions are massless.

One way to understand the smallness of the symmetry breaking is to begin a discussion of the quark model as proposed by Gell-Mann and Zweig (Cheng and Li, 1984). Just as the constituents of nuclei are taken to be protons and neutrons, a doublet or two-component representation of SU(2), so we take the fundamental constituents of neutrons, protons, and other elementary particles to be quarks, the triplet or three-dimensional representation of SU(3).

The three species of quarks are known as the up, the down and the strange quarks or the u, d and s quarks, often written as q_i with $i = 1, 2, 3$. Mesons are bound states of a quark and an antiquark, while baryons are bound states of three quarks. The strong interaction that binds these quarks is taken to be invariant under SU(3) × SU(3) (more about this in the next section) as is the kinetic energy term, so that the only breaking of the symmetry in the Lagrangian, or equivalently the Hamiltonian, is due to the quark masses:

$$\mathscr{L} = \mathscr{L}_{symm} + \mathscr{L}_{mass} = \mathscr{L}_{symm} + m_i \bar{q}_i q_i. \quad (57)$$

The success of the quark model has been amply confirmed with the caveat that a free quark has never been seen (more about this in the next section). The values of the quark masses have been found to be

$$m_\mathrm{u} \sim 4 \text{ MeV}, \quad m_\mathrm{d} \sim 7 \text{ MeV}, \quad m_\mathrm{s} \sim 130 \text{ MeV}; \quad (58)$$

so one sees why SU(2) × SU(2) is such a good symmetry [and even why SU(3) and SU(3) × SU(3) are relatively good symmetries]. In the limit of $m_\mathrm{u} = m_\mathrm{d} = 0$, SU(2) × SU(2) is an exact symmetry

of the Lagrangian, broken spontaneously by the vacuum with the appearance of a massless triplet of Goldstone bosons, the pions. In reality the pions have small masses, proportional to m_u, m_d. It is the smallness of the pion mass which is the true measure of the symmetry breaking.

Of course another puzzle then rears its head, namely how does one build a 940 MeV object out of three objects with masses in the 5–10 MeV range? There is no simple answer to this: quark dynamics create so-called constituent quarks of approximately 300 MeV, while the current-algebra quarks, as described by $\mathscr{L}$, have much smaller values of mass. A major effort to unravel these questions has been embarked on by studies of quarks on lattices (Creutz, 1983).

8 Gauge Symmetries

So far we have considered rotations of fields by parameters θ_a that are independent of space-time. Can we generalize this notion? The problem is seen immediately by looking at the free Lagrangian of a spin-$\frac{1}{2}$ particle,

$$\mathscr{L} = \overline{\psi}(x)[i\gamma^\mu \partial_\mu - m]\psi(x), \tag{59}$$

which is invariant under the transformation $\psi(x) \to e^{i0(x)}\psi(x)$ only for constant θ. If the rotation is space-time dependent, the derivative in $\mathscr{L}$ will introduce extra terms. If we consider not a free theory, but one with minimal coupling to a spin-1 gauge boson, the Lagrangian is invariant under space-time dependent transformations. Minimal coupling means that

$$\partial_\mu \to \partial_\mu + igA_\mu(x), \tag{60}$$

where g is a dimensionless coupling constant. The Lagrangian is invariant under a rotation of ψ by $\theta(x)$ if at the same time we shift

$$A_\mu(x) \to A_\mu(x) - \frac{\theta(x)}{g}. \tag{61}$$

Having introduced the gauge field (photon), we must now include in $\mathscr{L}$ a term for the kinetic energy: this must also be invariant under the gauge transformation and takes the form

$$\mathscr{L}_{\mathscr{A}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{62}$$

with the so called field strength defined as

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{63}$$

A further requirement for local gauge invariance is that the photon be massless because a mass term in $\mathscr{L}$ for the photon would have to be of the form $M^2 A_\mu A^\mu$ and hence would not be invariant under the translation of $A_\mu(x)$ given in Eq. (61).

Invariance under local symmetry transformations requires the existence of minimally coupled massless gauge bosons.

The example we treated above corresponds to a U(1) gauge transformation, namely the fields were only rotated by a phase $\theta(x)$. Suppose now that we want the more general transformations, as illustrated in Eqs. (9) and (11), to be dependent on space-time, i.e., $\theta_a \to \theta_a(x)$. The solution to this problem was given by Yang and Mills (1954), who showed how to maintain invariance of the Lagrangian when the gauge fields themselves form a multiplet of the group G.

The prescription for invariance under an infinitesimal transformation $\delta\theta_a = \varepsilon_a(x)$ whereby

$$\psi_i(x) \to [1 + it_a\varepsilon_a(x)]_{ij}\psi_j(x), \tag{64}$$

is that ordinary derivatives be replaced by covariant derivatives

$$\partial^\mu \to D^\mu = \partial^\mu + igt_a A_a^\mu \tag{65}$$

where g is the dimensionless coupling constant, and that the gauge fields transform as

$$A_a^\mu(x) \to A_a^\mu(x) - \partial^\mu \frac{\varepsilon_a(x)}{g} - C_{abc}\varepsilon_b(x)A_c^\mu(x). \tag{66}$$

For continuous groups it is sufficient to specify the transformations under infinitesimal transformations. From the above we see that the number of gauge bosons must equal the number of generators of the group G, or equivalently the gauge bosons must belong to the adjoint representation of G. For instance, if G is the unitary group $SU(N)$, there have to be $N^2 - 1$ gauge bosons, i.e., three for $SU(2)$ and eight for $SU(3)$.

As is the case for electromagnetism, the gauge bosons must be massless. The field strength needs to be of the form

$$F_a^{\mu\nu} = \partial^\mu A_a^\nu - \partial^\nu A_a^\mu - gC_{abc}A_b^\mu A_c^\nu \tag{67}$$

in order for the Lagrangian to be invariant.

The prototype of a gauge theory is the theory of the strong interactions known as quantum chromodynamics. It assumes that each type of quark comes in three species known as colors and that there is a gauged $SU(3)$ symmetry group that acts on them. There is a red, a white, and a blue up type quark and similarly three down quarks and three strange quarks. Since the gauge group is $SU(3)$, there must be eight massless spin-1 particles, the gauge bosons. These are commonly called gluons.

The field strength has terms proportional to the square of the field A_a^μ, and therefore the so-called kinetic term, proportional to the field strength squared, has terms proportional to the third and fourth powers of the field A_μ. This means that $\mathscr{L}$ has nonlinear interactions of the gauge field. This is only true when we have a so called non-Abelian symmetry group, i.e., one for which $C_{abc} \neq 0$. For electromagnetism such terms are not present.

The nonlinear interactions of the gluons cause the effective quark–gluon coupling constant (coupling constant as modified by radiative corrections) to have a very different behavior than the electromagnetic coupling constant. In fact the quark–gluon coupling becomes smaller as one increases the momentum transferred to the gluon; for electromagnetism the electron's coupling to the photon becomes smaller as one decreases the momentum transferred to the photon. These changes in the effective coupling constants are quantum effects due to the dielectric behavior of the vacuum. They are not present in the corresponding classical field theories.

Conversely, quark interactions at small momentum transfer are very strong, leading to quark confinement: only $SU(3)$ color singlet states are observable. For instance, the quark–antiquark potential grows with separation while the electron–positron potential decreases. Since quark interactions become weaker at large momentum transfer, the quarks become almost free, a property known as asymptotic freedom. Both aspects have been confirmed experimentally (Aitchison and Hey, 1989).

As of now the evidence indicates that both the $U(1)$ symmetry of electromagnetism and $SU(3)$ of color are unbroken, or equivalently, unbroken local $SU(3) \times U(1)$ is a symmetry of nature.

Gravity is also an unbroken gauge symmetry of sorts. The unbroken symmetry

is general covariance, the conserved quantity is the energy-momentum tensor, and the gauge boson is the graviton, a quantum fluctuation of the metric tensor (Weinberg, 1972). Intuitively it is not surprising that if the quantum field that couples to a conserved current has spin 1, the quantum field that couples to a conserved second rank tensor should have spin 2.

The notion that gravity and electromagnetism might be reunited was of course Einstein's dream. This concept is presently being pursued by theorists exploring string theory (Gross, 1995), but that is outside the scope of this article.

We have seen that three of the four fundamental forces (electromagnetic strong, and gravity) seem to be described by gauge theories with massless gauge bosons as mediators. What about the fourth known force, the weak force that describes beta decay, neutrino interactions, *etc.* Is it like the other three? The answer is that it is indeed similar to the other three forces in that there is a gauge invariance associated with the weak interactions, but there is also an important difference in that, although the Lagrangian is invariant under the corresponding transformations, the physical vacuum is not. This implies, as we shall see in the next section, that the corresponding gauge bosons are massive. In turn, this leads to the conclusion that the force generated by the weak interactions is short range, as opposed to the electromagnetic or gravitational forces.

Equation (48) gives an effective Hamiltonian describing the weak interactions, but we now understand that this is only valid to second order in perturbation theory for an underlying theory with interaction

$$\mathscr{L}_{weak} = gJ^{a\mu}A_{a\mu} \tag{68}$$

where $a = 1, 2, 3$ is the label for a set of three currents associated with the three gauge bosons of the weak interactions. The g and A^{μ} in the above equation are not meant to be confused with those we wrote for the strong interactions.

The breakthrough in understanding the connection of such a model with gauge theories came about with the development of the so-called Glashow–Weinberg–Salam model (Weinberg, 1995). It was realized that the underlying theory was a gauge theory with SU(2) invariance and hence three massless spin-1 gauge bosons and three associated conserved currents.

The quarks and leptons, which constitute the fermionic components of the theory, are placed in doublets and singlets of SU(2), and, of course, the Lagrangian is invariant under SU(2) symmetry transformations. Having the currents conserved is only possible if all fermion masses are zero, since the divergences of the weak currents are, as we have seen, proportional to sums and differences of fermion masses. This is achieved by putting all left chiral fermions in doublets and right chiral fermions in singlets (mass terms connect left to right chiral fields) and of course, SU(2) invariance forbids direct coupling of doublets to singlets.

The resulting model is elegant, but far from describing nature since neither the mediators of the weak interactions nor the fermions are massless. On the other hand, introducing mass terms in the Lagrangian would break the local symmetry and destroy the renormalizability and hence the finiteness of the theory.

The resolution of this problem using the technique of spontaneous symmetry breaking is explained in the next section. It succeeds in giving masses to the three gauge bosons of the weak interactions: two of them, called $W^{\pm}$, are charged and the remaining one, the Z^0, is neutral. In summary, we say that the so-called

"standard model" of the strong, weak and electromagnetic interactions is described by a SU(3) × SU(2) × U(1) gauge theory, spontaneously broken to SU(3) × U(1). At present the "standard model" includes as fermionic building blocks three negatively charged leptons, e^-, μ^-, τ^-, each with its associated neutrino, three charge-$\frac{2}{3}$ quarks, the u, c, and t quarks, and three charge-$\left(-\frac{1}{3}\right)$ quarks, the d, s, and b quarks. In addition each and every quark comes in three so-called colors, corresponding to the triplet representation of the unbroken color SU(3) group.

An active field of research is the study of whether there exists some much larger gauge group, e.g., SU(5), which spontaneously breaks to the standard-model invariance group (Aitchison and Hey, 1989).

9 The Higgs Mechanism

There is one very important exception to the rule that spontaneously broken symmetries lead to Goldstone modes. This is the so-called Higgs mechanism, anticipated and discussed in the context of condensed matter physics by Y. Nambu and by P. W. Anderson (Weinberg, 1995). If the theory includes minimally coupled massless gauge bosons, they necessarily interact with the scalar bosons because the latter's kinetic energy terms need to have ordinary derivatives replaced by covariant derivatives. For instance, for O(3), we find

$$\begin{aligned}\partial^\mu \phi_a \partial_\mu \phi_a &\to D^\mu \phi_a D_\mu \phi_a \\ &= (\partial_\mu \phi_a + ig\varepsilon_{abc} A_{\mu,b}\phi_c) \\ &\quad \times (\partial^\mu \phi_a + ig\varepsilon_{ade} A^\mu_d \phi_e). \quad (69)\end{aligned}$$

If we now introduce into the above equation the vacuum expectation value of ϕ_a, namely the O(3) vector (0, 0, v), we find an effective mass term for the gauge bosons

$$g^2 v^2 \varepsilon_{ab3}\varepsilon_{ad3} A_{\mu,b} A^\mu_d. \quad (70)$$

From this, we see that the third component field A^μ_3 has no mass term since $\varepsilon_{33a} = 0$ while the other two fields have acquired a mass proportional to gv. This acquisition of mass is puzzling since a massless gauge boson like the photon has only two polarization degrees of freedom while a massive gauge boson can have a longitudinal polarization as well as the two transverse polarizations.

The resolution of the puzzle lies in the Higgs mechanism. We apparently acquired two new degrees of freedom corresponding to the longitudinal modes of the gauge bosons. If we study the problem somewhat more carefully, we find that two massless scalar bosons can be removed from the Lagrangian by a gauge transformation and hence disappear from the spectrum. The colloquial expression is that they have been eaten by the gauge bosons.

Recapitulating, we see that spontaneous symmetry breaking leads to massless scalar bosons: if the scalar bosons are coupled minimally to massless gauge bosons, some or all of the massless scalars may disappear from the spectrum, replaced by the longitudinal modes of the now massive gauge bosons. Gauge bosons coupled to currents associated with unbroken symmetries remain massless. In the O(3) case rotational symmetry about the third axis is unbroken and A^μ_3 remains massless.

For the theory of electroweak interactions our starting gauge theory is SU(2) × U(1) with four massless gauge bosons. In addition we have a complex SU(2) scalar

doublet, which transforms nontrivially under U(1) as well, so that there are four independent scalar fields (read and imaginary parts of the doublet). The fermions are left-handed chiral SU(2) doublets and right-handed chiral SU(2) singlets.

To begin with, the gauge bosons are massless, the fermions are massless [left-right couplings are forbidden by SU(2) invariance], and the scalars have a negative squared mass. The symmetry is spontaneously broken from SU(2) × U(1) to simply a U(1) symmetry. This leads to one massive scalar and three massless scalars. The latter are eaten by the four gauge bosons, three of which become massive, and one, the photon, stays massless. The three massive gauge bosons are the $W^{\pm}$ and the Z^0, the mediators of the charged and neutral weak interactions.

The fermions also acquire a mass, since Lagrangian interaction terms of the form

$$f\overline{q}_{L,a}\phi_a q_R \tag{71}$$

are allowed, and, once ϕ_a acquires a vacuum expectation value equal to v, fermions get masses proportional to fv. The predictions of what has come to be known as the standard model are extraordinarily successful. The $W^{\pm}, Z^0$ have been discovered at exactly the predicted masses, decay modes agree, *etc.* The one remaining unconfirmed piece of the puzzle is the detection of the single massive scalar, commonly called the Higgs boson.

The Higgs phenomenon has its counterpart in condensed matter physics as well. Let us illustrate this again with the example of super-conductivity. The gauge symmetry is U(1), with the photon as the gauge boson. The scalar field is the order parameter, the complex scalar electron-pair wave function, which also transforms nontrivially under U(1) since it has charge −2. The order parameter has a nonzero vev, signaling a phase transition, and the phase of the order parameter is a massless excitation, corresponding to a Goldstone mode. The latter disappears from the spectrum, being eaten by the photon. Inside a superconducting medium, the photon is essentially massive: this is known as the Meissner effect (Weinberg, 1995).

10
Experimental Limits on Symmetry Breaking

We have discussed at some length the notion of symmetry. Let us now consider, at least briefly, some of the experimental limits on conservation laws. Begin with discrete symmetries; *TCP* follows, as we stated, from general principles of relativistic quantum field theory. The best limits on *TCP* invariance come from the equality of the K^0 and the $\bar{K}^0$ masses. The experimental limits on the mass difference Δ are that

$$\frac{\Delta}{m_{K^0}} \leq 10^{-18} \tag{72}$$

at a 90% confidence level. Unless otherwise indicated the experimental results cited in this section are all obtained from the Particle Data Group's compilation (Particle Data Group, 1994). *TCP* also predicts the equality of lifetimes *etc.* for particles and antiparticles. Tests have been performed to compare lifetimes for electrons and positrons, $\mu^{\pm}$ mesons, and so forth. As an example, the lifetimes of the latter are given by

$$\frac{\tau_{\mu}^{+}}{\tau_{\mu}^{-}} = 1.00002 \pm 0.00008. \tag{73}$$

Another example is the equality of the anomalous magnetic moments of the electron and the positron. If *TCP* invariance holds, T and CP symmetries

are presumably equal, but again it is worth testing them separately. Experiments in atomic, molecular, nuclear, and particle physics are being conducted to test these symmetries. At present the most sensitive test of time-reversal invariance is probably the limit on the electric dipole moment (edm) of the neutron. The present limit for the neutron is that its edm $< 1.1 \times 10^{-25}$ e cm. The bound on the electron edm is comparable to that of the neutron and is a more stringent test of unification theories. It is worth mentioning that a neutron nonzero edm requires violation of both T and P. Nuclear physics searches for T nonconservation are also being pursued (Haxton et al., 1994).

CP symmetry is a vast topic (Jarlskog, 1988). At present the best limits on CP-invariance violation come from the analyses of K decay. In the neutral kaon system the strength of CP-invariance violation relative to the standard weak interactions is characterized by a dimensionless small parameter $\varepsilon \simeq 10^{-3}$. It is crucial to determine if ε is so small because CP-invariance violations are very weak or because they involve mixing with heavier quarks, which are only minor constituents of the kaons.

Large experimental facilities in Japan (KEK) and the USA (SLAC) are presently under construction that will provide further tests of CP nonconservation and begin to address these questions. An example is the so called BABAR facility at the Stanford Linear Accelerator which will study the $B^0 - \bar{B}^0$ system (B^0 is a bound state of a b quark and either a d or an s antiquark).

Baryon-number conservation is presumed to be violated at some level if a grand unified theory that incorporates all known gauge theories exists. The present best limit on baryon-number conservation comes from the stability of the proton, which is known to have a lifetime greater than 10^{31} years.

Baryon-number nonconservation is one of the three key ingredients to the generation of a baryon asymmetry in the early universe, as was pointed out by Andrei Sakharov in a prescient 1967 paper (Kolb and Turner, 1990). The other two are CP-invariance violation and a departure from thermal equilibrium. Baryon asymmetry is, of course, present since our known universe is preponderantly made of matter, not of antimatter, but a wholly successful explanation of the asymmetry has not yet been advanced.

Lepton number is tested in neutrinoless double beta decay, particularly in isotopically enriched germanium (Avignone, 1995). The limits state the ratio of neutrinoless decays to neutrino decays to be less than one in a thousand.

The separateness of muon number and electron number is tested by the nonobservation of $\mu \to e + \gamma$, which has a branching ratio of $< 5 \times 10^{-11}$ to the total decay rate of the muon. The dominant muon decay is into a state of an electron, a neutrino, and an anti-neutrino. Though it is not obvious, lepton numbers are automatically conserved if neutrinos are massless (technically it would be sufficient if all neutrinos were mass degenerate). There are numerous ongoing searches for evidence of nonzero neutrino mass. Lepton mixing and nonzero neutrino masses may first be observed in so-called neutrino oscillation experiments. In these one species of neutrino, say an electron neutrino, is produced but it then oscillates between that and a second species, e.g., a muon neutrino, while travelling from the source to the observer.

Local gauge invariance requires massless gauge bosons. One obvious verification is the masslessness of the photon. Present

limits are

$$m_\gamma < 3 \times 10^{-27}\ \text{eV}. \qquad (74)$$

Trying to set limits on gluon (the presumed mediator of the strong interactions) masses, on the other hand, is very difficult, since gluons are not directly observable. In fact, observation of a free gluon or a free quark would represent a departure from the presently held view that these particles are only observable when combined with other like such particles to form color singlets. Incidentally, all recent searches for free quarks have yielded null results.

Conservation of electric charge is presumed to hold. If it were not, we could imagine an electron decaying into a neutrino and a photon. Present limits on the electron lifetime are $\tau_e > 2.7 \times 10^{23}$ years.

One prominent symmetry we have not mentioned in the text so far is supersymmetry, the unique allowed extension of space-time translation symmetry to a group that includes particle transformations. The symmetry requires paired multiplets, necessarily degenerate in mass, of fermions and bosons. Thus, if supersymmetry exists in nature, it must be spontaneously broken. Some current theories suggest that the partners of the known elementary particles, differing in spin by one half unit, should lie in the mass range of a few hundred GeV. It is therefore an interesting question for the next round of accelerator experiments whether this extended symmetry could be a reality (Wess and Bagger, 1983).

Glossary

Baryon: A strongly interacting particle obeying Fermi statistics, such as the proton or the neutron.

Bravais Lattice: The elementary specification of the location of the units in an infinite crystalline lattice, making manifest the periodic ordering and the translational symmetry.

Charge Conjugation: An operation that transforms a particle into its antiparticle. A theory which incorporates particles and antiparticles in a symmetric manner may then be invariant under the operation of charge conjugation or said to have charge-conjugation symmetry. Similar statements can be made about time-reversal and parity symmetries.

Gauge Boson: A particle of spin 1 that couples to matter in such a way as to maintain gauge invariance, the invariance under continuous space-time dependent transformations. The photon is an example of a gauge boson.

Goldstone Boson: A massless particle of spin 0 that arises when a continuous symmetry is broken spontaneously.

Group: A set of elements and a rule for combining them, commonly called group multiplication. The product of any two group elements must itself be a group element, multiplication must be associative, a unique identity element must exist, and every element must have a unique inverse such that the product of an element and its inverse yields the identity element. Groups may be continuous, such as the full rotation group, or discrete, such as the group of rotations about the z axis by 90°, which has only four elements: rotations through 90, 180, 270 and 360°.

Lepton: The name given to fermionic elementary particles with no strong interactions, as for example the electron and its neutrino.

Quantum Field: The generalization of the concept of a particle satisfying a quantum mechanical equation of motion. Particles are the quantum excitations of the appropriate field, e.g., the photon is an excitation of the electromagnetic field. Quantum field theory allows naturally antiparticle as well as particle creation and annihilation.

Quark: The presumed structureless, i.e., pointlike, elementary fermionic constituents of strongly interacting particles such as the neutron and the proton. Quarks are believed to be bound together by the mediators of the strong force, conventionally called gluons.

Spontaneous Symmetry Breaking: A situation that arises when the Lagrangian and hence the equations of motion of a system are invariant under a continuous symmetry, but the vacuum is not. This comes about because the minimum of the potential and hence the stable point corresponds to a nonzero expectation value of a scalar field.

Standard Model: The term commonly used for the model of matter described as consisting of pointlike fermionic constituents (quarks and leptons) interacting by means of gauge bosons, which in turn are self interacting. The model features an invariance under a set of specified symmetry transformations, corresponding to the group $SU(3) \times SU(2) \times U(1)$.

List of Works Cited

Adler, S., Dashen, R. (1968), *Current Algebras*, New York: Benjamin Press.

Aitchison, I. J., Hey, A. G. (1989), *Gauge Theories in Particle Physics*, Philadelphia: Adam Hilger.

Ashcroft, N., Mermin, D. (1976), *Solid State Physics*, Philadelphia: Saunders.

Avignone, F. (1995), in: *Particle and Nuclear Astrophysics and Cosmology in the Next Millennium*, E. Kolb, R. Peccei (Eds.), Singapore: World Scientific.

Cheng, T. P., Li, L. F. (1984), *Gauge Theory of Elementary Particle Physics*, Oxford, UK: Clarendon Press.

Christenson, J., Cronin, J., Fitch, V., Turlay, R. (1964), *Phys. Rev. Lett.* **13**, 138–142.

Commins, E., Bucksbaum, A. (1983), *Weak Interactions of Leptons and Quarks*, Cambridge: Cambridge University Press.

Greutz, M. (1983), *Quarks, Gluons and Lattices*, Cambridge: Cambridge University Press.

Fradkin, E. (1991), *Field Theories of Condensed Matter Systems*, New York: W. A. Benjamin.

Gross, D. J. (1995), "Symmetry in Physics: Wigner's Legacy," *Phys. Today* **48**(12), 47–55.

Haxton, W., Horing, A., Musolf, M. (1994), *Phys. Rev. D* **50**, 3422–3454.

Itzykson, C., Zuber, J. B. (1980), *Quantum Field Theory*, New York: McGraw-Hill.

Jarlskog, C. (1988), *CP Symmetry*, Singapore: World Scientific.

Kolb, E. W., Turner, M. S. (1990), *The Early Universe*, New York: Addison – Wesley.

Landau, L. D., Lifshitz, E. M. (1981), *Quantum Mechanics- Non Relativistic Theory*, London: Pergamon Press.

Lax, M. J. (1974), *Symmetry Principles in Solid State and Molecular Physics*, New York: Wiley.

Lee, T. D., Yang, C. N. (1956), *Phys. Rev.* **104**, 254–266.

Particle Data Group, M. Aguilar-Benitez et al., (1994) *Phys. Rev. D* **50**, 1173–1827.

Steinhardt, P. J., Ostlund, S. (1987), *Quasi-Crystals*, New York: World Scientific.

Streater, R. F., Wightman, A. S. (1968), PCT, *Spin and Statistics and All That*, New York: Benjamin Press.

Weinberg, S. (1972), *Gravitation and Cosmology*, New York: John Wiley.

Weinberg, S. (1995), *The Quantum Theory of Fields*, Cambridge: Cambridge University Press.

Wess, J., Bagger, J. (1983), *Supersymmetry and Supergravity*, Princeton: Princeton University Press.

Wigner, E. P. (1959), *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*, New York: Academic Press.

Yang, C. N., Mills, R. L. (1954), *Phys. Rev.* **96**, 191–205.

Further Reading

Coleman, S. (1989), *Aspects of Symmetry*, Cambridge UK: Cambridge University Press. A review of modern topics in quantum field theory, with particular emphasis on the gauge theory and its geometric underpinnings.

Lee, T. D. (1981), *Particle Physics and Introduction to Field Theory*, Chur, Switzerland: Harwood Academic. A review of field theory and how it is used in elementary particle physics.

Mermin, D. (1992), *Rev. Mod. Phys.* **64**, 3–122. A learned review of space groups in crystallography.

Peskin, M., Schroeder, R. (1995), *An Introduction to Quantum Field Theory*, Reading, MA: Addison-Wesley. A modern reference for field theory and elementary particle physics.

Weyl, H. (1931), *The Theory of Groups and Quantum Mechanics*, New York: Dover Publications. A classic introduction to group theory and its applications to quantum mechanics.

Wigner, E. P. (1967), *Symmetries and Reflections: Scientific Essays of Eugene P. Wigner*, Bloomington: Indiana University Press. An overview of the notions of symmetry.

Topology

S. P. Smith
Department of Mathematics and Computer Science, California State University, Hayward, California, USA

Introduction ***588***
1 Point-Set Topology ***591***
1.1 Basic Concepts *592*
1.2 Continuity *594*
1.3 Connectedness and Compactness *594*
2 Algebraic Topology ***596***
2.1 Some Basic Tools *597*
2.2 Homotopy *599*
2.3 Homology *601*
2.4 Cohomology *605*
3 Differential Topology ***606***
3.1 Manifolds and Bundles *607*
3.2 Vector Fields and the Poincaré–Hopf Index Theorem *609*
3.3 Differential Forms and de Rham Cohomology *610*
3.4 Morse Theory *611*
4 A Brief Guide to Further Reading ***613***
4.1 Point-Set Topology *613*
4.2 Algebraic Topology *613*
4.3 Differential Topology *613*
4.4 Physical Applications *614*
Glossary ***614***
Further Reading ***615***

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

Introduction

Topology is the study of properties that are invariant under continuous deformation. If an object A can be continuously deformed into an object B, and B can be continuously deformed into A, then A and B are considered topologically equivalent. For example, a two-dimensional circular disk A is equivalent to a square B since each may be continuously deformed into each other. Such a disk is not, however, equivalent to the circle that is its boundary, since there is no way to deform a disk to a circle without introducing a discontinuity, i.e., a hole or tear, in the disk. Topology is sometimes called "rubber sheet geometry." It ignores metric aspects of a space such as distance, angle, and area, and concentrates on the aspects associated with the relative position of its points. The word "topology" comes from the Greek words "topos" and "logos" and means "analysis of place"; in some older literature, topology is called *analysis situs*.

"Cut-and-paste" techniques are common methods used to construct spaces in topology. For example, imagine a square sheet of completely flexible and stretchable material sitting in the plane with the corner points labeled A, B, C, and D starting with the upper left-hand corner and moving clockwise as in Fig. 1. Stretching the square out lengthwise and gluing the top edge AB to the bottom edge DC, with A attaching to D and B attaching to C, gives an orientable space that is a portion of a cylinder; it has two sides, and its boundary is equivalent to two circles. In Fig. 1 sets of arrows are placed on identified sides to indicate the orientation used in gluing. If the edge AB is given a 180° twist before the gluing so that A attaches to C and B attaches to D, then the resulting surface is the *Möbius strip*, which is nonorientable: it has only one side, and its boundary is equivalent to a single circle. Walking once around the central circle of a Möbius strip, a traveler returns to the same point but in an upside-down position. In the same way, the usual *torus* T is obtained from the above square by gluing the side AB to the side DC and the side BC to the side AD. The first gluing produces the cylinder as before, which is then stretched out and bent around for the second gluing. The torus is equivalent to the surface of a doughnut or the surface of a ball with a single hole drilled through it. The *Klein bottle* K is obtained by gluing AB to DC as before, but giving BC a 180° twist before gluing it to AD so that B attaches to D, and C attaches to A. The *real projective plane* $\mathbf{R}P^2$ results from twisting both AB and BC 180° before gluing, so that AB glues to CD and CB glues to AD with A attaching to C and B attaching to D. Of course, after the gluings for the torus and the Klein bottle, the points A, B, C, and D become a single point. The Klein bottle and real projective plane are nonorientable surfaces, and also both of these surfaces have self intersections when considered as subsets of $\mathbf{R}^3 = \{(x_1, x_2, x_3) | x_1, x_2, x_3 \in \mathbf{R}\}$. The above construction of $\mathbf{R}P^2$ can be viewed as taking a two-dimensional disk and gluing each point of its boundary to its antipodal point on the boundary. $\mathbf{R}P^2$ is the set of lines through the origin in $\mathbf{R}^3$. In general, *n-dimensional real projective space* $\mathbf{R}P^n$ is the set of all one-dimensional subspaces (i.e., lines through the origin) in $\mathbf{R}^{n+1}$, and the *n-dimensional complex projective space* $\mathbf{C}P^n$ is the set of all one-dimensional subspaces in $\mathbf{C}^{n+1}$. The dimensions in these projective spaces are taken with respect to the underlying scalars; thus $\mathbf{C}^3$ is three-dimensional as a complex manifold but six-dimensional as a real manifold. The ordinary sphere S^2 can also be obtained from

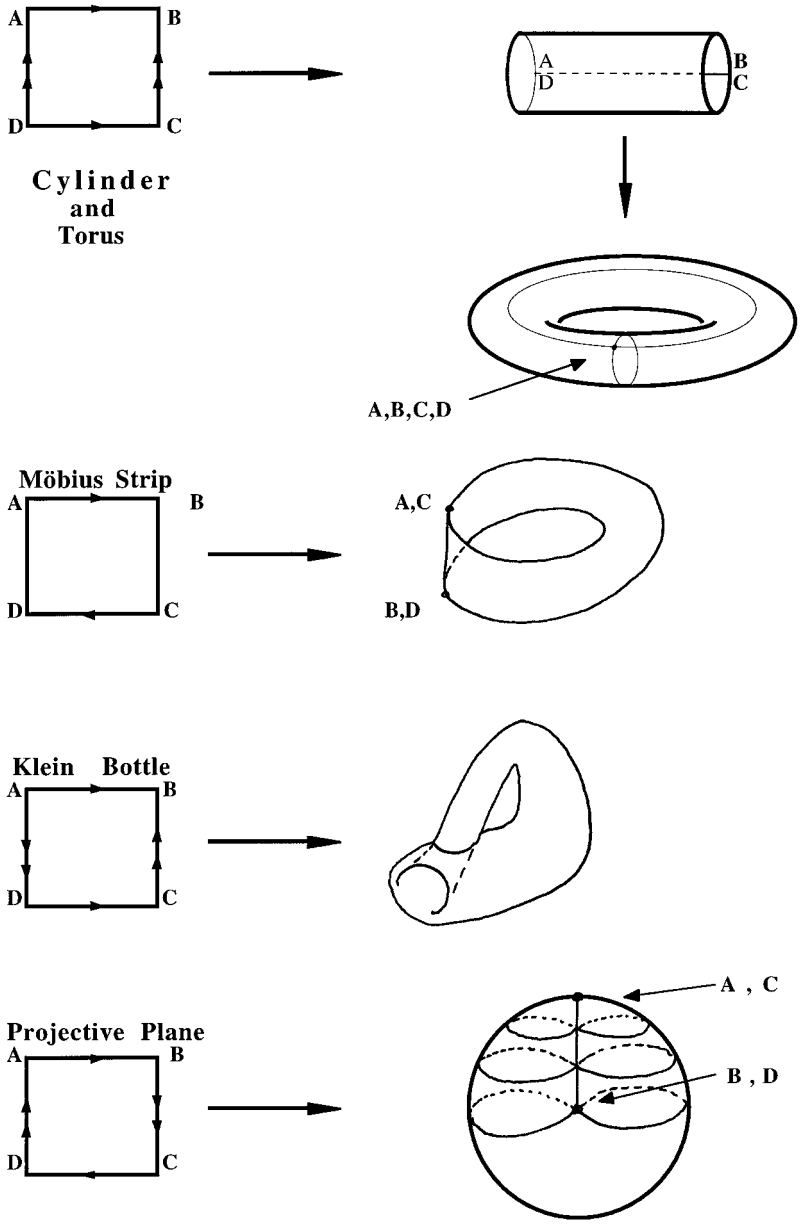


Fig. 1 Cut-and-paste methods for standard two-dimensional spaces

the square ABCD by simply collapsing the boundary of the square to a point. If the sides of a regular octagon are identified as in Fig. 2, the result is a two-holed torus.

After the spaces $\mathbf{R}^n$, the spheres are the most commonly encountered spaces in topology and are best considered simply as the set of points in $\mathbf{R}^{n+1}$ that

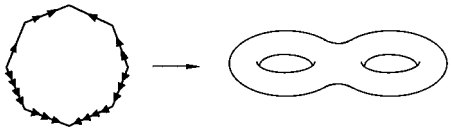


Fig. 2 Gluing the indicated sides produces the two-holed torus

are at unit distance from the origin; i.e., the *n-dimensional sphere* is $S^n = \{x \in \mathbf{R}^{n+1} | \ \|x\| = 1\}$. Each n-sphere S^n is the boundary of the unit $(n+1)$-*dimensional ball* $B^{n+1} = \{x \in \mathbf{R}^{n+1} | \ \|x\| \leq 1\}$. Notice that the 0-sphere S^0 consists of two points and is therefore the only sphere that is not connected. The *g-holed torus* T_g is the two-dimensional space equivalent to the surface of a three-dimensional ball with g non-intersecting holes drilled through it. Using the above spaces, large numbers of examples can be constructed using additional operations such as Cartesian products and connected sums. Given spaces M and N, the *Cartesian product* $M \times N$ is defined to be $\{(p, q) | p \in M \ \& \ q \in N\}$. If M and N have the same dimension n, the *connected sum* $M \# N$ is constructed by removing an n-dimensional ball from each of M and N, and then gluing the resulting spaces together along the boundaries of the removed balls. Thus, the Cartesian product of two circles $S^1 \times S^1$ is the two-dimensional torus T, and the connected sum $T_g \# T_k$ of a g holed torus and a k-holed torus is a $(g+k)$-holed torus.

Important isolated problems of a topological nature were treated by Descartes and Euler (the Descartes–Euler characteristic for polyhedra and the Königsberg bridge problem), and Riemann emphasized the importance of topological properties in his studies of complex analysis and algebraic functions. The key property of a Riemann surface is its genus, which counts the number of "holes" associated to the surface. Thus a two-dimensional sphere has genus 0, and a two-dimensional torus such as the surface of a doughnut has genus 1. With Cantor's development of set theory, mathematics entered a new era. More attention was paid to foundational questions, and a higher level of rigor was demanded. The real-number system had been carefully developed by Dedekind and Cantor in a constructive way, and mathematicians saw an opportunity to place all of mathematics on an axiomatic foundation similar to Euclid's *Elements*. The central areas in this new development were topology and abstract algebra. A very large part of twentieth century mathematics is associated with this program.

This development seems at first to be very remote from physics and the sciences. It is more like mathematical housecleaning: making more precise definitions, giving clearer statements and proofs of theorems, finding the weakest hypotheses that give a result, determining whether a property is metric or topological in nature, etc. However, it is important to note that most of the key questions arose from analysis of tools used in physical applications. Cantor's set theory came from his study of the convergence of Fourier series; Sophus Lie developed the theory of continuous groups to make the connection between symmetries and methods of solution of differential equations; Poincaré's deep study of the question of the "stability of the solar system" led him to realize that complicated systems were very difficult to treat quantitatively and that chaotic behavior was possible. In studying periodic orbits in celestial mechanics, Poincaré formulated a key difficulty as a topological conjecture, Poincaré's last geometric theorem, which was proved by G. D. Birkhoff.

It was clear at this point that topology was essential to the qualitative analysis of dynamical systems. This was one of the major motivating forces behind the initial development of topology and remains so. Further, the development of general relativity and the associated area of differential geometry brought into focus the key relation between local and global aspects of spaces. To illustrate: Suppose that the given space is a two-dimensional sphere in three-space. Locally – for example, in the Northern hemisphere – there exist an infinite number of harmonic functions, i.e., solutions of the Laplace equation $\Delta f = 0$, even if the value of f is prescribed at the North Pole. However, globally only one such harmonic function exists; it must be a constant function. The topology of the sphere forces each function that is harmonic on the entire sphere to be a constant function. Similarly, it is possible to construct a unit-length tangent-vector field on the Northern hemisphere, but globally such a vector field is impossible. However, the two-dimensional torus does admit such a global unit-length tangent-vector field, and this torus is the only orientable surface without boundary with this property. This result follows from the famous Poincaré–Hopf index theorem discussed in Sec. 3.2. Topology plays an essential role in quantifying these global aspects of spaces and is therefore essential in global analysis and boundary-value problems. It is important to note that in applications, topology is usually combined with other areas of mathematics such as algebra and differential geometry.

This article discusses the three most prominent branches of topology: point-set topology, algebraic topology, and differential topology. These areas are considered to be closest to applications in physics.

1 Point-Set Topology

Point-set topology, also called general topology, provides the basic foundation for mathematical analysis, the portion of mathematics most directly connected with the calculus and the theory of differential and integral equations. The key concepts are continuity and convergence. The subject assumes only basic set theory. Point-set topology begins with topological spaces and continuous functions between such spaces. The basic properties of such spaces and functions are studied, and then gradually new concepts such as connectedness and compactness are added and examined. This step-by-step development leads to a clearer understanding of the role each hypothesis plays in a given result and often points the way to important extensions and generalizations. For example, the extreme-value theorem may be stated in a simple case as follows: A real-valued function that is continuous on a closed bounded interval $[a, b] \subset \mathbf{R}$ attains its maximum and minimum values at some points of the interval. Further analysis of this theorem shows that two basic ideas, continuity and compactness, are essential hypotheses for this theorem. The more general theorem is then this: a real-valued function that is continuous on a compact topological space attains its maximum and minimum values at some points of that space.

A very concise summary of essential definitions and results follows. This list is not comprehensive; it contains only the most commonly encountered definitions and results of point-set topology. These are particularly important for a clear understanding of functional analysis as used in physics. For example, point-set topology is basic to the theory of Fourier

transforms, Sobolev spaces, finite-element methods, and the theory of distributions such as the Dirac delta function. The theory of distributions was a major motivation for the development of the theory of topological vector spaces.

Throughout this article, "iff" means "if and only if" and frequently indicates that a definition is being given.

1.1 Basic Concepts

A collection T of subsets of a set X is a *topology* on X iff

1. the empty set $\emptyset$ and the set X are in T;
2. T is closed under arbitrary unions; i.e., given any collection of subsets of X that are in T, their union is in T;
3. T is closed under finite intersections; i.e., given any finite collection of subsets of X that are in T, their intersection is in T.

A *topological space* (X,T) is a set X with a topology T on X. It is common to abbreviate (X,T) to X and *topological space* to *space*. If several topological spaces are involved, T may be replaced by T_X or another symbol such as S for clarity.

To illustrate, let $X = \{a, b, c\}$ be a set with three elements; then $T_1 = \{\emptyset, X\}$ and $T_2 = \{\emptyset, \{a\}, \{b\}, \{a, b\}, X\}$ are topologies on X, but $T_3 = \{\emptyset, \{a\}, \{b\}, X\}$ is not since condition 2 is not satisfied. In general, for any set X, there are two standard topologies that can always be defined on X: the *discrete topology* consisting of the collection of all subsets of X, and the *indiscrete* or *trivial topology* that consists of only $\emptyset$ and X itself.

If (X,T) is a topological space, then the elements of T are called the *open sets* of the space. A subset A of a topological space (X,T) is *closed* iff its complement $X - A$ is open. In the above example (X, T_2), $\{c\}$ is a closed set since its complement $\{a, b\}$ is open. Notice that open and closed are not opposites; a set can be both open and closed (e.g., X in T_1) or neither open nor closed (e.g., $\{a\}$ in T_1). The above topological spaces illustrate the definition clearly but are too artificial to be of much use in physics. A much more interesting example is given by $X = \mathbf{R}^n$, the n-dimensional vector space of all n-tuples of real numbers, with the *metric topology*. This is one of the most important topological spaces, and it is worth studying carefully. To define this topology some preliminary ideas are needed. For each pair of points a and b in $\mathbf{R}^n$, let $d(a, b)$ be the usual Euclidean distance from a to b. The *open ball* $B_r(p)$ centered at p and of radius $r > 0$ in $\mathbf{R}^n$ is defined to be $\{x \in \mathbf{R}^n | d(x, p) < r\}$, i.e., the set of all points strictly within distance r of p. Using these open balls the metric topology on $\mathbf{R}^n$ is defined as follows: A subset A is *open* in $\mathbf{R}^n$ iff for each $p \in A$, $B_r(p) \subset A$ for some $r > 0$. Intuitively, a subset A is open iff there is "space in all directions" in A around each of its points. This "space in all directions" is essential for the definition of key analytic concepts such as continuity and differentiation that require that the value of a function at a point be compared with values at all nearby points. Applying the definition is usually simple on an intuitive level. Thus, in the real line $\mathbf{R}$ the "open" interval (0,1) is an open set, the "closed" interval $[-1, 2]$ is a closed set since its complement, $(-\infty, -1) \cup (2, \infty)$, is an open set, and the half-open interval $(-3, 5]$ is neither open nor closed.

In general, a metric topology may be defined on any set that has an appropriate notion of distance, i.e., a *metric*, defined on it. Often the metric is given by a more elaborate structure on the set such as an inner product or a norm. Inner products,

norms, and metrics are formally defined as follows: Let V be a vector space over F where F is either the real or the complex number field, and let M be a set. Then $\langle , \rangle : V \times V \to F$ is an *inner product* on V iff

$$\langle A, B\rangle = \overline{\langle B, A\rangle} \text{ for all } A, B, \in V,$$

$$\langle A, bB + cC\rangle = b\langle A, B\rangle + c\langle A, C\rangle$$

$$\text{for all } A,\ B, C \in V \text{ and } b, c \in F, \text{ and}$$

$$A \in V \text{ and } A \neq 0 \text{ imply that } \langle A, A\rangle > 0.$$

An *inner product space* $(V, \langle , \rangle)$ is a vector space V together with an inner product $\langle , \rangle$ defined on it.

Next, $\| \cdot \| : V \to \mathbf{R}$ is a *norm* on V iff

$$\|aA\| = |a|\|A\| \text{ for all } A \in V \text{ and } a \in F,$$

$$\|A + B\| \leq \|A\| + \|B\|$$

$$\text{for all } A, B \in V, \text{ and}$$

$$A \in V \text{ and } A \neq 0 \text{ imply that } \|A\| > 0.$$

A *normed space* $(V, \| \cdot \|)$ is a vector space V together with a norm $\| \cdot \|$ defined on it.

Lastly, $d: M \times M \to \mathbf{R}$ is a *metric* on M iff

$$d(x, y) = d(y, x) \text{ for all } x, y \in M,$$

$$d(x, z) \leq d(x, y) + d(y, z)$$

$$\text{for all } x, y, z \in M, \text{ and}$$

$$d(x, y) \geq 0 \text{ for all } x, y \in M$$

$$\text{and } d(x, y) = 0 \text{ iff } x = y.$$

A *metric space* (M,d) is a set M together with a metric d defined on it. Each of these structures gives the vector space V a natural metric topology since each inner product $\langle , \rangle$ induces a norm $\| \cdot \|$, which in turn induces a metric via the rules $\|A\| = \sqrt{\langle A, A\rangle}$ and $d(A, B) = \|A - B\|$.

If (X,T) is a topological space, and A is a subset of X, then A inherits a topology S from X as follows: B is *open in* A iff $B = A \cap U$ for some open set U in X. Briefly, this is written $B \in S$ iff $B = A \cap U$ for some $U \in T$. This topology is called the *relative topology* on A. It is also called the *induced* or *subspace topology* on A. Thus in the metric topology on $X = \mathbf{R}$ above, the interval (1,2] is not open in X, but it is open in $A = [0, 2]$ since $(1, 2] = A \cap U$, with $U = (1, 3)$ for example. A subset $N \subset X$ is a *neighborhood* of a point $p \in X$ iff $p \in U \subset N$ for some open set $U \in T$. If, in addition, N is an open set, then N is called an *open neighborhood* of p.

Let A be a subset of a topological space (X,T). $p \in X$ is an *interior point* of A iff $p \in U \subset A$ for some $U \in T$, i.e., A is a neighborhood of p. $p \in X$ is an *exterior point* of A iff p is an interior point of $X - A$. $p \in X$ is a *closure* or *limit point* of A iff each neighborhood of p contains at least one point of A. $p \in X$ is a *boundary point* of A iff p is a closure point of both A and $X - A$. $p \in X$ is an *accumulation point* of A iff each neighborhood of p contains at least one point of A other than p. $p \in A$ is an *isolated point* of A iff p is not an accumulation point of A. The *interior* of A is the set of all interior points of A and is denoted int$A$. It is also the union of all the open subsets of A and is the largest open set contained in A. The *closure* of A is the set of all closure points of A and is denoted cl$A$. It is also the intersection of all the closed subsets that contain A and is the smallest closed set containing A. The *boundary* of A is the set of all boundary points of A and is denoted bd$A$. It is equal to cl$A$ − int$A$. In many cases these concepts behave as the names indicate. Thus, for the metric topology on $\mathbf{R}$, the set $A = [0, 1)$ has the open interval (0,1) as its interior, the closed interval [0,1] as its closure, and the set $\{0, 1\}$ consisting of the two "boundary points" as its boundary. However, these simple definitions contain some subtleties. For

example, the subset of all rational numbers in [0,1] has interior equal to the empty set, and both the closure and the boundary are equal to the closed interval [0,1]. The notations A°, $\bar{A}$, and ∂A are frequently used for int$A$, cl$A$, and bd$A$, respectively. A is *dense* in B iff $B \subset \mathrm{cl}A$. For example, the set $\mathbf{Q}$ of rational numbers is dense in the set $\mathbf{R}$ of real numbers, and the set of polynomials with rational coefficients is dense in the space $C^0([0, 1], \mathbf{R})$ of all continuous real-valued functions on [0,1] with the sup norm $\|f\| = \max\{|f(x)| \ |x \in [0, 1]\}$. This notion of denseness is essential in approximation theory.

A topological space (X,T) is *Hausdorff* iff for each pair of points a and b there exist disjoint open sets U and V such that $a \in U$ and $b \in V$. Hausdorff spaces are also called T_2 spaces. Hausdorff is the most common of the several *separation* conditions that may be imposed on a topological space. All metric spaces are Hausdorff spaces.

The above definitions give the flavor of point-set topology. It is abstract and very set theoretic. It is important to keep in mind some of the primary examples that are discussed in the following such as surfaces in $\mathbf{R}^3$ with the induced metric topologies, topological and differentiable manifolds, and especially function spaces such as Hilbert and Banach spaces.

1.2 Continuity

The δ-ε definition for continuity given in calculus is as follows: a function $f: \mathbf{R} \to \mathbf{R}$ is *continuous at the point p* iff for each $\varepsilon > 0$ there exists a $\delta_\varepsilon > 0$ such that $|x - p| < \delta_\varepsilon$ implies that $|f(x) - f(p)| < \varepsilon$. The function is said to be *continuous on* $A \subset \mathbf{R}$ iff f is continuous at each point of A. Intuitively, continuous functions are allowed to stretch or contract portions of their domains, but no breaks or tears are permitted. The above definition is easily generalized to functions between metric spaces by simply replacing the absolute values by the appropriate distances, but it does not apply to more general topological spaces in which no such distance is given. The following definition generalizes the notion of continuity to topological spaces and is equivalent to the δ-ε definition above in the metric-space case. A function $f: (X, T_X) \to (Y, T_Y)$ is *continuous on* X iff $f^{-1}(U)$ is open for each $U \subset Y$; i.e., inverse images of open sets are open. Here the inverse image $f^{-1}(U)$ is defined to be $\{x \in X | f(x) \in U\}$; inverse images are defined for all functions, invertible or otherwise. The above function f is *continuous on* $A \subset X$ iff f restricted to A with the relative topology is continuous on A. It is common to abbreviate *continuous function* to *map*. Two topological spaces (X, T_X) and (Y, T_Y) are considered to be *topologically equivalent* or *homeomorphic* iff there exists a bijective function $f: (X, T_X) \to (Y, T_Y)$ that is continuous on X and that has an inverse f^{-1} that is continuous on Y. Such a function f is called a *homeomorphism*. For the purposes of topology, homeomorphic spaces are considered to be the same and must have identical topological properties.

1.3 Connectedness and Compactness

Compactness, connectedness, and path connectedness are among the most important topological properties and occur very frequently in topological analysis. Let A be a subset of a topological space (X, T_X). A collection of open subsets of X is an *open cover* of A iff A is contained in their union. A *finite subcover* of such an open cover is any finite subcollection that also contains

A in its union. $A \subset X$ is *compact* iff each open cover of A has a finite subcover. For example, the open cover of the real line $\mathbf{R}$ given by $\{(n-1,\ n+1) | n \text{ is an integer}\}$ has no finite subcover; so $\mathbf{R}$ is not compact. Open covers are often denoted $\{U_i\}_{i \in I}$, where each U_i is an open set in the cover and I is an index set. $A \subset X$ is *connected* iff there do not exist open sets U and V in X with $A \subset U \cup V$, $A \cap U \neq \emptyset$, $A \cap V \neq \emptyset$, and $A \cap U \cap V = \emptyset$. Equivalently, $A \subset X$ is *connected* iff the only subsets of A that are both open and closed in the relative topology of A are $\emptyset$ and A. $A \subset X$ is *path connected* iff for each pair of points p and q in A there is a continuous function $\gamma: [0, 1] \to A$ such that $\gamma(0) = p$ and $\gamma(1) = q$. Such a continuous γ is called a *path* in A from p to q. Path-connected spaces are necessarily connected, but not vice versa. If $f: (X, T_X) \to (Y, T_Y)$ is a function continuous on X and $A \subset X$ is compact, then the image $f(A)$ is also compact; similarly, if A is connected (resp. path connected), then $f(A)$ is also connected (resp. path connected). Two important theorems result from these concepts. Continuity plus compactness gives the extreme-value theorem: If $f: X \to \mathbf{R}$ is continuous on X, and X is compact, then f must attain its maximum and minimum values at some points of X. Continuity plus connectedness gives the intermediate-value theorem: If $f: X \to \mathbf{R}$ is continuous on X, and X is connected, and $f(a) < z < f(b)$ for some a and b in X and some real number z, then there is a $c \in X$ with $f(c) = z$. Two simple consequences of the intermediate-value theorem are the simple fixed-point theorem: Each continuous function $f: [0, 1] \to [0, 1]$ must have a fixed point, i.e., there is an $x \in [0, 1]$ such that $f(x) = x$; and the heated-ring theorem: Given any continuous function $f: S^1 \to \mathbf{R}$, there exists at least one pair of antipodal points $x, x' = -x$ such that $f(x) = f(x')$. If S^1 is considered as the unit circle in the complex plane, and f is taken to be the temperature at each point, then this means that $f(x) = f(-x)$ for some $x \in S^1$; i.e., there are always at least two opposite points with the same temperature.

Continuity and compactness also combine to produce uniform continuity. A function $f: X \to \mathbf{R}$ is *uniformly continuous* on a metric space (X,d) iff for each $\varepsilon > 0$ there is a $\delta_\varepsilon > 0$ such that $x, y \in X$ and $d(x, y) < \delta_\varepsilon$ imply $|f(x) - f(y)| < \varepsilon$. The key point here is that δ_ε depends on ε but not on the points x and y involved. The important result is then that if $f: X \to \mathbf{R}$ is continuous on X, and X is compact, then f is uniformly continuous on X.

Compactness has several equivalent formulations, each appropriate to various applications. The Heine–Borel theorem says that $A \subset \mathbf{R}^n$ is compact iff A is closed and bounded. Here A is *bounded* iff A is contained in some sufficiently large open ball $B_r(p)$. This criterion is extremely easy to use, but it fails for most function spaces, which are usually infinite dimensional. For these more general cases, further conditions are usually required. As an illustration, consider the space $C^0([0, 1], \mathbf{R})$ given in Sec. 1.1. The Arzela–Ascoli theorem says: $A \subset C^0([0, 1], \mathbf{R})$ is compact iff A is closed, bounded, and equicontinuous. Here A is *equicontinuous* iff for each $\varepsilon > 0$ there is a $\delta_\varepsilon > 0$ such that $x, y \in [0, 1]$ and $|x - y| < \delta_\varepsilon$ imply $|f(x) - f(y)| < \varepsilon$ for all $f \in A$. The essential part of this definition is that the same δ_ε works for all $f \in A$. Equicontinuity extends the notion of uniform continuity and is therefore sometimes called *uniform uniform continuity*.

A sequence $x_1, x_2, x_3, \ldots$ of points in a topological space X *converges* to $p \in X$ iff

for each neighborhood U of p there exists an integer N_U such that $x_k \in U$ for all $k \geq N_U$. To make this definition useful, it is necessary to require that the space X be Hausdorff; otherwise a sequence may converge to two distinct points. For example, in the space $X = \{a, b\}$ with the indiscrete topology $\{\emptyset, X\}$, the sequence $a, b, a, b, a, b, a, \ldots$ converges to both a and b. This pathological behavior does not occur in Hausdorff spaces. A sequence $x_1, x_2, x_3, \ldots$ in a metric space M is a *Cauchy sequence* iff for each $\varepsilon > 0$ there exists an integer N_ε such that $d(x_m, x_n) < \varepsilon$ for all $m, n \geq N_\varepsilon$. A metric space M is *complete* iff each Cauchy sequence in M converges to a point of M. A metric space M is *totally bounded* iff for each $\varepsilon > 0$ there exists a finite collection of open balls of radius ε that contains M in their union; i.e., they form a finite open cover of M. A complete normed space is called a *Banach space*. A complete inner-product space is called a *Hilbert space*.

Four equivalent criteria for compactness are given as follows: A is compact iff

1. each open cover of A has a finite subcover (topological version);
2. each sequence in A has a subsequence that converges to a point of A (*sequential compactness* version);
3. A is complete and totally bounded (metric-space version); or
4. each infinite subset of A has an accumulation point in A (*Bolzano–Weierstrass property*).

The following properties are used frequently. Each closed subset of a compact topological space is compact. Each compact subset of a Hausdorff topological space is closed. The Tychonoff theorem states that the Cartesian product of an arbitrary number of compact topological spaces is compact.

An essential tool in the piecing together of local objects to form a global object is a *partition of unity*. Some definitions are needed first. A topological space X is *paracompact* iff X is Hausdorff and each open cover of X has a locally finite refinement. A *locally finite refinement* of an open cover $\{U_i\}_{i \in I}$ is an open cover $\{V_j\}_{j \in J}$ such that each V_j is contained in some U_{i_j} and each point of X is in at most a finite number of the V_j's. Given an open cover $\{U_i\}_{i \in I}$ of the space X, a family of continuous functions $\varphi_i: X \to [0, 1]$ is a *partition of unity* subordinate to the cover $\{U_i\}_{i \in I}$ iff $\mathrm{supp}(\varphi_i) \subset U_i$ for each $i \in I$, and for each $p \in X$, $\varphi_i(p) = 0$ for all but a finite number of indices i and $\Sigma_{i \in I} \varphi_i(p) = 1$. Here $\mathrm{supp}(\varphi)$ denotes the *support* of $\varphi: X \to \mathbf{R}$ and is defined to be the closure of the subset of X on which φ is nonzero, i.e., $\mathrm{cl}\{p \in X | \varphi(p) \neq 0\}$. The basic theorem in this regard is this: Each paracompact topological space admits a partition of unity. This result is important for the theory of manifolds (see Sec. 3.1). It implies that each paracompact manifold can be given a Riemannian metric (see GEOMETRICAL METHODS, Sec. 2.3). Essentially, the coordinate charts on the manifold give induced metrics on parts of the manifold, which are then pieced together via an associated partition of unity.

2 Algebraic Topology

The key idea in algebraic topology is to assign algebraic invariants to topological spaces in such a way that homeomorphic spaces have the same algebraic invariants. It is hoped that the algebraic information given by groups, rings, vector spaces, and homomorphisms is then easier to analyze. For example, if two spaces have different invariants, then they cannot be equivalent.

This gives a method for distinguishing many spaces.

Physically, the most direct motivation for algebraic topology comes from potential theory. The fundamental problem of electrostatics is to find the equilibrium charge-density distribution on the surface of a charged conductor. This difficult problem is replaced by the simpler and in many cases more relevant problem of the determination of the electric field $\mathbf{E}$ produced by such a charge distribution. This field problem is in turn reduced to a function problem via the introduction of a potential function φ. The equation $\mathrm{div}\mathbf{E} = 0$ then becomes the Laplace equation, $\Delta\varphi = 0$, with solution given by the line integral $\varphi(X) = \int_\gamma \mathbf{E} \cdot d\mathbf{l}$, where γ is any smooth path from a fixed point A to the variable point X. The path γ is also required to lie inside the region exterior to the conductor. The point A is usually taken to be a point at infinity or some point assumed to have zero potential. If the field $\mathbf{E}$ is given by $\mathbf{E} = (P, Q, R)$, and the line element $d\mathbf{l}$ is denoted by (dx,dy,dz), then $\varphi(X) = \int_\gamma Pdx + Qdy + Rdz$ expresses the potential as the integral of the differential one-form $\alpha = Pdx + Qdy + Rdz$ over a path from A to X. Further, the condition curl $\mathbf{E} = 0$ implies that the one-form α is closed (i.e., $d\alpha = 0$), and therefore the Stokes theorem implies that the line integral above will give the same value on any smooth path γ' connecting A to X that can be continuously deformed into γ without crossing the surface of the conductor. In such a case the paths γ and γ' are said to be *homotopic* to each other within the region exterior to the conductor. Two examples are useful here. If the conductor is a solid ball, then each pair of paths from arbitrary points A and X and lying in the region exterior to the ball are homotopic to each other in that region. Such a region is called *simply connected*. An equivalent form of this condition is that each closed path (*loop*) beginning and ending at a given point A should be continuously deformable within the region to the point A. For a nonsimply connected example, consider a solid torus (doughnut-shaped) conductor; then a loop at A that passes once through the hole in the torus is not deformable to the point A. This results, in general, in multivalued potential functions.

2.1 Some Basic Tools

Helmholtz's theorem also requires simply connected regions. A vector field $\mathbf{V}$ with curl $\mathbf{V} = 0$ in a simply connected region can be expressed as a gradient of appropriate potential functions, and a vector field $\mathbf{V}$ with $\mathrm{div}\mathbf{V} = 0$ in a simply connected region can be expressed as a curl of appropriate vector potentials. These basic results are simplified and generalized by the theory of differential forms. For simplicity consider the case of $\mathbf{R}^3$ with the usual Cartesian coordinate system. A *0-form* f is simply a smooth real-valued function on $\mathbf{R}^3$, such as $f(x, y, z) = x^2y + \sin(xyz^2)$. A *1-form* α has the form $\alpha = Pdx + Qdy + Rdz$, and a *2-form* β has the form $\beta = Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy$, where P, Q, and R are 0-forms. Finally, a *3-form* γ has the form $\gamma = fdx \wedge dy \wedge dz$, where f is a 0-form. The *wedge product* symbol $\wedge$ is used to emphasize that this multiplication is skew commutative; i.e., $dy \wedge dx = -dx \wedge dy$ and, in particular, $dx \wedge dx = 0$; $dy \wedge dz \wedge dy = 0$, etc. As a consequence, all differential forms of degree higher than the dimension of the underlying space are zero. The properties of these forms are

based on those of determinants where interchanging any two distinct rows reverses the sign of the determinant. Determinants measure signed lengths, areas, volumes, and hypervolumes; differential forms measure infinitesimal signed lengths, areas, volumes, and hypervolumes and, therefore, appear as integrands in integrals giving such quantities. The wedge product generalizes the cross product to spaces of arbitrary dimension. The key operator in differential forms is the exterior derivative d. Letting E^p denote the real vector space of all p-forms on $\mathbf{R}^3$, this derivative maps each p-form α to a $(p+1)$-form $d\alpha$, i.e., $d\colon E^p \to E^{p+1}$. Thus, for each 0-form f, $df = f_x dx + f_y dy + f_z dz$, where the subscripts denote partial derivatives; for each 1-form $\alpha = Pdx + Qdy + Rdz$, $d\alpha = (R_y - Q_z)dy \wedge dz + (P_z - R_x)dz \wedge dx + (Q_x - P_y)dx \wedge dy$; for each 2-form $\beta = Pdy \wedge dz + Qdz \wedge dx + Rdx \wedge dy$, $d\beta = (P_x + Q_y + R_z)dx \wedge dy \wedge dz$; and $d\gamma = 0$ for each 3-form γ. A key property of this derivative is $d^2 = 0$; i.e., $d(d\alpha) = 0$ for each p-form α. A p-form α is *closed* iff $d\alpha = 0$, and α is *exact* iff $\alpha = d\varphi$ for some $(p-1)$-form φ. The $d^2 = 0$ property implies that each exact form is also a closed form. If a vector field $\mathbf{X} = (P, Q, R)$ is identified with the 1-form α above, then curl $\mathbf{X}$ corresponds to $d\alpha$; and if $\mathbf{X}$ is identified with the 2-form β above, then div $\mathbf{X}$ corresponds to $d\beta$. The basic vector analysis identities curl $\nabla\mathbf{X} = 0$ and div curl $\mathbf{X} = 0$ are special cases of $d^2 = 0$. In physics, the vector fields corresponding to 1-forms are called "polar" vector fields, and the vector fields corresponding to 2-forms are called "axial" vector fields. Using this formalism, classic formulas such as the fundamental theorem of calculus:

$$\int_a^b f'(x)dx = f(b) - f(a),$$

Green's theorem:

$$\int\int_G (Q_x - P_y)dx \wedge dy = \int_{\partial G} Pdx + Qdy,$$

the Stokes theorem:

$$\int\int_S (\text{curl}\,\mathbf{X} \cdot \mathbf{n}dS) = \int_{\partial S} \mathbf{X} \cdot d\mathbf{s},$$

and the divergence theorem:

$$\int\int\int_D \text{div}\,\mathbf{X}dV = \int\int_{\partial D} \mathbf{X} \cdot \mathbf{n}dS$$

are special cases of the "generalized Stokes theorem," which has the form

$$\int\int\cdots\int_M d\alpha = \int\cdots\int_{\partial M} \alpha$$

for each n-dimensional orientable manifold M with boundary ∂M and $(n-1)$-form α on M. Thus, integration of the right sort of n-form, namely, an exterior derivative, over a given region reduces to an integral over the boundary of the region. In a sense this results from the fact that whatever enters or leaves a region must cross the boundary. The differential-forms formalism and the above theorems are very directly connected to basic techniques of potential theory and are therefore relevant to electromagnetism and in particular to Maxwell's equations.

The constructions in algebraic topology are designed to measure the topological complexity of a given space relative to the above considerations. For example, if T is a cylindrical tube with boundary ∂T given by two circles C_1 and C_2 with proper orientations, and α is a given closed 1-form on M, then $\int_{C_1} \alpha = \int_{C_2} \alpha$. Thus, C_2 is equivalent to C_1 in these circumstances. Two such spaces that form the boundary of a third space of one

higher dimension are said to be *homologous* and are considered to be equivalent in *homology*. For example, any two disjoint circles on the two-dimensional sphere S^2 form the boundary of an annular region on S^2, and so the homology group (as defined in Sec. 2.3) $H_1(S^2)$ of S^2 is 0. The *de Rham cohomology* of M (see Sec. 3.3) is defined directly in terms of the differential-forms structure as the quotient of the vector space of closed p-forms on M by the vector space of exact p-forms on M for each p and therefore measures the extent to which each closed form can be represented as an exact form. In the more general case, cohomology is defined by taking the appropriate notion of a *dual* of the homology of the space.

Some of the most famous theorems provable via algebraic topological methods are the following (let n be a positive integer, B^n be the unit ball in $\mathbf{R}^n$, and S^n be the unit sphere in $\mathbf{R}^{n+1}$):

Jordan-Brouwer separation theorem: Each compact connected hypersurface X in $\mathbf{R}^n$ divides $\mathbf{R}^n$ into two open sets: the "inside" of X and the "outside" of X. The special case with X a simple closed curve in $\mathbf{R}^2$ is called the Jordan curve theorem. A *hypersurface* in $\mathbf{R}^n$ is an $(n-1)$-dimensional submanifold of $\mathbf{R}^n$.

No-retraction theorem: There is no continuous map from B^n to its boundary S^{n-1}.

Brouwer fixed-point theorem: If $f: B^n \to B^n$ is continuous, then f has a fixed point; i.e., there exists a point $p \in B^n$ such that $f(p) = p$.

Borsuk–Ulam theorem: If $f: S^n \to \mathbf{R}^n$ is continuous, then $f(x) = f(-x)$ for some $x \in S^n$; i.e., f takes the same value on at least one pair of antipodal points.

Ham-sandwich theorem: Let A, B, C be three bounded subsets of $\mathbf{R}^3$ each of which has a volume; then there exists a plane that bisects each of the three sets. (This theorem generalizes to the case of n bounded sets with hypervolume and guarantees the existence of a hyperplane in $\mathbf{R}^n$ that bisects each of them.)

To illustrate via a physical example: Imagine a fluid flowing inside a torus-shaped container in such a way that each point of the fluid is moving counterclockwise around the torus as in Fig. 3; then the Brouwer fixed-point theorem implies that this flow must have a closed (periodic) orbit. Here the B^n involved is a disk placed perpendicular to the central circle of the torus, and the map f associates to each point on this disk the return point obtained by flowing once around the torus. This application is close to one of Poincaré's original motivations for topology. He wanted to prove the existence of periodic orbits in phase-space flows associated with celestial mechanics. The intercepting space is called a *Poincaré section*.

2.2 Homotopy

Intuitively, two paths from point P to point Q and lying in a region R are *homotopic*

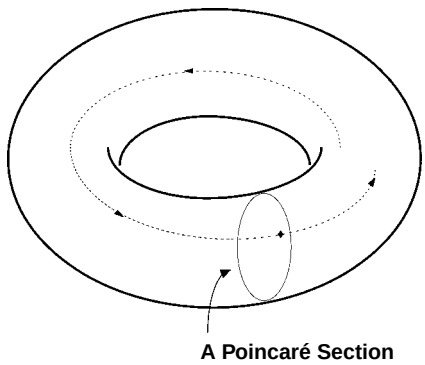


Fig. 3 Placing a Poincaré section in a fluid flow in a torus

iff each can be continuously deformed into the other without leaving the region. This notion clearly depends on the region involved. For example, consider the paths $\gamma_1 =$ the upper semicircle and $\gamma_2 =$ the lower semicircle connecting the points $P = (-1, 0)$ and $Q = (1, 0)$ in the plane $\mathbf{R}^2$. These are homotopic in the region $\mathbf{R}^2$, but they are not homotopic in the punctured plane, $\mathbf{R}^2 - \{(0, 0)\}$. Thus, if the one-form α has no singularities on or inside the unit circle, then $\int_{\gamma_1} \alpha = \int_{\gamma_2} \alpha$; but for the one-form

$$\alpha = \frac{-ydx + xdy}{x^2 + y^2},$$

which has a singular point at the origin, these two line integrals have distinct values. The *fundamental group* measures the complexity of a space in this regard. In Fig. 4 the underlying space is a torus, and the paths γ_1 and γ_2 are homotopic, but neither path is homotopic to the path γ_3.

Let I be the unit interval [0,1]. A *path* from P to Q in a topological space M is a continuous function $f: I \to M$ with $f(0) = P$ and $f(1) = Q$. Two paths f and g from P to Q in M are *homotopic* iff there exists a continuous function $F: I \times I \to M$ such that $F(0, t) = P$, $F(1, t) = Q$, $F(s, 0) = f(s)$, and $F(s, 1) = g(s)$ for all s and t in I. Thus F maps the unit square

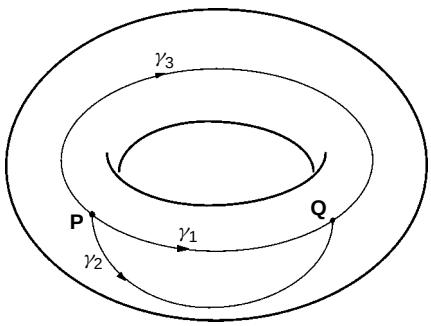


Fig. 4 Path γ_1 is homotopic to γ_2, but not to γ_3

$I \times I$ into M so that all points on the left edge go to P, all points on the right edge go to Q, the bottom edge is mapped to the path f, and the top edge is mapped to the path g. This is illustrated in Fig. 5. This definition makes precise the notion of "continuous deformation" of f to g in M. Notice that all paths are reparametrized if necessary so as to have domain I. To define the fundamental group of M, a single point x_0 in M is fixed, and only *loops*, i.e., paths with initial and terminal points at x_0, are considered. Two loops are considered *equivalent* iff each can be deformed into the other within M. In particular, a loop is *trivial* iff it can be deformed to the basepoint x_0; i.e., it is homotopic to the constant path at x_0. The *fundamental group* of M with *basepoint* x_0 is defined to be the set of equivalence classes of such loops at x_0. If M is path connected, then the fundamental group is independent of the choice of basepoint. The group product operation $*$ in the fundamental group is defined as follows: Given two equivalence classes $[f]$ and $[g]$ of loops at x_0, pick representative loops, f and g, from each class; reparametrize f to have domain [0,1/2] and g to have domain [1/2,1], and define a new loop, denoted $f \cdot g$, which is just f followed by g; then the product is defined by $[f] * [g] = [f \cdot g]$. This operation is independent of the choices made and gives the set of equivalence classes a group structure. If M is path connected and the fundamental group of M is 0, i.e., the group has only one element, then M is said to be *simply connected*. The fundamental group of M is also called the *Poincaré group* or *first homotopy group* of M and is denoted $\pi_1(M, x_0)$ or just $\pi_1(M)$ if M is path connected. Simple examples are $\pi_1(\mathbf{R}^n) = 0$ for all n, $\pi_1(S^1) = \mathbf{Z}$ for the circle S^1, and $\pi_1(S^n) = 0$ for all spheres S^n with $n \geq 2$. The real projective plane has

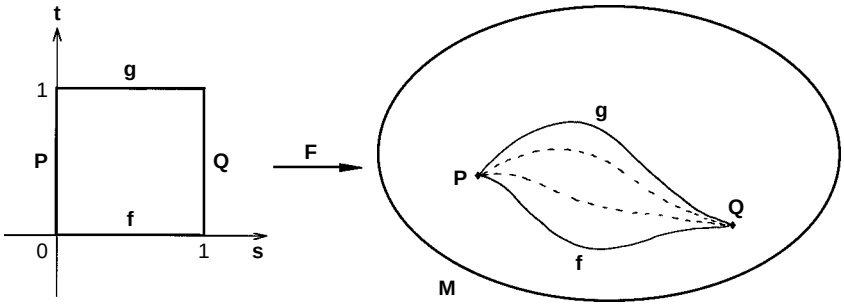


Fig. 5 A homotopy F between paths f and g

fundamental group $\mathbf{Z}_2$ as does the rotation group $R(3) = \mathrm{SO}(3)$. To see this, consider $R(3)$ as a solid ball of radius π in $\mathbf{R}^3$ with each point on its boundary identified to its antipodal point; i.e., each rotation in $\mathbf{R}^3$ with an axis corresponds to a point in the ball along that axis at a distance given by the angle of rotation about that axis. The identity rotation corresponds to the center of the ball, and each point on the boundary is a rotation through π radians and corresponds to a rotation through $-\pi$ radians, which is the antipodal point. The loop starting at the origin, going upward to the North pole, which is identified with the South pole, and then continuing upward to the origin is not homotopic to a point; yet the same loop traversed twice is deformable to a point. This is the topological property of $R(3)$ that is associated to "spin" and "spinor" representations in quantum mechanics.

Each loop with basepoint x_0 can be considered as a map of the unit circle $S^1 \subset \mathbf{R}^2$ into M that sends (1,0) to the point x_0. $\pi_1(M)$ is then the set of equivalence classes of such maps. In an analogous fashion, higher homotopy groups $\pi_n(M)$ are defined to be the sets of equivalence classes of maps of n-spheres S^n into M, which map a chosen fixed point of S^n to the basepoint x_0. These groups detect higher-dimensional "holes" in the given space M. For example, let M be a solid ball of radius 3 with a concentric inner ball of radius 1 removed. M is simply connected, and so $\pi_1(M) = 0$. However, $\pi_2(M) \neq 0$ since no two-sphere that encloses the inner hole can be collapsed to a point inside of M. In general, $\pi_1(M)$ is non-Abelian. However, the higher-dimensional counterparts $\pi_n(M)$ are Abelian groups for $n \geq 2$.

One of the most important unsolved problems in topology is the Poincaré conjecture. Poincaré asked if a compact three-dimensional space with trivial fundamental group had to be homeomorphic to the three-sphere. Higher-dimensional versions of this conjecture have been proved to be true by Smale for all dimensions greater than 4 and by Freedman in dimension 4, yet the original case remains unresolved. This problem continues to play a very important role in the development of topology.

2.3 Homology

There are several ways to define homology groups for a given topological space M. Examples include combinatorial, simplicial, cellular, and singular homology. Each method has some advantages, and for the majority of cases the resulting groups are identical. Simplicial homology

was very popular in the early development of algebraic topology in part because it is easy to visualize spaces as constructed step by step with standard pieces such as points, line segments, triangles, and tetrahedra that are called simplices and are more precisely defined later. Thus, any triangulated surface could be approximated by its associated simplicial complex. For example, a two-sphere is homeomorphic to a (hollow) tetrahedron. Using such simplicial approximations, the homology groups can be computed via linear algebra. Unfortunately, the computations are unwieldy for complexes with large numbers of simplices. Singular homology is more frequently encountered and is the only case treated here. Let $\mathbf{Z}$ be the integers, $\mathbf{R}$ be the real numbers, and $\mathbf{C}$ be the complex numbers.

The *standard p-simplex* Δ_p is defined by

$$\Delta_p = \{(t_0, t_1, \ldots, t_p) \in \mathbf{R}^{p+1} | t_0 + t_1 + \cdots + t_p = 1 \text{ and each } t_i \geq 0\}.$$

See Fig. 6. The *vertices* of Δ_p are $e_0 = (1, 0, 0, \ldots, 0)$, $e_1 = (0, 1, 0, \ldots, 0), \ldots,$ $e_p = (0, 0, 0, \ldots, 1)$, which form the standard orthonormal basis in $\mathbf{R}^{p+1}$. Thus Δ_1 is the one-dimensional line segment from (1,0) to (0,1) in $\mathbf{R}^2$, and Δ_2 is the two-dimensional triangle with vertices at (1,0,0), (0,1,0), and (0,0,1) in $\mathbf{R}^3$. $[e_0, e_1, \ldots, e_p]$ is a convenient notation for Δ_p. A p-simplex with its vertices listed in such an order is called an *oriented p-simplex*. Two oriented p-simplices are *equivalent* iff one can be obtained from the other by an even permutation of the vertices, e.g., $[e_2, e_0, e_1]$ and $[e_0, e_1, e_2]$ are equivalent oriented standard 2-simplices.

A *singular p-simplex* in a topological space X is simply a continuous function $f: \Delta_p \to X$. Thus the points of X may be thought of as the images of singular 0-simplices, and paths in X correspond to the images of singular 1-simplices. The word *singular* is used to indicate that the dimension of the image may be less than the dimension p of the domain Δ_p. So points can in fact be the images of 1-simplices, 2-simplices, etc., and the image of any p-simplex can be regarded as the image of a q-simplex if q is larger than p. The boundary elements of each p-simplex break up into lower-dimensional simplices called *faces*, e.g., the triangle Δ_2 has three one-dimensional faces given by its edges and three zero-dimensional faces given by its vertices. For each $0 \leq i \leq p$, the ith *face* of the singular p-simplex f is the singular $(p-1)$-simplex $F_i(f): \Delta_{p-1} \to X$ defined by $F_i(f)(t_0, t_1, \ldots, t_{p-1}) =$

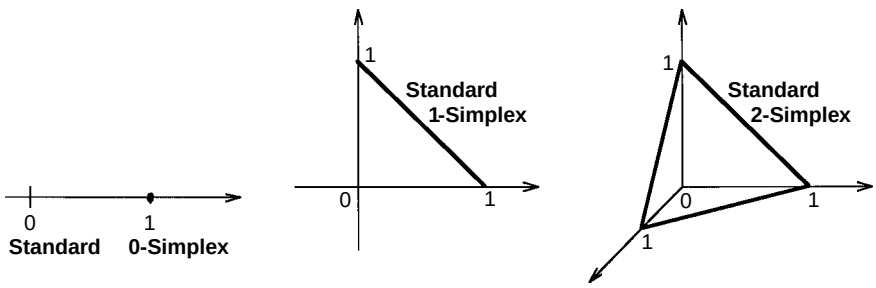


Fig. 6 Standard zero-, one-, and two-dimensional simplices

$f(t_0, t_1, \ldots, t_{i-1}, 0, t_i, \ldots, t_{p-1})$. The *boundary operator* ∂_p is defined on each singular p-simplex f by $\partial_p f = F_0(f) - F_1(f) + \cdots + (-1)^p F_p(f)$. This agrees with the notion of boundary used in the Stokes theorem. To illustrate, let $f\colon \Delta_2 \to X$ be a singular 2-simplex, so that f maps the triangle $[e_0, e_1, e_2]$ in $\mathbf{R}^3$ to X. The boundary $\partial_2 f$ is then given by the combination of three 1-simplices: $\partial_2 f = F_0(f) - F_1(f) + F_2(f)$, where $F_0(f)$ corresponds to the map f restricted to side $[e_1, e_2]$, $F_1(f)$ corresponds to the map f restricted to side $[e_0, e_2]$, and $F_2(f)$ corresponds to the map f restricted to side $[e_0, e_1]$. The negative sign attached to $F_1(f)$ indicates that the orientation is reversed, so that $\partial_2 f$ simply traverses the boundary of f in the expected manner. See Fig. 7.

As seen above, the boundary operators give sums of simplices. The notion of a *p-chain* is introduced to define these correctly as well as to discuss more complicated subsets of X and to develop the algebraic structure of the theory. The pth *singular chain group* $C_p(X)$ of X is defined to be the free Abelian group generated by the set of all singular p-simplices in X; i.e., c is a *singular p-chain* in X iff $c = n_1 f_1 + n_2 f_2 + \cdots + n_k f_k$ for some integers $n_1, n_2, \ldots, n_k$ and singular p-simplices $f_1, f_2, \ldots, f_k$. So a singular p-chain is simply a finite integer-weighted sum of singular p-simplices. Intuitively, imagine such a sum as a collection of parametrized regions over which an integration is performed. An integer weight k indicates that the value of the integral over that region is counted k times; a negative weight indicates that the integral is taken with the opposite orientation, and therefore the sign of the value is reversed. The boundary operator extends to such singular p-chains by $\partial_p c = n_1 \partial_p f_1 + n_2 \partial_p f_2 + \cdots + n_k \partial_p f_k$. This gives a sequence of boundary operators $\partial_p\colon C_p(X) \to C_{p-1}(X)$ defined for each $p \geq 1$. ∂_0 is defined by $\partial_0 c = 0$ for each singular 0-chain c. These operators satisfy $\partial_p \circ \partial_{p+1} = 0$ for all $p \geq 0$; i.e., the boundary of a boundary is always 0 where by convention 0 represents the special degenerate singular chain whose image is the empty set. This important condition is usually abbreviated to $\partial^2 = 0$.

A singular p-chain $z \in C_p(X)$ is called a *p-cycle* in X iff $\partial_p z = 0$. $b \in C_p(X)$ is called a *p-boundary* in X iff $b = \partial_p c$ for some $c \in C_{p-1}(X)$. The group of p-cycles in X is denoted by $Z_p(X)$, and the group of p-boundaries in X is denoted by $B_p(X)$. By convention, $Z_0(X)$ is equal to $C_0(X)$. The condition $\partial^2 = 0$ implies

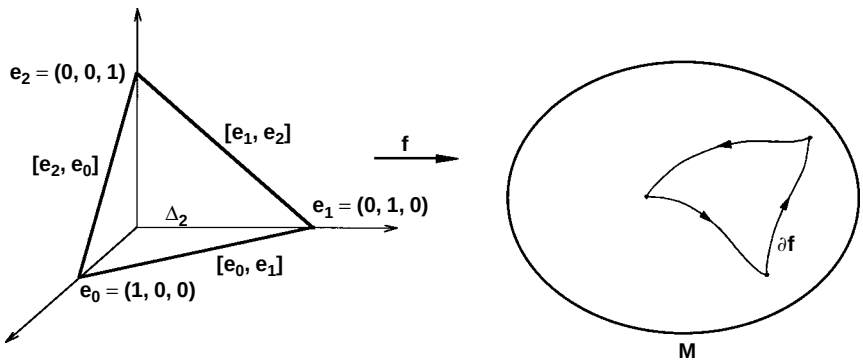


Fig. 7 A singular 2-simplex f and its boundary ∂f

that each p-boundary is also a p-cycle, and so $B_p(X) \subset Z_p(X)$ for each p. For each $p \geq 0$ the pth *singular homology group with integer coefficients* of the topological space X is defined to be the factor group $H_p(X, \mathbf{Z}) = Z_p(X)/B_p(X)$. Thus a sequence of Abelian groups is assigned to each space X. "Integer coefficients" refers to the integer weights in the chain groups; other coefficient groups may also be used.

Here are a few famous examples. Assume $n \geq 1$.

For n-dimensional real spaces $\mathbf{R}^n$,

$$H_p(\mathbf{R}^n, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0, \text{ and 0 otherwise.}$$

For the n-dimensional spheres S^n,

$$H_p(S^n, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0 \text{ or } n, \text{ and 0 otherwise.}$$

For the real projective plane, $\mathbf{R}P^2$:

$$H_p(\mathbf{R}P^2, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0, \mathbf{Z}_2 \text{ if } p = 1, \text{ and 0 otherwise.}$$

Here $\mathbf{Z}_2$ is the Abelian group with exactly two elements. In general, $\mathbf{Z}_n$ denotes the cyclic group with exactly n elements. For the two-dimensional Klein bottle K:

$$H_p(K, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0, \mathbf{Z} \oplus \mathbf{Z}_2 \text{ if } p = 1, \text{ and 0 otherwise.}$$

For the two-dimensional torus $T = S^1 \times S^1$:

$$H_p(T, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0 \text{ or } 2, \mathbf{Z}^2 = \mathbf{Z} \oplus \mathbf{Z} \text{ if } p = 1, \text{ and 0 otherwise.}$$

For the two-dimensional g-holed torus T_g:

$$H_p(T_g, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0 \text{ or } 2, \mathbf{Z}^{2g} \text{ if } p = 1, \text{ and 0 otherwise.}$$

Here $\mathbf{Z}^{2g}$ denotes the direct sum of $2g$ copies of $\mathbf{Z}$. The nonnegative integer g is called the *genus* of T_g. For the n-dimensional tori $T^n = S^1 \times \cdots \times S^1$:

$$H_p(T^n, \mathbf{Z}) = Z^{C(n,p)} \text{ for all } 0 \leq p \leq n, \text{ and 0 otherwise,}$$

where $C(n, p)$ is the binomial coefficient.

For the n-dimensional real projective spaces, $\mathbf{R}P^n$:

$$H_p(\mathbf{R}P^n, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0 \text{ or } n \text{ with } n \text{ odd, } \mathbf{Z}_2 \text{ if } p \text{ is odd and } 0 < p < n, \text{ and 0 otherwise.}$$

For the $2n$-dimensional complex projective spaces, $\mathbf{C}P^n$:

$$H_p(\mathbf{C}P^n, \mathbf{Z}) = \mathbf{Z} \text{ if } p = 0, 2, 4, \ldots, 2n, \text{ and 0 otherwise.}$$

In most cases of interest such as compact manifolds, the above homology groups are *finitely generated Abelian groups*; i.e., they can be put into the *cyclic normal form*

$$\mathbf{Z}^m \oplus \mathbf{Z}_{t_1} \oplus \mathbf{Z}_{t_2} \oplus \cdots \oplus \mathbf{Z}_{t_k}$$

for some integer $m \geq 0$ and integers $t_1, t_2, \ldots, t_k$ such that each $t_i > 1$ and t_i divides t_{i+1} for each $1 \leq i < k$. Here $\mathbf{Z}^m$ denotes the direct sum of m copies of $\mathbf{Z}$, and $\mathbf{Z}_n$ denotes the cyclic group of order n.

For $H_p(X, \mathbf{Z})$ in the above form, the integer m is called the pth *Betti number*, β_p, of X, and $t_1, t_2, \ldots, t_k$ are called the pth *torsion coefficients* of X, and the number k of such torsion coefficients is denoted τ_p. The *Euler – Poincaré characteristic* $\chi(X)$ is given by the alternating sum of the Betti numbers of X:

$$\chi(M) = \sum_{p \geq 0} (-1)^p \beta_p.$$

The homology groups, Betti numbers, torsion coefficients, and Euler–Poincaré characteristic are topological invariants of the space X; i.e., these invariants are the same for homeomorphic spaces. However, these invariants may also be the same for some nonhomeomorphic spaces such as $\mathbf{R}$ and $\mathbf{R}^2$. So these invariants are generally more useful in proving spaces are not homeomorphic than in proving them homeomorphic. If M and N are compact manifolds of dimension m and n, respectively, then $\chi(M \times N) = \chi(M) \times \chi(N)$, and if $m = n$, then $\chi(M \# N) = \chi(M) + \chi(N) - [1 + (-1)^n]$, where $M \times N$ is the Cartesian product and $M \# N$ is the connected sum of M and N.

In the 1920s Lefschetz defined the *relative homology* groups of a *pair* (X,A) of topological spaces with $A \subset X$. Intuitively, this corresponds to the homology of the space obtained from X by collapsing A to a point. Formally, the group of *relative p-cycles mod A* is $Z_p(X, A) = \{c \in C_p(X) | \partial_p c \in C_{p-1}(A)\}$, and the group of *relative p-boundaries mod A* is $B_p(X, A) = \{c \in C_p(X) | c - c' \in B_p(X)$ for some $c' \in C_p(A)\}$. The pth *relative singular homology group of the pair* (X,A) is then defined to be the factor group $H_p(X, A) = Z_p(X, A)/B_p(X, A)$ for each $p \geq 0$. One very important consequence of relative homology is the *long exact sequence in homology*. For each pair (X,A) two inclusion maps of pairs are naturally defined: $i: (A, \emptyset) \to (X, \emptyset)$ and $j: (X, \emptyset) \to (X, A)$. These induce natural maps i_* and j_* in homology, and combining these with appropriate induced boundary operators ∂_* yields the long exact homology sequence:

$$\cdots \longrightarrow H_{p+1}(X, A) \xrightarrow{\partial_*} H_p(A) \xrightarrow{i_*} H_p(X) \xrightarrow{j_*} H_p(X, A) \xrightarrow{\partial_*} H_{p-1}(A) \longrightarrow \cdots$$

Such a sequence $\cdots \to G \xrightarrow{f} H \xrightarrow{g} K \to \cdots$ of groups connected by group homomorphisms is *exact* at the group H iff the image of f equals the kernel of g; i.e., $f(G) = \{a \in H | g(a) = e_K\}$, where e_K is the identity element in the group K. Thus, the long exact sequence in homology is exact at each group in the sequence. Further, each continuous map $f: (X, A) \to (Y, B)$ between pairs induces natural homomorphisms between their associated long exact sequences. This structure is very important in the theoretical development and application of algebraic topology in general. Similar long exact sequences also exist for homotopy and cohomology. To illustrate briefly a typical use of the structure of exact sequences, consider such a sequence

$$\cdots \longrightarrow F \xrightarrow{f} G \xrightarrow{g} H \xrightarrow{h} K \longrightarrow \cdots.$$

If the group F is the zero group, then the map g must be injective; if the group K is the zero group, then the map g must be surjective; thus if both F and K are zero groups, then the map g is an isomorphism, and the groups G and H are isomorphic.

2.4 Cohomology

Cohomology groups are additional algebraic structures associated to a given space. Like homology, there are various methods for defining and computing cohomology such as simplicial, cellular, singular, Čech, and de Rham. Singular cohomology proceeds by constructing dual objects called cochain groups and their associated coboundary operators and performing quotient group operation as in homology. As usual, for the spaces typically found in physics, such as manifolds, all of these methods produce essentially the same results. The cohomology groups of a space

are dual to the homology groups in a sense similar to that of the duality between Dirac's bra and ket vectors, or between contravariant and covariant tensors. The most commonly encountered cohomology in physics is the de Rham cohomology associated with the differential forms on a manifold, and this is discussed in Sec. 3.

3
Differential Topology

Differential topology is the area of topology most directly associated with physics. The basic objects are smooth manifolds, and the functions of interest are smooth maps between such manifolds. Vector fields, differential forms, differential operators, and integration are naturally the key tools. The subject is very closely tied to differential geometry and global analysis, and these subjects usually appear together in applications. This section has much in common with the article GEOMETRICAL METHODS. This section treats manifolds, bundles, vector fields and differential forms, de Rham cohomology, and Morse theory. Manifolds are the natural answer to the question: What sort of spaces should be the domains and ranges of functions? It is clear that higher-dimensional spaces are needed for problems involving several variables such as dynamical systems. Also spaces with curvature are necessary to treat surfaces in Euclidean spaces as well as general relativity. Coordinate systems are required in order to apply the techniques of calculus and analysis. These allow problems in the space to be transferred to subsets of $\mathbf{R}^n$, analyzed, and then mapped back to the original space. Manifolds are spaces designed with these objectives in mind. Bundles, more precisely called fiber bundles, are structures built on manifolds. They answer the question: Where are objects such as vector fields, differential forms, and general tensor fields located? Perhaps the most familiar example in physics is phase space. Here the underlying "base" manifold is configuration space: the set of allowable positions or generalized positions for the system. A simple planar pendulum, for example, has the circle S^1 as its configuration space and phase space given by an infinite cylinder. This infinite cylinder is the "cotangent" bundle of S^1. The double planar pendulum has configuration space given by a two-dimensional torus T, and its phase space is the cotangent bundle of T, which is four-dimensional. Two of these dimensions give the position of the pendulum bobs, and the remaining two give their momenta. If the generalized momenta are replaced by generalized velocities, then the appropriate bundle is the tangent bundle of the configuration space M. Velocities are given by vector fields on M and are therefore located in the tangent bundle, whereas momenta are given by 1-forms on M and so are located in the cotangent bundle.

The topology of a manifold M places restrictions on the behavior of the functions, vector fields, and differential forms defined on M. For example, each smooth tangent vector field on a 2-sphere must be zero at some point. Intuitively, at each point in time there must be some point on the Earth at which the (horizontal component of the) wind velocity is zero. The Poincaré – Hopf index theorem makes this relation precise. Similarly, the theorems of de Rham and Hodge use de Rham cohomology to quantify the topological restrictions on the differential forms on manifolds. One of most beautiful areas of differential topology is Morse theory, which Marston Morse

developed in connection with his study of geodesics on manifolds. The original motivation for this work came from the work of Poincaré and Birkhoff in dynamical systems where the motion of the system is described by geodesics. Morse theory quantifies the topological restrictions on the critical points of smooth nondegenerate functions. For example, the theory guarantees that each smooth nondegenerate function on a g-holed torus must have at least $2g + 2$ critical points.

Characteristic classes are specific cohomology classes assigned to bundles. They were designed to answer two questions: (1) When is a bundle trivial, i.e., a simple Cartesian product of the base manifold and the fiber? and (2) When are two bundles equivalent? An important area of application is the Atiyah – Singer index theorem that quantifies the number of solutions of certain operators on manifolds. Although mentioned briefly in this article, these topics require more extensive machinery than is reasonable in this discussion. See Sec. 4 for appropriate references.

3.1
Manifolds and Bundles

The notion of a manifold is a generalization of that of a surface in $\mathbf{R}^3$ such as a two-dimensional sphere. It is not possible to cover the sphere with a single coordinate system; at least two are needed. If each coordinate system is viewed as an observer, then in the regions in which their coordinate systems overlap, they should be able to compare data, i.e., effectively transform each coordinate system into that of any overlapping system. Since not only positions but also velocities and accelerations should be transformable, the changes of coordinates are required to be *smooth;* i.e., the functions and all their derivatives of all orders are required to be continuous. Typical examples are the configuration and phase spaces used in classical mechanics. These are usually higher-dimensional manifolds.

An *n-dimensional differentiable manifold* M is a locally Euclidean topological space. More precisely, M has a covering by *coordinate charts*, which are homeomorphisms $\varphi_i\colon U_i \to \varphi_i(U) \subset \mathbf{R}^n$ for each $i \in I$ such that each U_i is an open subset of M, $M \subset U_{i\in I} U_i$, and on the overlap between the images of each pair of charts, the function $\varphi_j \circ \varphi_i^{-1}\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ is a smooth function. See Fig. 8. Coordinate charts are also called coordinate systems, local charts, or just charts. On a manifold, calculus is done by transferring the process back to $\mathbf{R}^n$ via the coordinate charts, performing the appropriate calculus operations, and then transferring the result back to the manifold. For example, a function $f\colon M \to N$ between manifolds M of dimension m and N of dimension n is smooth iff $\psi \circ f \circ \varphi^{-1}$ is smooth as a map from an open set of $\mathbf{R}^m$ to an open set of $\mathbf{R}^n$ for each pair of charts φ on M and ψ on N. A smooth function $\varphi\colon M \to N$

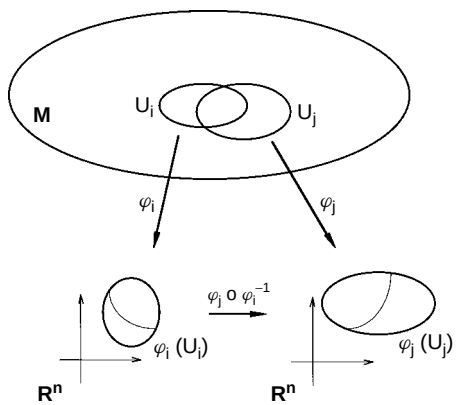


Fig. 8 Manifold charts ϕ_i and ϕ_j must have smooth transition functions $\phi_j \circ \phi_i^{-1}$ on overlaps $\phi_i(U_i \cap U_j)$

with a smooth inverse is called a *diffeomorphism*. A manifold M is *orientable* iff a covering by coordinate charts $\{\varphi_i\}_{i\in I}$ can be chosen for M in such a way that the determinants of the Jacobian matrices of the overlap maps $\varphi_j \circ \varphi_i^{-1}$ are positive for each $i, j \in I$. Manifolds for which such a covering does not exist are called *nonorientable*. Among two-dimensional surfaces, the sphere and g-holed tori are orientable, whereas the Klein bottle and real projective plane are nonorientable. Very frequently, manifolds are assumed to be Hausdorff or even paracompact topological spaces, and this is the case for almost all cases typically encountered. If, in the above definition of differentiable manifold, $\mathbf{R}^n$ is replaced by $\mathbf{C}^n$ and the overlap maps $\varphi_j \circ \varphi_i^{-1}$ are required to be complex analytic, then the space M is called an *n-dimensional complex manifold*, and the tools of complex analysis are appropriate. *Manifolds with boundary* are defined by allowing two kinds of coordinate charts: the usual manifold style charts as given originally, plus special charts adapted for boundary points as follows. Let $\mathbf{R}^n_+ = \{(x_1, x_2, \ldots, x_n) | x_n \geq 0\}$ be the standard *upper half-space* in $\mathbf{R}^n$ with the induced topology. Then each point of a *manifold with boundary* is required to be homeomorphic to an open subset of either $\mathbf{R}^n$ or $\mathbf{R}^n_+$, and where charts overlap, their transition functions must be smooth. In Fig. 9 the manifold with boundary is a torus with a disk removed at the top. Each point such as P on the boundary requires a half-space chart as illustrated by ψ. The points such as Q in the interior of M have usual manifold-style charts. The boundary of M is denoted ∂M. Typical examples are the Möbius strip and the upper hemisphere of the 2-sphere, each of which has boundary homeomorphic to the circle S^1, and the closed ball in $\mathbf{R}^n$ with boundary given by the $(n-1)$-sphere S^{n-1}. In

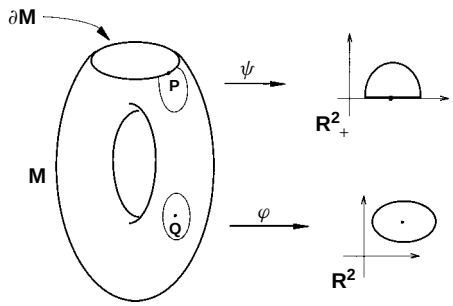


Fig. 9 Manifolds with boundary require "half-space charts" ψ for boundary points such as P

discussions involving manifolds both with and without boundary, manifolds without boundary are often called *closed manifolds*.

Although the initial motivation for manifolds came from spaces such as surfaces sitting in Euclidean spaces, the definition of a manifold does not assume this, and so it is an interesting question whether such abstractly given manifolds can be considered as embedded in $\mathbf{R}^n$ for n sufficiently large. The Whitney embedding theorem guarantees that this is the case, i.e., each paracompact manifold of dimension $n \geq 1$ has an embedding into $\mathbf{R}^{2n}$. Thus the Klein bottle always intersects itself when considered as a subset of $\mathbf{R}^3$, but it can be placed in $\mathbf{R}^4$ without such a self intersection.

Manifolds are the natural choice for the complicated geometrical structures that occur in physics. In addition to the configuration and phase spaces mentioned above, Minkowski spacetime, the universe of general relativity if singularities are ignored, and all Lie groups such as the rotation groups, Lorentz groups, unitary groups, and Poincaré groups are manifolds. However, structures such as vector fields, differential forms, and other tensors defined on such spaces require more general constructions, namely, bundles built

on such manifolds. Bundles (see GEOMETRICAL METHODS) are special manifolds that are generalizations of Cartesian products. $\mathbf{R}^2$ is the Cartesian product of $\mathbf{R}$ and $\mathbf{R}$ and so has a global product structure. In a bundle only a local product structure is required. The most commonly used bundles on a manifold M are the tangent bundle TM, the cotangent bundle T^*M, the exterior product bundles $\boldsymbol{\Lambda}^p M$, and more generally the tensor bundles $T^r_s M$. These are vector bundles; i.e., the fibers are vector spaces and the sections of these bundles give the vector fields, differential forms, and tensors on M. A *section* of a bundle is simply a smooth map that assigns to each point of the base manifold an element of the fiber over that point. Thus, a section of the tangent bundle of M is a (tangent) vector field on M. The two basic formulations of classical mechanics, Lagrangian and Hamiltonian, are associated to the bundles TM and T^*M, respectively. For example, if M is the configuration space of the system, then the phase space is the cotangent bundle T^*M, and the Hamiltonian is a map $H: T^*M \to \mathbf{R}$. In gauge theories vector bundles are replaced by the more general principal fiber bundles in which each fiber is a Lie group appropriate to the symmetries of the fields involved. Characteristic classes are special cohomology classes giving topological information about bundles. These can be used, for example, to determine whether a given bundle has a global product structure, or to distinguish various bundles since bundles with different characteristic classes must be different. The Euler class, Stiefel–Whitney classes, Pontryagin classes, and Chern classes are standard characteristic classes. Secondary characteristic classes such as the Chern–Simons classes also play a role in gauge theories and mechanics.

3.2 Vector Fields and the Poincaré–Hopf Index Theorem

Viewed in terms of the above considerations, a (tangent) vector field on a manifold M is a section of the tangent bundle of M. Each point of M at which a vector field X is zero is called a *zero* of X, and the set of such zeros is denoted Z_X. If a zero p of X is isolated, i.e., p is the only zero of X in some neighborhood of p, then an integer *index* $i_X(p)$ can be assigned to p. Consider first the vector field X in $\mathbf{R}^2$ defined by $X(x, y) = (x^2 - y^2, 2xy)$. X has the origin, $p = (0, 0)$, as its only zero. Let C be the circle of radius 1 centered at p, and consider the normalized unit vector field $U = X/\|X\|$ restricted to C. As C is traversed once counterclockwise, the tip of U performs exactly two counterclockwise rotations about its origin; therefore the index of X is 2. This X is not the field of a dipole, but it does have the same flow lines. Similarly, $X(x, y) = (x^3 - 3xy^2, 3x^2y - y^3)$ has the same flow lines as a quadrupole and index 3. The vector field $X(x, y) = (x, -y)$ has the flow lines of a saddle and index -1. For the case of a general planar vector field $X(x, y) = (P(x, y), Q(x, y))$ with an isolated zero at the origin p, the index may be computed by the formula

$$i_X(p) = \frac{1}{2\pi} \int_C \frac{PdQ - QdP}{P^2 + Q^2},$$

where C is a circle centered at p with radius small enough so that p is the only zero of X inside C, and C is traversed counterclockwise. In the n-dimensional case, the index is defined to be the degree of the map of a small $(n-1)$-dimensional sphere S^{n-1}, centered at p and containing no other zero of X, to itself produced by appropriately normalizing X restricted to this S^{n-1}. This index is always an integer

and is independent of the choice of radius of the sphere provided p is the only zero of X interior to that sphere.

The Poincaré–Hopf index theorem states that if X is a vector field having only isolated zeros on a compact orientable manifold M of dimension n, then

$$\sum_{p \in Z_x} i_X(p) = \chi(M).$$

Thus, among the orientable surfaces, the sum of the indices of the zeros of such a vector field must be 2 for the sphere and $2 - 2g$ for the g-holed torus. In particular, only the usual torus can admit a nonzero (tangent) vector field.

3.3 Differential Forms and de Rham Cohomology

It is hard to overemphasize the importance of differential forms in differential geometry and topology. Although they are just completely alternating covariant tensor fields on manifolds and are therefore part of the general theory of tensors, the simplicity of their structure and operations plus the direct connection with the underlying geometry of the space have given them a very special place in these subjects. Examples of forms in $\mathbf{R}^3$ were given in Sec. 2.1 above. See GEOMETRICAL METHODS for the more general case as well as definitions of the exterior derivative operator d, the Hodge star operator *, and the codifferential operator δ, which is the adjoint of d with respect to the inner product $(,)$ defined by $(\alpha, \beta) = \int_M \alpha \wedge^* \beta$ for each pair of p-forms α and β on M. Here E^pM denotes the vector space of all p-forms on M. A Riemannian or pseudo-Riemannian metric is required for the definition of both * and δ. If M has dimension n, then $d: E^pM \to E^{p+1}M, ^*: E^pM \to E^{n-p}M$, and $\delta: E^pM \to E^{p-1}M$ for each p. The $d^2 = 0$ condition again holds, i.e., $d(d\alpha) = 0$ for all $\alpha \in E^pM$, as well as $\delta^2 = 0$. The Laplace–Beltrami operator $\Delta: E^pM \to E^pM$ is then defined by $\Delta = d\delta + \delta d$. This generalizes the usual Laplace operator $\Delta = \nabla^2$, which is the divergence of the gradient on functions. The Laplace operator certainly has a claim to being the most important and the most studied operator in mathematical physics. It occurs in the Laplace equation, Poisson equation, heat equation, wave equation, wave-equation forms of Maxwell equations, and Schrödinger equation among others. The underlying reason is that it measures the deviation from equilibrium. So it is very natural to seek a generalization of this operator. However, it is important to be careful about the sign of the operator. The usual conventions for the Hodge star operator give the negative of the divergence of the gradient. This agrees with Maxwell's convention; he called this Δf the *concentration* of f. As a result of this convention the spectrum (set of eigenvalues) of Δ on a compact manifold is nonnegative. Δ is then a nonnegative, elliptic, self-adjoint, second-order operator on each E^pM. Also, Δ acting on one-forms corresponds to the negative of the usual Laplacian, $\nabla \mathrm{div} X - \mathrm{curl}\ \mathrm{curl}\ X$, on vector fields X as used in vector analysis. It is important to notice that since a metric is needed to define the Hodge star operator *, and therefore also the codifferential operator δ, the Laplace–Beltrami operator involves the geometry of the manifold as well as its topology.

Again, $\alpha \in E^pM$ is *closed* iff $d\alpha = 0$, and $\alpha \in E^pM$ is *exact* iff $\alpha = d\varphi$ for some $\varphi \in E^{p-1}M$. Setting $Z^p(M)$ equal to the set of closed p-forms, and $B^p(M)$ equal to the set of exact p-forms, $d^2 = 0$ implies that $B^p(M) \subset Z^p(M)$, and the *de Rham cohomology* vector spaces are defined by

$H^p_{\mathrm{dR}}(M, \mathbf{R}) = Z^p(M)/B^p(M)$ for each p. The *Poincaré lemma* says that each p-form that is closed in a neighborhood equivalent to an open ball is exact in that neighborhood; the de Rham cohomology measures the extent to which this holds globally on the manifold M. The de Rham theorem says that this cohomology agrees with that derived via the more topological methods such as singular and Ĉech. Further, on each compact orientable manifold M of dimension n, the Hodge star operator induces an isomorphism called *Poincaré duality* between $H^p_{dR}(M, \mathbf{R})$ and $H^{n-p}_{dR}(M, \mathbf{R})$ for each p. The Hodge theorem gives an orthogonal decomposition $E^pM = \mathscr{H}^pM \oplus dE^{p-1}M \oplus \delta E^{p+1}M$, where $\mathscr{H}^pM$ denotes the set $\{\alpha \in E^pM | \Delta\alpha = 0\}$ of *harmonic* p-forms, $dE^{p-1}M$ is the set of *exact* p-forms, and $\delta E^{p+1}M$ is the set of *coexact* p-forms on M. Further, the dimension of the space of harmonic p-forms is $\beta_p(M)$, the pth Betti number of M. The Hodge theorem is a generalization of Helmholtz's theorem that represents vector fields in $\mathbf{R}^3$ as sums of curls and gradients. More sophisticated global theorems such as the Hirzebruch–Riemann–Roch theorem and the general Atiyah–Singer index theorem utilize the theory of characteristic classes.

3.4 Morse Theory

Morse theory relates the critical-point structure of a smooth function on a manifold to the topology of the manifold. Given a smooth function $f: A \to \mathbf{R}$ where A is an open subset of $\mathbf{R}^n$, a point $p \in A$ is a *critical point* of f iff

$$\frac{\partial f}{\partial x_i}(p) = 0$$

for each $i = 1, 2, \ldots n$; i.e., the gradient of f is 0 at p. Such a critical point of f is *nondegenerate* iff the determinant of the Hessian matrix $Hf(p)$ is nonzero. The *Hessian* $Hf(p)$ is the symmetric matrix of second derivatives of f evaluated at p; i.e.

$$(Hf(p))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(p).$$

The *index* k of a critical point p is the number of negative eigenvalues of $Hf(p)$. For example, $f(x) = 3x^5 - 5x^3 + 1$ has a degenerate critical point at 0, and nondegenerate critical points at -1 and 1 with indices 1 and 0, respectively. The Rosenholtz function $f(x, y) = 4x^2e^y - 2x^4 - e^{4y}$ has exactly two critical points, located at $(-1, 0)$ and (1,0), both nondegenerate and of index 2. This remarkable function has two local maxima and no other critical points; its graph has two mountains but no saddle points. Index-0 critical points correspond to local minima, and critical points of maximal index (equal to the dimension of the domain of the function) correspond to local maxima; the critical points of intermediate index behave like saddle points.

The Morse lemma states that for each nondegenerate critical point p of index k of f there is a neighborhood U of p and a set of coordinates $\{t_1, t_2, \ldots, t_n\}$ in U such that f can be expressed as $f(t_1, t_2, \ldots, t_n) = f(p) - t_1^2 - t_2^2 - \cdots - t_k^2 + t_{k+1}^2 + \cdots + t_n^2$ in U. So f has a *normal form* in some neighborhood of each of its nondegenerate critical points. This implies that each nondegenerate critical point is isolated. These local notions easily extend to smooth real-valued maps on manifolds, although it is worth noting that in general a Riemannian metric is needed to define the Hessian at noncritical points. A standard example is the height function f defined

on a two-dimensional torus M placed on end on the x-y plane in $\mathbf{R}^3$ as in Fig. 10. The value $f(p)$ is just the height (the z coordinate) of the point p above the $z = 0$ plane. This function has four critical points located at a, b, c, and d, all nondegenerate: the bottom point a with index 0, the two saddle points b and c each of index 1, and the top point d with index 2. For the torus laid flat on the plane, the height function has two circles of critical points at top and bottom, and each of these critical points is degenerate; however, the slightest tilting of the torus from this position is sufficient to give a height function with all nondegenerate critical points. A smooth function $f: M \to \mathbf{R}$ is a *Morse function* on M iff each critical point of f is nondegenerate. The Morse functions are dense in the set of smooth functions on M in the following sense: Given any smooth function $f: M \to \mathbf{R}$, there exist arbitrarily nearby Morse functions.

Morse, following the lead of his advisor G. D. Birkhoff, found a sequence of inequalities giving topological restrictions

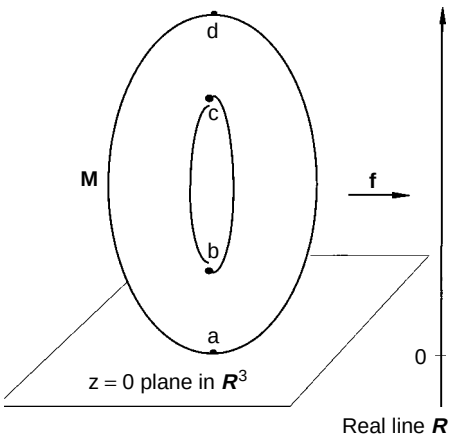


Fig. 10 The height function f has critical points at a, b, c, and d. The minimum at a has index 0, the saddle points at b and c have index 1, and the maximum at d has index 2

on the number c_k of critical points of index k for a general Morse function f on a compact manifold M of dimension n. Letting β_p denote the pth Betti number of M, the *Morse inequalities* are

$$c_0 \geq \beta_0$$

$$c_1 - c_0 \geq \beta_1 - \beta_0$$

$$c_2 - c_1 + c_0 \geq \beta_2 - \beta_1 + \beta_0$$

$$\vdots$$

$$\sum_{0 \leq i \leq p} (-1)^{p-i} c_i \geq \sum_{0 \leq i \leq p} (-1)^{p-i} \beta_i$$

$$\vdots$$

$$\sum_{0 \leq i \leq n} (-1)^{n-i} c_i = \sum_{0 \leq i \leq n} (-1)^{n-i} \beta_i.$$

Notice that the last relation is in fact always an equality. Thus the alternating sum of the *critical numbers* c_p of f equals the Euler–Poincaré characteristic $\chi(M)$. These relations immediately imply $c_p \geq \beta_p$ for each p. These inequalities were extended by Pitcher to include the influence of the torsion coefficients of M as follows: For each p,

$$c_p \geq \beta_p + \tau_p + \tau_{p-1},$$

where τ_p denotes the number of torsion coefficients in the cyclic normal form of the pth homology group $H_p(M, \mathbf{Z})$, as discussed in Sec. 2.3. As a simple application, consider a surface M sitting in $\mathbf{R}^3$. For almost all points p not on M, the distance function from p to each of the points of M is a Morse function on M (the other points are called *focal points*), and the critical points of this function correspond to lines from p to M that are perpendicular to M. Thus the Morse inequalities imply that the number of normals to M from any

nonfocal point p is greater than or equal to the sum of the Betti numbers of M.

4
A Brief Guide to Further Reading

Topology has a vast literature with texts ranging from very elementary to very specialized. This section is designed to select a relatively small number of readily available books in each of the areas covered in this article. These books are usually intended to be used in courses and therefore are rather condensed. More physics-oriented texts such as Felsager (1983), Nash and Sen (1983), von Westenholz (1981), Eguchi et al. (1980), Nash (1991), Göckeler and Schücker (1987), and Nakahara (1990) are demanding texts, but they usually keep physical applications in the forefront. Browsing through these texts for an initial orientation is strongly recommended.

4.1
Point-Set Topology

Many of the classic texts on point-set topology were designed to provide a foundation for analysis. Good examples are Simmons (1963), Wilansky (1970), and Kelley (1955). Texts by Munkres (1975), Sieradski (1992), Jänich (1984), Dugundji (1966), and Hocking and Young (1961) also include discussion of algebraic-topological topics such as homotopy and homology. Most algebraic-topology texts include brief discussions of point-set topology topics. Armstrong (1983) and Kinsey (1993) are good examples at an undergraduate level. Most functional-analysis texts also contain relevant material; see, for example, Naylor and Sell (1982), Kreyszig (1978), Reed and Simon (1980), and Lang (1993).

4.2
Algebraic Topology

Seifert and Threlfall (1980), Veblen (1931), Cairns (1961), and Wallace (1968) are older texts that treat simplicial homology. Armstrong (1983), Kinsey (1993), and Jänich (1984) are good undergraduate-level introductions. Munkres (1984), Greenberg and Harper (1981), Rotman (1988), Massey (1991), Dold (1980), and Spanier (1981) are popular graduate-level texts. Hilton and Wiley (1960), Mac Lane (1963), and Vick (1994) treat homology theory, and Hu (1959) is a basic early reference for homotopy theory.

4.3
Differential Topology

Guillemin and Pollack (1974), Milnor (1965), Dodson and Poston (1991), and Wallace (1968) are good introductions. Bröcker and Jänich (1982), Hirsch (1976), Kosinski (1993), Lang (1995), and Warner (1983) are popular graduate-level texts. Springer (1981) and Weyl (1955) are classic introductions to Riemann surface theory. Steenrod (1951), Chern (1979), Kobayashi and Nomizu (1963, 1969), Gilkey (1995), Bott and Tu (1982), Milnor and Stasheff (1974), Hirzebruch (1966), and Husemoller (1994) are standard references for bundles and characteristic classes. Milnor (1963), Hirsch (1976), Bott (1994), and Kosinski (1993) discuss Morse theory. Eguchi et al. (1980), Gilkey (1995), Nash (1991), and Nakahara (1990) are good references for the Atiyah–Singer index theorem. Choquet-Bruhat et al. (1982, 1989) is a good general reference that covers a wide range of topics in differential topology and geometry particularly relevant for physics. Chern (1979), Wells

(1980), and Hübsch (1992) treat complex manifolds.

4.4 Physical Applications

Flanders (1963) is the classic reference for differential forms. Schutz (1980) and von Westenholz (1981) are recommended for more advanced applications. Felsager (1983), Nash and Sen (1983), Eguchi et al. (1980), Nash (1991), Göckeler and Schücker (1987), and Nakahara (1990) are highly recommended for physicists. Birkhoff (1927) is a classic text in dynamical systems; more recent mathematically oriented texts are Arnol'd (1989), Abraham and Marsden (1978), and Abraham et al. (1988) for classical mechanics and dynamical systems, and Atiyah (1990), Green et al. (1987), and Hübsch (1992) for relations to string theories.

Glossary

Accumulation Point: (see Sec. 1.1)
Balls: (see Introduction)
Banach Space: (see Sec. 1.3)
Basepoint: (see Sec. 2.2)
Betti Numbers: (see Sec. 2.3)
Bolzano–Weierstrass Property: (see Sec. 1.3)
Boundary: (see Sec. 1.1)
Boundary Operator: (see Sec. 2.3)
Boundary Point: (see Sec. 1.1)
Bounded Subset: (see Sec. 1.3)
Bundle: (see Sec. 3.1)
Cartesian Product: (see Introduction)
Cauchy Sequence: (see Sec. 1.3)
Chain: (see Sec. 2.3)
Chain Group: (see Sec. 2.3)
Charts: (see Sec. 3.1)
Closed Differential Form: (see Secs. 2.2 and 3.3)
Closed Manifold: (see Sec. 3.1)
Closed Set: (see Sec. 1.1)
Closure: (see Sec. 1.1)
Closure Point: (see Sec. 1.1)
Coexact Differential Form: (see Sec. 3.3)
Compact Space: (see Sec. 1.3)
Complete Metric Space: (see Sec. 1.3)
Complex Manifold: (see Sec. 3.1)
Complex Projective Spaces: (see Introduction)
Connected Space: (see Sec. 1.3)
Connected Sum: (see Introduction)
Continuity: (see Sec. 1.2)
Convergent Sequence: (see Sec. 1.3)
Coordinate Charts: (see Sec. 3.1)
Critical Numbers of a Map: (see Sec. 3.4)
Critical Point: (see Sec. 3.4)
Cyclic Normal Form: (see Sec. 2.3)
De Rham Cohomology: (see Sec. 3.3)
Dense: (see Sec. 1.1)
Diffeomorphism: (see Sec. 3.1)
Differential Form: (see Secs. 2.1 and 3.3)
Equicontinuity: (see Sec. 1.3)
Euler–Poincaré Characteristic: (see Sec. 2.3)
Exact Differential Form: (see Secs. 2.1 and 3.3)
Exact Sequence: (see Sec. 2.3)
Exterior Point: (see Sec. 1.1)
Finite Subcover: (see Sec. 1.3)
Finitely Generated Abelian Group: (see Sec. 2.3)
First Homotopy Group: (see Sec. 2.2)
Focal Points: (see Sec. 3.4)
Fundamental Group: (see Sec. 2.2)
***g*-Holed Torus:** (see Introduction)
Harmonic Differential Form: (see Sec. 3.3)
Hausdorff Space: (see Sec. 1.1)
Hessian: (see Sec. 3.4)
Hilbert Space: (see Sec. 1.3)
Homeomorphic: (see Sec. 1.2)
Homeomorphism: (see Sec. 1.2)
Homologous: (see Sec. 2.1)
Homology Group: (see Sec. 2.3)
Homotopic Paths: (see Secs. 2.1 and 2.2)

Hypersurface: (see Sec. 2.1)
Index of a Zero of a Vector Field: (see Sec. 3.2)
Inner Product: (see Sec. 1.1)
Inner Product Space: (see Sec. 1.1)
Interior: (see Sec. 1.1)
Interior Point: (see Sec. 1.1)
Isolated Point: (see Sec. 1.1)
Klein Bottle: (see Introduction)
Laplace–Beltrami Operator: (see Sec. 3.3)
Limit Point: (see Sec. 1.1)
Locally Finite Refinement: (see Sec. 1.3)
Long Exact Sequence: (see Sec. 2.3)
Loop: (see Sec. 2.2)
Manifold: (see Sec. 3.1)
Map: (see Sec. 1.2)
Metric Space: (see Sec. 1.1)
Metric Topology: (see Sec. 1.1)
Möbius Strip: (see Introduction)
Morse Function: (see Sec. 3.4)
Neighborhood: (see Sec. 1.1)
Nondegenerate Critical Point: (see Sec. 3.4)
Nonorientable Manifold: (see Sec. 3.1)
Norm: (see Sec. 1.1)
Normed Space: (see Sec. 1.1)
Open Ball: (see Sec. 1.1)
Open Cover: (see Sec. 1.3)
Open Neighborhood: (see Sec. 1.1)
Open Set: (see Sec. 1.1)
Orientable Manifold: (see Sec. 3.1)
Oriented Simplex: (see Sec. 2.3)
Pair of Topological Spaces: (see Sec. 2.3)
Paracompact: (see Sec. 1.3)
Partition of Unity: (see Sec. 1.3)
Path: (see Sec. 1.3)
Path Connected Space: (see Sec. 1.3)
Poincaré Duality: (see Sec. 3.3)
Poincaré Group: (see Sec. 2.2)
Poincaré Section: (see Sec. 2.1)
Real Projective Spaces: (see Introduction)
Relative Homology: (see Sec. 2.3)
Relative: p-Boundary (see Sec. 2.3)
Relative: p-Cycle (see Sec. 2.3)
Relative Topology: (see Sec. 1.1)
Separation Conditions: (see Sec. 1.1)
Sequential Compactness: (see Sec. 1.3)
Simplex: (see Sec. 2.3)
Simply Connected: (see Secs. 2 and 2.2)
Singular Homology: (see Sec. 2.3)
Singular Simplex: (see Sec. 2.3)
Smooth Map: (see Sec. 3.1)
Space: (see Sec. 1.1)
Spheres: (see Introduction)
Subspace Topology: (see Sec. 1.1)
Sup Norm: (see Sec. 1.1)
Support: (see Sec. 1.3)
Topological Space: (see Sec. 1.1)
Topologically Equivalent: (see Sec. 1.2)
Topology: (see Sec. 1.1)
Torsion Coefficients: (see Sec. 2.3)
Torus: (see Introduction)
Totally Bounded Metric Space: (see Sec. 1.3)
Trivial Loop: (see Sec. 2.2)
Vector Field: (see Sec. 3.2)
Zero of a Vector Field: (see Sec. 3.2)

Further Reading

Abraham, R., Marsden, J. (1978), *Foundations of Mechanics*, 2nd ed., New York: Addison-Wesley.

Abraham, R., Marsden, J., Ratiu, T. (1988), *Manifolds, Tensor Analysis, and Applications*, 2nd ed., New York: Springer-Verlag.

Armstrong, M. (1983), *Basic Topology*, New York: Springer-Verlag.

Arnol'd, V. (1989), *Mathematical Methods of Classical Mechanics*, 2nd ed., New York: Springer-Verlag.

Atiyah, M. (1990), *The Geometry and Physics of Knots*, Cambridge, U.K.: Cambridge Univ. Press.

Azcárraga, J. de, Izquierdo, J. (1996), *Lie Groups, Lie Algebras, Cohomology and Some Applications in Physics*, Cambridge, U.K.: Cambridge Univ. Press.

Birkhoff, G. (1927), *Dynamical Systems*, New York: American Mathematical Society.

Bott, R. (1994), *Collected Papers*, 4 volumes, Basel: Birkhäuser Verlag.

Bott, R., Tu, L. (1982), *Differential Forms in Algebraic Topology*, New York: Springer-Verlag.

Bredon, G. (1993), *Topology and Geometry*, New York: Springer-Verlag.
Bröcker, T., Jänich, K. (1982), *Introduction to Differential Topology*, Cambridge, U.K.: Cambridge Univ. Press.
Cairns, S. (1961), *Introductory Topology*, New York: Ronald Press Company.
Chern, S.-S. (1979), *Complex Manifolds without Potential Theory*, 2nd ed., New York: Springer-Verlag.
Choquet-Bruhat, Y., DeWitt-Morette, C., Dillard-Bleick, M. (1982, 1989), *Analysis, Manifolds, and Physics*, 2 volumes, Amsterdam: North-Holland.
Dodson, C., Poston, T. (1991), *Tensor Geometry*, New York: Springer-Verlag.
Dold, A. (1980), *Lectures on Algebraic Topology*, 2nd ed., New York: Springer-Verlag.
Dugundji, J. (1966), *Topology*, Boston: Allyn and Bacon.
Eguchi, T., Gilkey, P., Hanson, A. (1980), "Gravitation, Gauge Theories and Differential Geometry," *Phys. Rep.* **66**, 213–393.
Felsager, B. (1983), *Geometry, Particles and Fields*, 2nd ed., Odense, Denmark: Odense Univ. Press.
Flanders, H. (1963), *Differential Forms*, New York: Academic.
Fulton, W. (1995), *Algebraic Topology*, New York: Springer-Verlag.
Gilkey, P. (1995), *Invariance Theory, the Heat Equation, and the Atiyah–Singer Index Theorem*, 2nd ed., Boca Raton, FL: CRC Press.
Göckeler, M., Schücker, T. (1987), *Differential Geometry, Gauge Theories, and Gravity*, Cambridge, U.K.: Cambridge Univ. Press.
Green, M., Schwarz, J., Witten, E. (1987), *Superstring Theory*, 2 volumes, Cambridge, U.K.: Cambridge Univ. Press.
Greenberg, M., Harper, J. (1981), *Algebraic Topology, A First Course*, New York: Benjamin-Cummings.
Guillemin, V., Pollack, A. (1974), *Differential Topology*, Englewood Cliffs, NJ: Prentice-Hall.
Hilton, P., Wylie, S. (1960), *Homology Theory*, Cambridge, U.K.: Cambridge Univ. Press.
Hirsch, M. (1976), *Differential Topology*, New York: Springer-Verlag.
Hirzebruch, F. (1966), *Topological Methods in Algebraic Geometry*, 3rd ed., New York: Springer-Verlag.
Hocking, J., Young, G. (1961), *Topology*, New York: Addison-Wesley.
Hu, S. (1959), *Homotopy Theory*, New York: Academic.
Hübsch, T. (1992), *Calabi-Yau Manifolds, A Bestiary for Physicists*, Singapore: World Scientific.
Husemoller, D. (1994), *Fibre Bundles*, 3rd ed., New York: Springer-Verlag.
Jänich, K. (1984), *Topology*, New York: Springer-Verlag.
Kelley, J. (1955), *General Topology*, New York: Van Nostrand.
Kinsey, L. (1993), *Topology of Surfaces*, New York: Springer-Verlag.
Kobayashi, S., Nomizu, K. (1963, 1969), *Foundations of Differential Geometry*, 2 volumes, New York: Wiley-Interscience.
Kosinski, A. (1993), *Differential Manifolds*, New York: Academic.
Kreyszig, E. (1978), *Introductory Functional Analysis with Applications*, New York: Wiley.
Lang, S. (1993), *Real and Functional Analysis*, New York: Springer-Verlag.
Lang, S. (1995), *Differential and Riemannian Manifolds*, New York: Springer-Verlag.
Mac Lane, S. (1963), *Homology*, New York: Springer-Verlag.
Massey, W. (1991), *A Basic Course in Algebraic Topology*, New York: Springer-Verlag.
Milnor, J. (1963), *Morse Theory*, Princeton: Princeton Univ. Press.
Milnor, J. (1965), *Topology from the Differentiable Viewpoint*, Charlottesville, Va: The University Press of Virginia.
Milnor, J., Stasheff, J. (1974), *Characteristic Classes*, Princeton, NJ: Princeton Univ. Press.
Munkres, J. (1975), *Topology, A First Course*, Englewood Cliffs, NJ: Prentice-Hall.
Munkres, J. (1984), *Elements of Algebraic Topology*, New York: Addison-Wesley.
Nakahara, M. (1990), *Geometry, Topology and Physics*, Bristol: Institute of Physics Publishing.
Nash, C. (1991), *Differential Topology and Quantum Field Theory*, New York: Academic.
Nash, C., Sen, S. (1983), *Topology and Geometry for Physicists*, New York: Academic.
Naylor, A., Sell, G. (1982), *Linear Operator Theory in Engineering and Science*, New York: Springer-Verlag.
Reed, M., Simon, B. (1980), *Methods of Modern Mathematical Physics*, Vol. 1, rev. and enl. ed., New York: Academic.
Rotman, J. (1988), *An Introduction to Algebraic Topology*, New York: Springer-Verlag.

Schlichenmaier, M. (1989), *An Introduction to Riemann Surfaces, Algebraic Curves and Moduli Spaces*, New York: Springer-Verlag.
Schutz, B. (1980), *Geometrical Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge Univ. Press.
Seifert, H., Threlfall, W. (1980), *A Textbook of Topology*, New York: Academic.
Sieradski, A. (1992), *An Introduction to Topology and Homotopy*, Boston: PWS-Kent.
Simmons, G. (1963), *Introduction to Topology and Modern Analysis*, New York: McGraw-Hill.
Spanier, E. (1981), *Algebraic Topology*, New York: Springer-Verlag.
Springer, G. (1981), *Introduction to Riemann Surfaces*, 2nd ed., New York: Chelsea Publishing Company.
Steenrod, N. (1951), *The Topology of Fibre Bundles*, Princeton: Princeton Univ. Press.
Stillwell, J. (1993), *Classical Topology and Combinatorial Group Theory*, 2nd ed., New York: Springer-Verlag.
Veblen, O. (1931), *Analysis Situs*, 2nd ed., Providence, RI: American Mathematical Society.
Vick, J. (1994), *Homology Theory*, 2nd ed., New York: Springer-Verlag.
von Westenholz, C. (1981), *Differential Forms in Mathematical Physics*, Amsterdam: North-Holland.
Wallace, A. (1957), *An Introduction to Algebraic Topology*, New York: Pergamon.
Wallace, A. (1968), *Differential Topology, First Steps*, New York: Benjamin.
Warner, F. (1983), *Foundations of Differentiable Manifolds and Lie Groups*, New York: Springer-Verlag.
Wells, R. (1980), *Differential Analysis on Complex Manifolds*, New York: Springer-Verlag.
Weyl, H. (1955), *The Concept of a Riemann Surface*, 3rd ed., New York: Addison-Wesley.
Wilansky, A. (1970), *Topology for Analysis*, Waltham, MA: Ginn and Company.

Variational Methods

G. W. F. Drake
Department of Physics, University of Windsor, Windsor, Ontario, Canada

Introduction ***620***
1 Techniques of the Calculus of Variations ***621***
1.1 Variations with Constraints *623*
1.2 Generalizations *624*
2 Applications to Classical Mechanics ***625***
2.1 Introductory Concepts *625*
2.2 Hamilton's Principle *626*
2.2.1 Conservative Systems and First Integrals *626*
2.3 The Hamilton–Jacobi Equation *627*
2.3.1 The Principle of Least Action *629*
2.3.2 Hamilton's Equations of Motion *630*
2.3.3 Canonical Transformations *631*
2.3.4 Interpretation of the Hamilton–Jacobi Equation *632*
2.4 Relativistic Generalization *633*
2.4.1 Inclusion of Electromagnetic Fields *633*
3 Applications to Quantum Mechanics ***634***
3.1 Variational Derivation of the Schrödinger Equation *634*
3.2 The Rayleigh–Schrödinger Variational Principle *635*
3.3 The Rayleigh–Ritz Variational Method *635*
3.3.1 Algebraic Solution for Linear Variational Parameters *636*
3.3.2 Extension to Excited States *637*
3.3.3 Variational Lower Bound *637*
3.3.4 Illustrative Results for Helium *638*
3.3.5 Extensions to More Complex Systems *641*
3.4 Variation–Perturbation Methods *642*
3.4.1 Variational Bounds *643*
3.4.2 Spectral Representations and Pseudostates *644*

Mathematical Tools for Physicists. Edited by George L. Trigg
© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
ISBN: 3-527-40548-8

3.4.3 Example: The Polarizability of Hydrogen *644*
3.4.4 Time-Dependent Perturbations *645*
4 The General Sturm–Liouville Problem *646*
4.1 The Oscillation Theorem *646*
4.2 Example: The Coulomb Problem *647*
5 Applications to Electrodynamics *648*
6 Feynman's Path Integral *649*
6.1 Relation to Classical Dynamics *651*
Acknowledgments *651*
Glossary *652*
List of Works Cited *654*
Further Reading *655*

Introduction

Variational principles derive from a certain aesthetic and metaphysical ideal of simplicity in the search for the principles underlying physical phenomena. The origins date back to the earliest Greek philosophers Thales (c. 600 B.C.) and Pythagoras (c. 550 B.C.). Aristotle (384–322 B.C.) clearly makes use of a variational principle to justify circular orbits for the planets when he says in *de Caelo II*

> Now of lines which return upon themselves, the line which bounds the circle is the shortest, and that movement is the swiftest which follows the shortest line.

This marks the first use of a "minimum" postulate, and the conclusion held sway until the time of Kepler (1571–1630). Hero of Alexandria (c. 125 B.C.) made the first rigorous use of a variational principle when he proved that when the angle of incidence equals the angle of reflection, the path taken by a ray of light from the object to the observer is shorter than any other possible path with fixed end points (see *Catoptrics* by Hero in Cohen and Drabkin, 1958). This later became Fermat's principle of least time in geometrical optics.

The belief that nature is in some sense "simple" and can be explained by some economically small number of postulates pervades the works of Galileo (1564–1642), Newton (1642–1727), and Leibniz (1646–1716). Although some of the early conclusions turned out to be scientifically unfounded, the philosophical basis for variational principles has great antiquity. They continue to guide the development of new physical theories at the most fundamental level and to provide powerful methods of practical computation. Perhaps most importantly, they bring out the structural analogies between superficially different phenomena and allow techniques developed in one field to be readily applied in another. At a fundamental level, practically all physical phenomena can be expressed in terms of variational principles that have a striking similarity.

The *calculus of variations* provides the basic mathematical tool for formulating and analyzing variational principles. The purposes of this article are first to give an overview of the calculus of variations, and then to discuss its application to a

variety of physical phenomena. There is a vast literature on both aspects, and only a few of the most important points can be covered in the space available. Only a few principal references are given, with further general references in the reading list at the end. The main emphasis is on applications to classical mechanics and bound-state problems in quantum mechanics. Except for a brief discussion of the Feynman path integral, applications to scattering problems are not covered. An informal and instructive introduction to variational methods can be found in Hildebrand and Tromba (1985).

1 Techniques of the Calculus of Variations

In its simplest form, the calculus of variations addresses the problem of finding the function $y(x)$ for which the integral

$$J = \int_{x_1}^{x_2} f(x, y, y_x)\, dx \tag{1}$$

is an extremum. The integrand $f(x, y, y_x)$ is some prescribed function of the indicated variables, where $y_x = dy/dx$, and x_1, x_2 are fixed end points. J is termed a *functional* of $y(x)$. As originally formulated by Euler, the problem is solved by considering infinitesimal variations $\delta y(x)$ about a particular path $y(x)$ connecting x_1 and x_2 (see Fig. 1) and demanding that

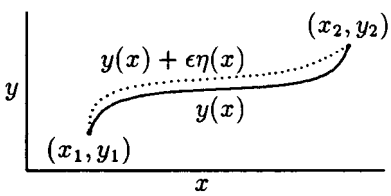


Fig. 1 Illustration of the actual path $y(x)$ and the varied path connecting fixed end points

the variation δJ induced in J vanish. For example, one might choose the variations to be $\delta y(x) = \varepsilon\eta(x)$, where $\eta(x)$ is an arbitrary function such that $\eta(x_1) = \eta(x_2) = 0$ (to make the variation vanish at the end points), and ε is a small parameter controlling the size of the variation. Then $\delta y_x(x) = \varepsilon d\eta(x)/dx$, and from a Taylor series expansion of $f(x, y, y_x)$ about $\varepsilon = 0$ in Eq. (1), the induced variation in J is

$$\delta J = \varepsilon \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y}\eta(x) + \frac{\partial f}{\partial y_x}\frac{d\eta(x)}{dx} \right] dx \tag{2}$$

up to terms of first order in ε. An integration of the second term by parts yields

$$\delta J = \varepsilon \frac{\partial f}{\partial y_x}\eta(x)\bigg|_{x_1}^{x_2} + \varepsilon \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x} \right] \times \eta(x)\, dx. \tag{3}$$

The first term vanishes by the assumption that $\eta(x) = 0$ at the end points. Since $\eta(x)$ is otherwise an arbitrary function, the condition $\delta J = 0$ can be fulfilled only if the integrand of the second term vanishes identically for $x_1 < x < x_2$; i.e.,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x} = 0. \tag{4}$$

This is the basic Euler–Lagrange equation. For purposes of compactness, the functional dependence of f on its various arguments is usually suppressed, as is done here.

Any solution to Eq. (4) satisfies the variational condition $\delta J = 0$. However, J itself could be a minimum, a maximum, or a point of inflection. One often knows from the geometrical nature of the problem being solved which case applies. Otherwise, it is necessary to extend the Taylor series expansion of $f(x, y, y_x)$ to terms of order ε^2 and determine the sign of

the second-order variation $\delta^{(2)}J$ (see, e.g., Courant and Hilbert, 1966). If $\delta^{(2)}J < 0$, then J is a maximum; if $\delta^{(2)}J > 0$, then J is a minimum; if $\delta^{(2)}J = 0$, then J lies at a point of inflection.

If the end points are not fixed, then $\eta(x)$ does not vanish there. In this case, Eq. (4) still applies, subject to the condition that $\partial f/\partial y_x = 0$ at the end points (see Jeffreys and Jeffreys, 1972).

An important special case of Eq. (4) occurs if $f(x, y, y_x)$ does not depend explicitly on x because then the integrating factor is simply y_x. After multiplying Eq. (4) through by y_x and using

$$\frac{df}{dx} = \frac{\partial f}{\partial y} y_x + \frac{\partial f}{\partial y_x}\frac{dy_x}{dx} \tag{5}$$

(since $\partial f/\partial x = 0$ by assumption), the Euler–Lagrange equation becomes

$$\frac{d}{dx}\left(y_x \frac{\partial f}{\partial y_x} - f\right) = 0, \tag{6}$$

and so

$$y_x \frac{\partial f}{\partial y_x} - f = \text{const.} \tag{7}$$

A classic example is provided by the brachistochrone (shortest time) problem first propounded by John Bernoulli in 1696. It was solved by both him and his brother James, as well as by Newton and Leibnitz. Consider a bead sliding without friction on a wire of arbitrary length connecting two fixed points (x_1, y_1) and (x_2, y_2) in a vertical plane, as shown in Fig. 2. The problem is to find the shape that minimizes the travel time as the bead slides from rest under the force of gravity; i.e., to find the function $y = y(x)$ such that the integral

$$\tau = \int_{x_1}^{x_2} \frac{ds}{v} \tag{8}$$

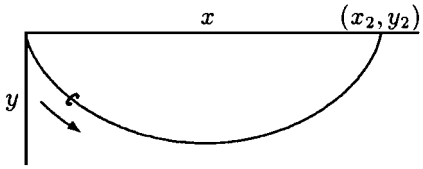


Fig. 2 The brachistochrone problem of a bead sliding without friction on a wire of arbitrary length in a vertical plane

for the travel time is a minimum. Here $ds = [(dy/dx)^2 + 1]^{1/2}dx$ is the element of arc length and v is the velocity. By conservation of energy, the velocity after falling a distance y (measured downward) is $v = \sqrt{2gy}$ (independent of x), where g is the acceleration due to gravity. The integral to be minimized is then

$$\tau = \int_{x_1}^{x_2} \frac{(y_x^2 + 1)^{\frac{1}{2}}}{(2gy)^{\frac{1}{2}}}\, dx. \tag{9}$$

With $f(x, y, y_x)$ defined by the above integrand, the Euler–Lagrange equation (7) then gives

$$\frac{y_x^2}{(y_x^2 + 1)^{\frac{1}{2}} y^{\frac{1}{2}}} - \frac{(y_x^2 + 1)^{\frac{1}{2}}}{y^{\frac{1}{2}}} = c^{\frac{1}{2}}, \tag{10}$$

and hence

$$y_x^{-1} \equiv \frac{dx}{dy} = \left(\frac{cy}{1 - cy}\right)^{\frac{1}{2}}, \tag{11}$$

where c is a constant of integration that determines the distance scale. Assuming that the bead starts from the origin, integration of this equation yields

$$cx = \sin^{-1}(cy)^{\frac{1}{2}} - (cy - c^2y^2)^{\frac{1}{2}}. \tag{12}$$

This is the equation of a cycloid symmetric about the minimum at $cy = 1$. The scale factor c is determined by the condition

that the curve pass through the second terminus (x_2, y_2). If $y_2 = 0$ (as in Fig. 2), then $c = \pi/x_2$.

1.1 Variations with Constraints

There are many classes of problems where the functional J must be made an extremum subject to a subsidiary condition of the form

$$\int_{x_1}^{x_2} f_1(x, y, y_x)dx = \text{const.} \tag{13}$$

The example of a hanging chain of fixed length is discussed in the following paragraph. Such problems can be handled by applying the variational procedure to the functional

$$g(x, y, y_x) = f(x, y, y_x) + \lambda_1 f_1(x, y, x_y), \tag{14}$$

where λ_1 is called a *Lagrange undetermined multiplier*. The resulting Euler – Lagrange equation is

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y_x} = -\lambda_1\left(\frac{\partial f_1}{\partial y} - \frac{d}{dx}\frac{\partial f_1}{\partial y_x}\right), \tag{15}$$

or, if f and f_1 are independent of x, the first integral is [cf. Eq. (7)]

$$y_x\frac{\partial f}{\partial y_x} - f + \lambda_1\left(y_x\frac{\partial f_1}{\partial y_x} - f_1\right) = \text{const.} \tag{16}$$

The idea is to solve Eq. (15) or Eq. (16) for a fixed but arbitrary value of λ_1. The equation of constraint provides the additional condition to determine λ_1 at the end of the problem, together with the two constants of integration.

As an example, consider the problem of finding the shape of a uniform hanging chain with both ends fixed. The shape is such that the potential energy due to gravity is a minimum, and so the quantity to be minimized is

$$J = \int_{x_1}^{x_2} \mu g y(y_x^2 + 1)^{\frac{1}{2}}\, dx, \tag{17}$$

subject to the constraint

$$\int_0^L ds \equiv \int_{x_1}^{x_2} (y_x^2 + 1)^{\frac{1}{2}} dx = L, \tag{18}$$

where μ is the mass per unit length and L the length. Thus $f = \mu g y(y_x^2 + 1)^{1/2}$ and $f_1 = (y_x^2 + 1)^{1/2}$. After dividing by μg and defining $\lambda = \lambda_1/\mu g$, Eq. (16) gives

$$(y + \lambda)\left[\frac{y_x^2}{(y_x^2 + 1)^{\frac{1}{2}}} - (y_x^2 + 1)^{\frac{1}{2}}\right] = -\frac{1}{c}. \tag{19}$$

As for the brachistochrone problem, this equation can be solved for $y_x^{-1} \equiv dx/dy$ and the result integrated to obtain

$$cy = -c\lambda + \cosh[c(x - a)], \tag{20}$$

where a is the second constant of integration. The three parameters a, c, and λ are determined by the three conditions that the curve pass through the points (x_1, y_1) and (x_2, y_2) at the ends of the chain, together with the equation of constraint. For example, if the two points are (0,0) and $(x_2, 0)$, then $a = x_2/2$, $c\lambda = \cosh(cx_2/2)$, and the equation of constraint becomes

$$\begin{aligned} L &= \int_0^{x_2} \{\sinh^2[c(x - \tfrac{1}{2}x_2)] + 1\}^{\frac{1}{2}} dx \\ &= \frac{2}{c}\sinh\left(\frac{cx_2}{2}\right). \end{aligned} \tag{21}$$

Solving this transcendental equation determines the remaining constant c. The quantity $T_0 = \mu g/c$ is the force of tension in the chain at the lowest point, where $y_x = 0$.

1.2 Generalizations

In applications to mechanics (see Sec. 2.1), the time t plays the role of the independent variable x marking the evolution of the system, but there are typically several dependent generalized coordinates q_i and $\dot{q}_i \equiv dq_i/dt$, $i = 1, \ldots, N$, in place of y and y_x, respectively. With this change in notation, f becomes a function of all the q_i's, $\dot{q}_i$'s, and t, and the generalization of Eq. (4) is

$$\frac{\partial f}{\partial q_i} - \frac{d}{dt}\frac{\partial f}{\partial \dot{q}_i} = 0, i = 1, \ldots, N. \quad (22)$$

There are thus N coupled Euler–Lagrange equations, one for each degree of freedom of the system.

If in addition there are several independent variables $t_1, t_2, \ldots, t_r$, then Eq. (22) is further generalized to read

$$\frac{\partial f}{\partial q_i} - \sum_{j=1}^{r} \frac{\partial}{\partial t_j}\frac{\partial f}{\partial(\partial q_i/\partial t_j)} = 0, \quad i = 1, \ldots, N. \quad (23)$$

There can also be several equations of constraint of the form of Eq. (13) with integrands $f_k(t_j, q_i, \dot{q}_i)$, $k = 1, \ldots, m$. In this case, m Lagrange undetermined multipliers λ_k are introduced, and the function f in Eq. (22) or (23) is replaced by

$$g(t_j, q_i, \dot{q}_i) = f(t_j, q_i, \dot{q}_i) + \sum_{k=0}^{m} \lambda_k f_k(t_j, q_i, \dot{q}_i). \quad (24)$$

Constraints that can be expressed in integrated form, such as Eq. (13), are said to be *holonomic* (wholly named or specified). However, problems often arise in mechanics involving *nonholonomic* constraints that can only be expressed in differential form, such as a relation between velocities. An example is the problem of a vertical disk of radius R rolling without slipping on a plane. Four coordinates are required – the (x,y) Cartesian coordinates of the point of contact between the disk and the plane, a spinning angle of rotation θ about a vertical axis, and a rolling angle ϕ about an axis perpendicular to the disk. If the plane of the disk is initially perpendicular to the x axis (i.e., $\theta = 0$), then the constraint of "not slipping" corresponds to the differential relations

$$dx = R\sin\theta\, d\phi, \quad (25)$$

$$dy = -R\cos\theta\, d\phi. \quad (26)$$

These equations cannot be integrated without knowing in advance θ and ϕ as a function of t. However, the method of Lagrange undetermined multipliers can still be applied. If the general form of the differential constraints is written as

$$\sum_{i=1}^{N} a_{k,i}\, dq_i + a_{k,t} dt = 0, \quad k = 1, \ldots, s, \quad (27)$$

where the coefficients $a_{k,i}$ are, in general, functions of the q_i's and $\dot{q}_i$'s, then the generalization of Eq. (22) for nonholonomic systems is

$$\frac{\partial f}{\partial q_i} - \frac{d}{dt}\frac{\partial f}{\partial \dot{q}_i} + \sum_{k=1}^{s} \lambda_k a_{k,i} = 0, \quad i = 1, \ldots, N. \quad (28)$$

This equation can still be used even if the constraints are in fact holonomic. The term $\Sigma_{k=1}^{s} \lambda_k a_{k,i}$ can be identified with the generalized forces of constraint. In general, the λ_k are now functions of the q_i's and $\dot{q}_i$'s. The term $a_{k,t}dt$ in Eq. (27) does not contribute to Eq. (28) because

the variations δq_i from the actual path are considered to occur at a particular instant of time (see Sec. 2.1).

As a simple illustration of the versatility of the method, consider the problem of finding the function $u(x,y,z)$ such that the square of its gradient is a minimum in a given volume of space. The problem is then to minimize

$$J = \int\int\int f\, dx\, dy\, dz, \tag{29}$$

with

$$f = (\nabla u)^2 = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2. \tag{30}$$

This can be regarded as an application of Eq. (23) with the three independent variables $t_1 = x, t_2 = y$, and $t_3 = z$, and a single degree of freedom ($N = 1$) with $q_1 = u$. Equation (23) then immediately gives

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{31}$$

which is Laplace's equation. The term $\partial f/\partial q_1 \equiv \partial f/\partial u$ does not contribute because f does not depend explicitly on u, only on the partial derivatives $\partial u/\partial x$, $\partial u/\partial y$, and $\partial u/\partial z$. This problem clearly illustrates the way in which functional derivatives are to be interpreted and should be carefully studied.

2 Applications to Classical Mechanics

This section draws together the connections of elements and techniques of classical mechanics with variational principles.

2.1 Introductory Concepts

Consider a classical system of n interacting particles having masses m_s located at positions $\mathbf{r}_s = (x_{s,1}, x_{s,2}, x_{s,3})$, and acted on by forces $\mathbf{F}_s = (F_{s,1}, F_{s,2}, F_{s,3})$ due to the other particles and any external forces, including forces of constraint. The evolution of the system is obtained by solving Newton's equations of motion

$$m_s \ddot{x}_{s,j} = F_{s,j},\ s = 1, \ldots, n;\ j = 1, 2, 3. \tag{32}$$

These equations are completely general. However, it is often more convenient to replace the $3n$ Cartesian coordinates $x_{s,j}$ by *generalized coordinates* q_i defined through a system of connection equations of the form

$$x_{s,j} = x_{s,j}(q_1, q_2, \ldots, q_{3n}). \tag{33}$$

The use of generalized coordinates is particularly effective in problems involving constraints. If the generalized coordinates are chosen such that their variations δq_i do no *virtual work* against the forces of constraint (i.e., δq_i is perpendicular to the *instantaneous* forces of constraint), then the number of independent q_i needed is reduced from $3n$ to $3n - m$, where m is the number of constraints.

To make these ideas concrete, consider the example of a bead sliding on a vertical wire hoop of radius R that is itself constrained to rotate about the z axis with angular velocity ω. In terms of the polar angles θ and $\phi = \omega t$, the connection equations are

$$\begin{aligned} x &= \sin\theta \cos\omega t, \\ y &= \sin\theta \sin\omega t, \\ z &= \cos\theta. \end{aligned} \tag{34}$$

The constraint of sliding on the hoop has reduced the number of independent coordinates from three to a single azimuthal angle θ. A variation $\delta\theta$ generates a displacement of the bead consistent with the instantaneous orientation of the hoop, but not consistent with the actual time evolution of the system, which includes the rotation of the hoop. Such variations are said to do no virtual work, and are called *virtual displacements*. From the connection equations (34), the kinetic energy of the bead is

$$\begin{aligned} T &= \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) \\ &= \tfrac{1}{2}m[(R\dot{\theta})^2 + (\omega R \sin\theta)^2]. \end{aligned} \tag{35}$$

If the system is conservative with a potential-energy function $V(x,y,z)$, then V can similarly be expressed in terms of θ.

2.2 Hamilton's Principle

In the absence of constraints, a direct transformation of Newton's equations of motion (32) from the $3n$ Cartesian coordinates $x_{s,j}$ to the $3n$ generalized coordinates q_i yields Lagrange's equations of motion

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} = Q_i, \tag{36}$$

where the Q_i are the generalized forces defined by

$$Q_i = \sum_{s=1}^{n} \mathbf{F}_s \cdot \frac{\partial \mathbf{r}_s}{\partial q_i}. \tag{37}$$

A comparison with Eqs. (3) and (22) shows immediately that with the identification $f = -T$, Lagrange's equations correspond to the variational condition

$$\int_{t_1}^{t_2} \left(\delta T + \sum_{i=1}^{3n} Q_i \delta q_i \right) dt = 0. \tag{38}$$

This is the most general form of Hamilton's principle in classical dynamics. The advantage gained is that $3n$ equations of motion have been consolidated into a single scalar variational condition that is invariant under coordinate transformation. In the absence of constraints, all the q_i can be varied independently so that each coefficient of δq_i must vanish separately, and Lagrange's equations are recovered. If m constraints are present, one need keep only $3n - m$ of the q_i whose virtual displacements δq_i are consistent with the instantaneous constraints, as in the rotating-hoop example of Sec. 2.1 where only a single parameter θ is required. One can still consider variations in the remaining q_i, even though they would violate the constraints. The only difference is that the corresponding Q_i are reinterpreted as the generalized forces required to maintain the constraints. They can be calculated by the method of Lagrange undetermined multipliers, as described in Sec. 1.2 [see especially Eq. (28)]. In this way, Lagrange's equations apply to the entire set, whether or not constraints are present. The invariance of Lagrange's equations and Hamilton's principle under coordinate transformation guarantees that if they are correct in Cartesian coordinates (as can easily be checked), they are correct in any other system of generalized coordinates.

2.2.1 Conservative Systems and First Integrals

If the system is conservative, then the Q_i are derivable from a potential function

$V(q_i)$ according to $Q_j = -\partial V/\partial q_i$, and Lagrange's equations reduce to

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0, \tag{39}$$

where $L = T - V$ is the Lagrangian. It is then immediately evident from Eq. (4) that Hamilton's principle becomes

$$\delta \int_{t_1}^{t_2} L dt = 0. \tag{40}$$

This is the most useful form of Hamilton's principle for theoretical discussion.

If T and V do not involve time explicitly, then the first integral corresponding to Eq. (6) is

$$\sum_i \dot{q}_i \frac{\partial T}{\partial \dot{q}_i} - T + V = \text{const.} \tag{41}$$

If, further, T is a homogeneous quadratic function of the $\dot{q}_i$, then the above becomes

$$T + V = \text{const.}, \tag{42}$$

where the constant can now be identified with the conserved energy E of the system.

For the example of a bead on a rotating hoop discussed in Sec. 2.1, the first integral gives

$$\begin{aligned} &\tfrac{1}{2}m[(R\dot{\theta})^2 - (\omega R \sin\theta)^2] + V(\theta) \\ &\quad = T + V - m(\omega R \sin\theta)^2 = \text{const.} \end{aligned} \tag{43}$$

The constant of the motion here differs from $E = T + V$ because the time-dependent forces of constraint do work on the system. Overall conservation of energy is recovered only when the work required to keep the hoop rotating at a constant rate is included. If N_z is the required torque, then, with the use of Eq. (43), the rate at which it does work is

$$\begin{aligned} N_z\omega &= \frac{d}{dt}(T + V) = \frac{d}{dt}m(\omega R \sin\theta)^2 \\ &= \omega \frac{dl_z}{dt}, \end{aligned} \tag{44}$$

where $l_z = m\omega(R\sin\theta)^2$ is the angular momentum of the bead about the z axis. Thus $N_z = dl_z/dt$ as expected.

2.3 The Hamilton–Jacobi Equation

The variations considered thus far are taken between fixed end points t_1 and t_2 and, therefore, necessarily do not correspond to a possible dynamical evolution of the system. The actual evolution between fixed end points is uniquely defined, at least with respect to local variations, and so the varied path is unphysical. (But it is not necessarily so for nonlocal variations. For example, two points on a Kepler orbit are connected by two paths, depending on which way around the particle goes.)

The Hamilton–Jacobi equation comes from consideration of a different kind of variation Δq_i along a possible dynamical path between points allowed to vary in both space and time. In this case, it is necessary to keep the integrated term in Eq. (3) (or its generalizations). Suppose that t_1 and t_2 are replaced by $t_1 + \delta t_1$ and $t_2 + \delta t_2$. There is then a corresponding variation δt in the arrival time at each point along the path, so that $\Delta q_i(t) \equiv \delta q_i(t + \delta t)$ is the variation in path evaluated at the modified arrival time.

For definiteness, suppose that the paths are parameterized according to

$$q_i(\varepsilon, t) = q_i(0, t) + \varepsilon \eta_i(t), \tag{45}$$

where $q_i(0, t)$ is the actual path and $\eta_i(t)$ is an arbitrary differentiable function not

assumed to vanish at the end points. Then $\delta q_i(t) = \varepsilon\eta_i(t)$ and

$$\begin{aligned}\delta q_i(t+\delta t) &= q_i(\varepsilon, t+\delta t) - q_i(0,t)\\ &\simeq q_i(0, t+\delta t) - q_i(0,t) + \varepsilon\eta_i(t)\\ &= \dot{q}_i\delta t + \delta q_i(t) \qquad (46)\end{aligned}$$

up to terms of first order in ε and δt. Application of the Δ variation to the integral in Hamilton's principle then yields

$$\begin{aligned}\Delta\int_{t_1}^{t_2} L\,dt &= \int_{t_1+\delta t_1}^{t_2+\delta t_2} L\,dt - \int_{t_1}^{t_2} L\,dt + \int_{t_1}^{t_2}\delta L\,dt\\ &= L\delta t\Big|_{t_1+\delta t_1}^{t_2+\delta t_2} + \sum_i \frac{\partial L}{\partial\dot{q}_i}\delta q_i\Bigg|_{t_1}^{t_2}\\ &\quad + \int_{t_1}^{t_2}\sum_i\left(\frac{\partial L}{\partial q_i} - \frac{d}{dt}\frac{\partial L}{\partial\dot{q}_i}\right)\delta q_i\,dt. \qquad (47)\end{aligned}$$

The last term vanishes by the assumption that the varied path is a possible dynamical path, and so Lagrange's equations of motion are satisfied. Interest therefore centers on the integrated terms. The first term is simply $L(\delta t_2 - \delta t_1)$. Using Eq. (46) to replace $\delta q_i(t)$ by $\delta q_i(t+\delta t)$ in the second term, these terms become

$$\Delta\int_{t_1}^{t_2} L\,dt = \left[\left(L - \sum_i \dot{q}_i\frac{\partial L}{\partial\dot{q}_i}\right)\delta t + \sum_i\frac{\partial L}{\partial\dot{q}_i}\delta q_i\right]_{t_1+\delta t_1}^{t_2+\delta t_2}. \qquad (48)$$

With the definitions

$$p_i = \frac{\partial L}{\partial\dot{q}_i}, \qquad (49)$$

$$H = \sum_i \dot{q}_i\frac{\partial L}{\partial\dot{q}_i} - L, \qquad (50)$$

$$S = \int_{t_1}^{t_2} L\,dt, \qquad (51)$$

where the integral for S is taken along any dynamical path, Eq. (48) can be written in the form

$$\Delta S = \left[-H\delta t + \sum_i p_i\delta q_i\right]_{t_1+\delta t_1}^{t_2+\delta t_2}. \qquad (52)$$

H is called the *Hamiltonian* for the system, p_i is the *canonical momentum*, and S is called *Hamilton's principal function*. For an ordinary conservative system, $H = T + V$ is the total energy.

Equation (52) may be interpreted as follows. Since by assumption the system evolves along a possible dynamical path from an initial configuration $(q_i)_1$ at time t_1 to a final configuration $(q_i)_2$ at time t_2, there will in general be only a single set of initial velocities $(\dot{q}_i)_1$ that satisfy these requirements (at least with respect to small variations). The same applies to the initial momenta $(p_i)_1$ since they are connected to the $(\dot{q}_i)_1$ through Eq. (49), and hence the $(\dot{q}_i)_1$ can be eliminated from the problem. The p_i's as well as the q_i's are therefore uniquely determined from the initial conditions at time t_1. Since Eq. (52) remains true as t_2 is varied, it follows that

$$\frac{\partial S}{\partial q_i} = p_i, \qquad (53)$$

$$\begin{aligned}\frac{\partial S}{\partial t} &= -H(q_i, p_i, t)\\ &= -H\left(q_i, \frac{\partial S}{\partial q_i}, t\right). \qquad (54)\end{aligned}$$

Equation (54) is called the *Hamilton–Jacobi* equation. It is a first-order partial differential equation in $N+1$ variables and does not involve S explicitly. As it stands, a complete solution involving

$N+1$ constants of integration is determined from the initial conditions

$$\left(\frac{\partial S}{\partial t}\right)_{t_1} = (H)_{t_1}, \quad \left(\frac{\partial S}{\partial q_i}\right)_{t_1} = -(p_i)_{t_1}. \tag{55}$$

The real significance of the Hamilton–Jacobi equation comes not from its practical utility in solving mechanical problems, but for the insight it gives into the structure of mechanics, and for applications involving the use of perturbation theory. The significance is further elaborated after a brief discussion of the principle of least action, Hamilton's equations of motion, and the theory of canonical transformations.

2.3.1 The Principle of Least Action

Assume that the system is conservative and holonomic so that H is a constant. The Maupertuis principle of least action follows by considering a restricted class of variations Δ_t that are identical to the Δ variations of Sec. 2.3, except that $\Delta_t q_i = \delta q_i(t+\delta t)$ is assumed to vanish at the end points; i.e., the system arrives at the same end point, but at the varied time $t+\delta t$, and H has the same constant value on the varied path. The varied path could be the same as the actual path except that it is traversed at a different rate. Under these conditions, Eq. (52) reduces immediately to

$$\Delta_t S = -H(\delta t_2 - \delta t)_1. \tag{56}$$

However, with the use of Eq. (50) to replace L by H, a direct evaluation of $\Delta_t S$ from Eq. (51) yields

$$\Delta_t S = \Delta_t \int_{t_2}^{t_2} \sum_i p_i \dot{q}_i \, dt - H(\delta t_2 - \delta t_1), \tag{57}$$

from which it follows that

$$\Delta_t \int_{t_1}^{t_2} \sum_i p_i \dot{q}_i \, dt = 0. \tag{58}$$

The quantity

$$W = \int_{t_1}^{t_2} \sum_i p_i \dot{q}_i \, dt \tag{59}$$

defines the classical action.

This is one of many possible ways of expressing the principle of least action. A purely geometrical form in which t is eliminated can be obtained as follows. If V is velocity-independent and T is a homogeneous quadratic function of the $\dot{q}_i$, then

$$\sum_i p_i \dot{q}_i = 2T = \sum_{j,k} M_{j,k}(q) \dot{q}_j \dot{q}_k, \tag{60}$$

where the $M_{j,k}(q)$ are the coefficients appearing in the kinetic-energy expression. The $M_{j,k}(q)$ can be regarded as the elements of a metric tensor in a curvilinear coordinate space such that the element of path length is

$$(d\rho)^2 = \sum_{j,k} M_{j,k}(q) dq_j dq_k. \tag{61}$$

The principle of least action can then be written in the form

$$\Delta_t \int_{t_1}^{t_2} T \, dt = 0 = \Delta \int_{\rho_1}^{\rho_2} \sqrt{T} d\rho, \tag{62}$$

or equivalently,

$$\Delta \int_{\rho_1}^{\rho_2} \sqrt{H - V} d\rho = 0. \tag{63}$$

For a single particle moving in a potential V, $d\rho$ is simply the element of arc length ds along the trajectory. In this form, the principle of least action is formally identical to Fermat's principle

of geometrical optics. One need merely identify $\sqrt{H-V}$ with a variable index of refraction $n(s) = c_{\text{vac}}/c(s)$, which is inversely proportional to the velocity of light in the medium. The path taken by a beam of light is then such that the travel time given by

$$\tau = \frac{1}{c_{\text{vac}}} \int_{s_1}^{s_2} n(s)ds \tag{64}$$

is a minimum (or, more strictly, an extremum). This justifies the solution to the brachistochrone problem in Sec. 1, and it demonstrates the formal equivalence between geometrical optics and the dynamics of conservative systems.

In a recent article, Gray et al. (1996) show that the Maupertuis principle of least action can be generalized to a form in which W is stationary with respect to varied paths on which H is held constant only *in the mean* and that there exists a reciprocal principle in which the mean value $\bar{H}$ is made stationary with respect to varied paths of constant W. They also show that the reciprocal principle can be derived directly from the classical limit of the Schrödinger variational principle of quantum mechanics (see Sec. 3.2). A similar reciprocity theorem for Hamilton's principle provides a set of four variational principles analogous to the four equilibrium principles of thermodynamics.

2.3.2 Hamilton's Equations of Motion

The basic approach in Lagrangian mechanics is to regard the N generalized coordinates $q_i(t)$ as the independent variables whose time dependence is determined by the N second-order Lagrangian equations of motion expressed by Eq. (39). The velocities $\dot{q}_i(t)$ enter only as derived quantities whose initial values, together with the initial $q_i(t)$, determine the required $2N$ constants of integration. The Hamiltonian approach differs in that the $\dot{q}_i(t)$ are eliminated in favor of the canonical momenta $p_i(t)$ defined by Eq. (49). The $p_i(t)$ are then elevated to an equal footing with the $q_i(t)$ so that the set $\{q_i, p_i | i = 1, \ldots, N\}$ forms a set of $2N$ independent variables satisfying a set of $2N$ coupled first-order differential equations called Hamilton's equations of motion. In what follows, we adopt the convention that a summation over repeated subscripts is implied, and q or p without a subscript stands for the entire set.

Hamilton's equations of motion can be derived from Hamilton's principle if Eqs. (49) and (50) are first used to rewrite Eq. (51) in the form

$$S = \int [p_i \dot{q}_i - H(q, p, t)]\, dt. \tag{65}$$

With *fixed* end points t_1 and t_2, independent variations δp_i and δq_i then induce the variation

$$\begin{aligned} \delta S &= \int_{t_1}^{t_2} \left(\dot{p}_i \delta q_i + \dot{q}_i \delta p_i \right. \\ &\quad \left. + \frac{\partial H}{\partial q_i} \delta q_i + \frac{\partial H}{\partial p_i} \delta p_i \right) dt \\ &= p_i \delta q_i \Big|_{t_1}^{t_2} + \int_{t_1}^{t_2} \left(-\dot{p}_i \delta q_i + \dot{q}_i \delta p_i \right. \\ &\quad \left. + \frac{\partial H}{\partial q_i} \delta q_i + \frac{\partial H}{\partial p_i} \delta p_i \right) dt. \end{aligned} \tag{66}$$

The integrated term vanishes by assumption. Equating to zero the coefficients of δp_i and δq_i then yields the set of equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \tag{67}$$

which are Hamilton's equations of motion. Thus, both Lagrangian and Hamiltonian mechanics are implied by Hamilton's principle, provided that the variations δp_i and δq_i can be regarded as independent. The

latter assumption is important because it underlies the entire approach of Hamiltonian mechanics. It is justified by the fact that the Hamiltonian form can be obtained by a direct transformation of variables from the Lagrangian form.

2.3.3 Canonical Transformations

Transformations of the type (33) are called *point transformations* because they involve only the coordinates q_i. An advantage of Hamiltonian mechanics is that, with the q's and p's regarded as independent variables, more general types of transformation can be constructed that map the set $\{q, p\}$ to a new set $\{Q, P\}$ defined by connection equations of the form

$$Q_j = Q_j(q, p, t), \quad P_i = P_i(q, p, t). \tag{68}$$

Such a *contact transformation* is said to be *canonical* if the form of Hamilton's equations is left invariant; i.e., there exists a transformed Hamiltonian function $K(Q,P,t)$ such that

$$\dot{Q}_j = \frac{\partial K}{\partial P_i}, \quad \dot{P}_i = -\frac{\partial K}{\partial Q_j} \tag{69}$$

are the correct equations of motion for the transformed variables. This will be true if both sets of coordinates and momenta satisfy their respective variational principles

$$\delta \int [p_i \dot{q}_i - H(q, p, t)]\, dt = 0, \tag{70}$$

$$\delta \int [P_i \dot{Q}_j - K(Q, P, t)]\, dt = 0. \tag{71}$$

The integrands need not be equal, but they can differ by at most the total time derivative of a function F, called the *generating function* for the transformation. We can thus write

$$p_i \dot{q}_i - H(q, p, t) = \lambda \left[P_i \dot{Q}_j - K(Q, P, t) + \frac{dF}{dt} \right]. \tag{72}$$

Values of the constant $\lambda \neq 1$ just correspond to a scale transformation, and so without loss of generality, we can take $\lambda = 1$. With the choice $F = F_1(q, Q, t)$, the set $\{q, Q\}$ is regarded as the $2N$ separately independent variables, and the quantity dF/dt in Eq. (72) can be replaced by

$$\frac{dF_1}{dt} = \frac{\partial F_1}{\partial q_i} \dot{q}_i + \frac{\partial F_1}{\partial Q_j} \dot{Q}_j + \frac{\partial F_1}{\partial t}. \tag{73}$$

Equating to zero the coefficients of $\dot{q}_i$ and $\dot{Q}_j$ then yields

$$p_i = \frac{\partial F_1}{\partial q_i}, \quad P_i = -\frac{\partial F_1}{\partial Q_j}, \tag{74}$$

$$K = H + \frac{\partial F_1}{\partial t}. \tag{75}$$

For example, the choice $F_1 = q_i Q_j$ interchanges the roles of the q's and p's (except for a sign change), with the result

$$p_i = Q_i, \quad P_i = -q_i, \quad K = H. \tag{76}$$

This demonstrates explicitly that the q's and p's stand on an equal footing as independent variables in Hamiltonian mechanics – their roles can be interchanged by a canonical transformation.

Other choices for the $2N$ independent variables can be achieved by application of one or more Legendre transformations (as is done in thermodynamics to change independent variables). Of particular relevance to the Hamilton–Jacobi equation is the choice $\{q, P\}$. One need merely

replace $F_1(q, Q)$ by $F_2(q, P) = F_1 + Q_i P_i$. Then Eq. (72) becomes

$$p_i \dot{q}_i - H(q, p, t) = -Q_i \dot{P}_i - K(Q, P, t) + \frac{dF_2}{dt}, \tag{77}$$

with

$$\frac{dF_2}{dt} = \frac{\partial F_2}{\partial q_i}\dot{q}_i + \frac{\partial F_2}{\partial P_i}\dot{P}_i + \frac{\partial F_2}{\partial t}, \tag{78}$$

and so

$$P_i = \frac{\partial F_2}{\partial q_i}, \quad Q_i = \frac{\partial F_2}{\partial P_i}, \tag{79}$$

$$K = H + \frac{\partial F_2}{\partial t}. \tag{80}$$

For example, the choice $F_2 = q_i P_i$ generates the identity transformation with $Q_i = q_i$ and $P_i = p_i$.

2.3.4 Interpretation of the Hamilton–Jacobi Equation

A comparison of Eq. (80) with Eq. (54) shows that the Hamilton–Jacobi equation can be regarded as a canonical transformation with $F_2 = S(q, P, t)$ such that the transformed Hamiltonian is $K = 0$. The transformed equations of motion are then

$$\frac{\partial K}{\partial P_i} = \dot{Q}_i = 0, \quad \frac{\partial K}{\partial Q_i} = -\dot{P}_i = 0. \tag{81}$$

Thus the transformed momenta $P_i = \alpha_i$ and the transformed coordinates $Q_i = \beta_i$ are all constants of the motion. The nature of the solution is now clear. Writing $S = S(q, \alpha, t)$, then the set of equations

$$p_i = \frac{\partial S(q, \alpha, t)}{\partial q_i}, \quad \beta_i = \frac{\partial S(q, \alpha, t)}{\partial \alpha_i} \tag{82}$$

evaluated at $t = t_1$ provides a set of $2N$ equations to determine $2N$ unknowns α_i and β_i in terms of the initial $(q_i)_1$ and $(p_i)_1$. For conservative systems, the remaining $(N+1)$th constant of integration is the energy E of the system. For this case, the time variable is separable, and S can be written in the form

$$S(q, \alpha, t) = W(q, \alpha) - Et, \tag{83}$$

where $W(q, \alpha)$ is called *Hamilton's characteristic function*. The crucial point is that *any* complete integral to the Hamilton–Jacobi equation generates a possible dynamical motion of the system since it is the generating function for a canonical transformation. This result is called Jacobi's theorem.

As a consequence of Jacobi's theorem, any complete integral S contains within it all possible trajectories of the system as a function of the initial conditions, rather than one particular trajectory. In fact, surfaces of constant S move through configuration space like wave fronts of constant phase such that the particle trajectories follow the orthogonal set of curves. To see this, consider the example of a particle moving in a potential V, expressed in Cartesian coordinates. Equations (53) for $\{q_1, q_2, q_3\} = \{x, y, z\}$ can then be written as the single equation

$$\nabla S = \nabla W = \mathbf{p}, \tag{84}$$

which shows that the momentum $\mathbf{p} = m\mathbf{v}$ is everywhere perpendicular to surfaces of constant W, and the particle velocity is $v = |\nabla W|/m$. The Hamilton–Jacobi equation for W in this case is

$$\left(\frac{1}{2m}\right)(\nabla W)^2 + V = E, \tag{85}$$

so that

$$|\nabla W| = \sqrt{2m(E - V)}. \tag{86}$$

For the case of a freely falling particle, surfaces of constant W are just horizontal

planes with the particle trajectories in the perpendicular direction. As time goes on, surfaces of constant S sweep through surfaces of constant W with a phase velocity given by

$$u = \frac{ds}{dt} = \frac{E}{|\nabla W|}, \tag{87}$$

where ds is a displacement in the direction normal to a surface of constant W. The above follows from the facts that the stationary phase condition $dS = 0$ corresponds to $dW = E\,dt$, and $dW = |\nabla W| ds$ is the change in W due to a displacement in the direction normal to the surface. Thus the phase velocity decreases as the particle velocity increases, just as is the case for the wave and particle pictures of light.

The above considerations in fact provide a wave picture of classical dynamics in the "geometrical optics" limit where the wavelength is infinitesimally small compared with the dimensions of the apparatus. The wave nature can then be ignored, and the trajectories determined by the principle of least action (or Fermat's principle in the case of geometrical optics).

2.4 Relativistic Generalization

For the case of a single particle acted on by forces derivable from a potential V, Hamilton's principle can be simply modified to incorporate the effects of special relativity. One need simply define

$$L = -mc^2\gamma - V, \tag{88}$$

where $\gamma = \sqrt{1 - v^2/c^2}$, and v is the velocity $|\dot{\mathbf{r}}|$ in a particularly chosen Lorentz frame. With this choice of L, Hamilton's principle and the Euler – Lagrange equations give the correct equations of motion

$$\frac{d}{dt}\left(\frac{m\dot{x}_i}{\gamma}\right) = -\frac{\partial V}{\partial x_i}. \tag{89}$$

The canonical momenta p_i and Hamiltonian H are given by

$$p_i = \frac{\partial L}{\partial \dot{x}_i} = \frac{m\dot{x}_i}{\gamma}, \tag{90}$$

$$H = p_i\dot{x}_i - L = T + V = E, \tag{91}$$

where $T = mc^2/\gamma$ is a relativistic generalization of the kinetic energy, including the rest-mass energy mc^2. After substituting for $\dot{x}_i$, H assumes the form

$$H = \sqrt{c^2p^2 + m^2c^4} + V. \tag{92}$$

2.4.1 Inclusion of Electromagnetic Fields

In general, an electromagnetic field is derivable from a scalar potential $\phi(\mathbf{r}, t)$ and a vector potential $\mathbf{A}(\mathbf{r},t)$ according to

$$\mathbf{E} = -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}, \tag{93}$$

$$\mathbf{B} = \nabla \times \mathbf{A}, \tag{94}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields. The equation of motion for a particle of charge q is then

$$\frac{d}{dt}\left(\frac{m\dot{x}_i}{\gamma}\right) = qE_i + \frac{q}{c}(\mathbf{v} \times \mathbf{B})_i, \tag{95}$$

which now contains a velocity-dependent force term. This equation follows from the Lagrangian

$$L = -mc^2\gamma - q\phi + \frac{q}{c}\mathbf{A} \cdot \mathbf{v}, \tag{96}$$

or its nonrelativistic counterpart with the term $-mc^2\gamma$ replaced by T. The canonical

momenta are then

$$p_i = \frac{\partial L}{\partial \dot{x}_i} = \frac{m\dot{x}_i}{\gamma} + \frac{q}{c}A_i. \tag{97}$$

A direct calculation shows that the Hamiltonian becomes [cf. Eq. (92)]

$$H = \sqrt{c^2[\mathbf{p} - \left(\frac{q}{c}\right)\mathbf{A}]^2 + m^2c^4} + V. \tag{98}$$

The same substitution $\mathbf{p} \to \mathbf{p} - (q/c)\mathbf{A}$ applies also in the nonrelativistic case. This simple prescription, together with $V = q\phi$, allows electromagnetic fields to be easily incorporated into the Lagrangian and Hamiltonian formulations of mechanics.

3 Applications to Quantum Mechanics

3.1 Variational Derivation of the Schrödinger Equation

The considerations of Sec. 2.3.4 suggest that the Hamilton–Jacobi equation of classical dynamics expresses the short-wavelength limit of an underlying wave equation, with surfaces of constant S identified as surfaces of constant phase. In fact, Eq. (85) already bears a superficial resemblance to the time-independent Schrödinger equation, but it does not yet have the form of a wave equation. Following Schrödinger, a suitable wave equation can be obtained by first making the substitution

$$W = iC \ln \Psi \Rightarrow \Psi = e^{\frac{iW}{C}} \tag{99}$$

into the Hamilton–Jacobi equation

$$\left(\frac{1}{2m}\right)(\nabla W^* \cdot \nabla W) + V = E, \tag{100}$$

generalized for complex W, to obtain

$$\left(\frac{C^2}{2m}\right)(\nabla\Psi^* \cdot \nabla\Psi) + (V - E)\Psi^*\Psi = 0. \tag{101}$$

The left-hand side can be integrated over all space, provided that $\int \Psi^*\Psi d^3r$ remains finite. Application of the Euler–Lagrange equations to make the integral stationary with respect to arbitrary independent variations of Ψ and Ψ^* then yields the Schrödinger equation

$$-\left(\frac{C^2}{2m}\right)(\nabla^2\Psi) + (V - E)\Psi = 0, \tag{102}$$

together with a similar equation for Ψ^*. The derivation is a simple extension of the one used to obtain Laplace's equation (31) in Sec. 1.2. The solution to Eq. (102) then determines the wave function $\Psi(\mathbf{r})$ for the system, subject to the constraint that $\int \Psi^*\Psi d^3r$ remain finite for bound systems; i.e., that Ψ is *normalizable*. Comparison with experiment shows that one should set $C = \hbar = h/2\pi$, where h is Planck's constant. Equation (102) can then be written in the form

$$H(q, p)\Psi = E\Psi, \tag{103}$$

where $H(q, p)$ is the Hamiltonian with the quantum-mechanical replacement $\mathbf{p} \to (\hbar/i)\nabla$. The constraint on $\int \Psi^*\Psi d^3r$ makes this an eigenvalue problem that determines the possible energies E of the system.

If there are n interacting particles, then the term $\nabla^2\Psi$ is to be replaced by $\Sigma_i \nabla_i^2 \Psi$, and V includes all the interaction potentials. Also, the various integrals over $d\mathbf{r}$ are replaced by multiple integrals over $d^3r_1 d^3r_2 \cdots d^3r_n$.

3.2 The Rayleigh–Schrödinger Variational Principle

Consider a bound system, or one that is contained in a finite box. Under these conditions, Eq. (101) can be integrated over all space and the term $\nabla\Psi^* \cdot \nabla\Psi$ integrated by parts to obtain

$$\int \Psi^*(H - E)\Psi d^3r = 0. \qquad (104)$$

The integrated term does not contribute under the assumed conditions because $\Psi(\mathbf{r}) \to 0$ sufficiently rapidly as $|\mathbf{r}| \to \infty$. The variational derivation of Sec. 3.1 guarantees that this integral is stationary with respect to arbitrary variations $\delta\Psi$ if Ψ satisfies the Schrödinger equation. However, the same variational condition can now be reinterpreted as the problem of making the integral $\int \Psi^* H\Psi d^3r$ stationary, subject to the constraint that

$$\int \Psi^*\Psi d^3r = \text{const.}, \qquad (105)$$

with E playing the role of a Lagrange undetermined multiplier. In this guise, one can say that E obtained from the Rayleigh quotient

$$E = \frac{\int \Psi^* H\Psi d^3r}{\int \Psi^*\Psi d^3r} \qquad (106)$$

is stationary. In fact, as discussed in the following section, E is a minimum under many circumstances.

3.3 The Rayleigh–Ritz Variational Method

The Schrödinger equation is a partial-differential equation that can be solved exactly only for certain special cases such as the Coulomb potential or the harmonic-oscillator potential. For arbitrary potentials, or for problems containing more than two bodies, the quantum-mechanical problem is no easier to solve than the corresponding classical one. In these cases, the Rayleigh–Schrödinger variational principle provides one of the most powerful methods for obtaining approximate eigenvalues E and wave functions Ψ.

Suppose one guesses by some means an approximate trial wave function Ψ_{tr} that conforms with the constraint of normalizability and approximates one of the exact solutions to

$$H\Psi_i = E_i\Psi_i, \quad i = 1, 2, \ldots. \qquad (107)$$

The index i labels the spectrum of exact solutions. In general, the eigenvalue spectrum will have both discrete and continuous parts. In the latter case, summations over i include integrations over the continuous part. The crucial point is that even though the Ψ_i are not known, they form a complete basis set of functions in terms of which the trial function Ψ_{tr} can be expanded. In analogy with Fourier series, one can therefore write

$$\Psi_{\text{tr}} = \sum_{i=1}^{\infty} c_i\Psi_i, \qquad (108)$$

where the c_i are the expansion coefficients. Let the eigenvalue spectrum be ordered so that $E_1 < E_2 < E_3 < \cdots$, and assume that all the Ψ_i and Ψ_{tr} are normalized to unity; i.e., using Dirac bra-ket notation for integrals,

$$\begin{aligned} &\langle\Psi_{\text{tr}}|H|\Psi_{\text{tr}}\rangle \equiv \int \Psi^* H\Psi d^3r = E_{\text{tr}}, \\ &\langle\Psi_i|\Psi_j\rangle = \delta_{i,j}, \\ &\langle\Psi_i|H|\Psi_j\rangle = E_i\delta_{i,j}. \end{aligned} \qquad (109)$$

Substituting Eq. (108) into (106), and using Eqs. (109), one then obtains

$$E_{\mathrm{tr}} = |c_1|^2 E_1 + |c_2|^2 E_2 + |c_3|^2 E_3 + \cdots \tag{110}$$

for the corresponding trial energy. Since, by assumption, $\langle \Psi_{\mathrm{tr}} | \Psi_{\mathrm{tr}} \rangle = 1$, it follows that

$$\sum_{i=1}^{\infty} |c_i|^2 = 1, \tag{111}$$

and so Eq. (110) can be rewritten in the form

$$\begin{aligned} E_{\mathrm{tr}} &= E_1 + |c_2|^2 (E_2 - E_1) \\ &\quad + |c_3|^2 (E_3 - E_1) + \cdots \\ &\geq E_1. \end{aligned} \tag{112}$$

Thus E_{tr} is an *upper bound* on the lowest eigenvalue E_1 for any normalizable Ψ_{tr}.

The basic idea of variational calculations then is to write Ψ_{tr} in some arbitrarily chosen mathematical form with variational parameters (subject to normalizability and boundary conditions at the origin and infinity), and then adjust the parameters to obtain the minimum value of E_{tr}. A lower E_{tr} is guaranteed to be closer to E_1. The power of the method stems both from this and the fact that, by the Rayleigh–Schrödinger variational principle, the error term linear in $\delta\Psi = \Psi_{\mathrm{tr}} - \Psi_1$ vanishes.

3.3.1 Algebraic Solution for Linear Variational Parameters

Suppose that Ψ_{tr} depends in some arbitrarily chosen way on a set of N variational parameters $a_1, a_2, \ldots, a_N$. [For example, in a one-dimensional case, one might choose $\Psi(r) = r^{a_1} e^{-a_2 r}$ with a_1 and a_2 regarded as nonlinear variational parameters.] Then the variational condition corresponds to the system of equations

$$\frac{\partial E_{\mathrm{tr}}}{\partial a_p} = 0, \quad p = 1, \ldots, N. \tag{113}$$

In general, this is a set of transcendental algebraic equations that cannot be solved exactly.

However, the minimization problem for the case of *linear* variational coefficients can be solved exactly by matrix diagonalization. For example, let $\{\chi_p | p = 1, \ldots, N\}$ be a finite basis set of N arbitrarily chosen functions (subject to the boundary conditions and normalizability) that need have nothing to do with the exact Ψ_i, and write Ψ_{tr} in the form

$$\Psi_{\mathrm{tr}} = \sum_{p=1}^{N} a_p \chi_p. \tag{114}$$

Now the variational parameters a_p enter linearly, and the set of variational conditions (113) becomes exactly equivalent to the N-dimensional generalized eigenvalue problem

$$\mathbf{Ha} = \lambda \mathbf{Oa}, \tag{115}$$

where $\mathbf{a}$ is a column vector of coefficients a_p, and $\mathbf{H}$ and $\mathbf{O}$ have matrix elements $H_{pq} = \langle \chi_p | H | \chi_q \rangle$ and $O_{pq} = \langle \chi_p | \chi_q \rangle$. There are N eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, of which the lowest is an upper bound to E_1.

Equation (115) is equivalent to the original Schrödinger equation (103) only if the basis set $\{\chi_p\}$ is complete, and in general this requires taking the limit $N \to \infty$. The significance of Eq. (115) is that it provides a computationally useful means of obtaining approximate solutions, even if the complete basis set of functions $\{\chi_p\}$ is truncated at some finite number N, and the lowest eigenvalue provides an upper bound that systematically decreases toward the exact E_1 as N is increased.

As discussed in the following section, the bounds apply not just to the ground state but also to the lower-lying excited states.

3.3.2 Extension to Excited States

By the Hylleraas – Undheim – MacDonald (HUM) theorem (see Hylleraas and Undheim, 1930; MacDonald, 1933), the remaining eigenvalues $\lambda_2, \lambda_3, \ldots$ are also upper bounds to the exact energies $E_2, E_3, \ldots$, provided that the spectrum is bounded from below. The HUM theorem is a consequence of the matrix eigenvalue interleaving theorem, which states that as the dimensions of **H** and **O** are progressively increased by adding an extra row and column, the N old eigenvalues λ_p fall between the $N+1$ new ones. Consequently, as illustrated in Fig. 3, all eigenvalues numbered from the bottom up must move inexorably downward as N is increased. Since the exact spectrum of bound states is obtained in the limit $N \to \infty$, no λ_p can cross the corresponding exact E_p on its way down. Thus $\lambda_p \geq E_p$ for every finite N.

If the exact Ψ_i can be formed from a linear combination of the χ_p included in the finite basis set, then the result of the variational calculation is the exact Ψ_i and E_i. Otherwise, one obtains the best variational approximation provided by the particular χ_p chosen. If the χ_p basis set becomes asymptotically complete as $p \to \infty$, then convergence to the correct answer is assured.

The HUM theorem no longer applies directly to the relativistic Dirac equation or similar problems because the spectrum is not bounded from below. However, finite basis-set methods can still be applied, provided that sufficient care is taken in their construction (see Drake and Goldman, 1988; Grant, 1996).

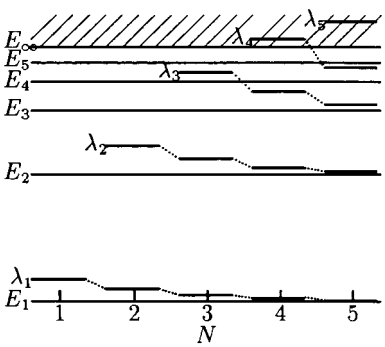


Fig. 3 Diagram illustrating the HUM theorem. The λ_p, $p = 1, \ldots, N$, are the variational eigenvalues for an N-dimensional basis set, and the E_i are the exact eigenvalues of H. The highest λ_p lie in the continuous spectrum of H

3.3.3 Variational Lower Bound

If the Rayleigh – Ritz method is applied to the integral

$$\int \Psi^*(H-E)(H-E_2^<)d^3r, \qquad (116)$$

then the quantity

$$E^< = \frac{\langle \Psi_{\rm tr}|H(H-E_2^<)|\Psi_{\rm tr}\rangle}{\langle \Psi_{\rm tr}|(H-E_2^<)|\Psi_{\rm tr}\rangle} \qquad (117)$$

is made stationary. If the quantity $E_2^<$ is chosen to be a lower bound on the energy E_2 of the first excited state, then by an argument similar to that leading to Eq. (112), $E^<$ is a lower bound on the ground-state energy E_1, called the Temple lower bound (Temple, 1928) and denoted by $E_1^<$. In fact, if $\Psi_{\rm tr}$ is expanded as in Eq. (108), then after some algebra, $E^<$ from Eq. (117) becomes

$$E_1^< = E_1 + \frac{\sum_{i=2}^{\infty} |c_i|^2 (E_i - E_1)(E_i - E_2^<)}{E_1 - E_2^< + D}, \qquad (118)$$

where

$$D = \sum_{i=2}^{\infty} |c_i|^2 (E_i - E_1) = E_1^{>} - E_1, \tag{119}$$

and $E_1^{>}$ is the variational upper bound on E_1. The numerator of the fraction in Eq. (118) is positive and the denominator is negative, provided that $E_1^{>} < E_2^{<}$, thus proving the bound. However, generally speaking, $E_1^{<}$ is much less accurate than $E_1^{>}$.

3.3.4 Illustrative Results for Helium

Application of the variational method to helium by Hylleraas (1928, 1929) played an important role in the early history of quantum mechanics because it provided the first test of the Schrödinger equation in a system more complicated than hydrogen. With its two electrons orbiting the nucleus, helium is the simplest atomic system that cannot be adequately described by the older Bohr – Sommerfeld quantum theory.

The Hamiltonian for helium (in the limit of infinite nuclear mass) is

$$H = -\sum_{i=1}^{2} \left(\frac{\hbar^2}{2m_e} \nabla_i^2 + \frac{e^2}{r_i} \right) + \frac{e^2}{r_{12}}, \tag{120}$$

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$ is the electron – electron separation and Ze is the nuclear charge. The e^2/r_{12} term represents the Coulomb repulsion between the two electrons. Without this term, the Schrödinger equation would be separable, and the exact solution (including permutational symmetry) would be of the form

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2) \pm \psi_1(\mathbf{r}_2)\psi_2(\mathbf{r}_1), \tag{121}$$

where $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$ are exactly known hydrogenic wave functions, depending on the state in question. The Hartree – Fock approximation corresponds to the best variational representation that can be written in the form of a separable product with $\psi_1(\mathbf{r})$ and $\psi_2(\mathbf{r})$ regarded as arbitrary functions of r. However, even this is in error for the ground-state energy of -2.903724 a.u. [the atomic unit (a.u.) of energy is $e^2/a_0 = 27.211\,396$ eV, where a_0 is the Bohr radius] by 0.0247 a.u., or 0.673 eV. This difference, called the *correlation energy*, is much larger than typical chemical energies.

To obtain a better representation, Hylleraas suggested constructing a trial solution of the form

$$\Psi_{\text{tr}} = \sum_{i,j,k} a_{ijk} r_1^i r_2^j r_{12}^k \exp(-\alpha r_1 - \beta r_2), \tag{122}$$

which is of the form of a hydrogenic product $\psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2)$, except that it contains explicit powers of r_{12} and is therefore not separable. The a_{ijk} are the linear variational parameters, and α and β are additional nonlinear parameters that can be separately varied to optimize the energy. Detailed formulas for the necessary matrix elements are given by Drake (1996). This basis set is provably complete in the limit of large $i_{\max}, j_{\max}$, and $k_{\max}$ (Klahn and Bingel, 1977). Typically, all combinations of powers are included in the basis set such that, for electrons with angular momenta l_1 and l_2,

$$i + j + k - l_1 - l_2 \leq \Omega, \tag{123}$$

where Ω is an integer that is progressively increased until adequate convergence is obtained. Without further truncation, the number of terms obtained in this way is

$$N = \tfrac{1}{6}(\Omega + 1)(\Omega + 2)(\Omega + 3). \tag{124}$$

The effect of including powers of r_{12} is dramatic and immediate. The

Tab. 1 Energies for the ground state of helium obtained with various powers of r_{12} in the basis set

r_{12} Terms	Energy (a.u.)	Error (eV)
No r_{12}	−2.879 029	0.672
r_{12}^2	−2.900 503	0.087 6
r_{12}^2, r_{12}^4	−2.902 752	0.026 4
r_{12}	−2.903 496	0.006 20
r_{12}, r_{12}^3	−2.903 700	0.000 65
All r_{12}	−2.903 724	0.000 00

Hartree – Fock approximation corresponds to the limit of large $i_{\max}$ and $j_{\max}$ with $k_{\max} = 0$. As shown in Table 1, an increase of $k_{\max}$ to 1 reduces the error in E to only 0.006 20 eV, thereby accounting for 99% of the correlation energy. The results in Table 1 also demonstrate that the odd powers of r_{12} are much more effective than the even powers. This can be understood from the fact that r_{12}^2 can be written in the form

$$r_{12}^2 = r_1^2 + r_2^2 - 2r_1 r_2 \cos\theta_{12}, \qquad (125)$$

where θ_{12} is the angle between the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Thus r_{12}^2 is only linear in $\cos\theta_{12}$, while an expansion of $r_{12} = (r_{12}^2)^{1/2}$ contains all powers of $\cos\theta_{12}$.

Calculations of this type have been performed by many authors (see Drake, 1993a, for a review), and they have reached a high degree of sophistication. Problems typical of all variational calculations are a dramatic and progressive loss of accuracy for the more highly excited states, and numerical linear dependence in the basis set as it is enlarged. These problems can be avoided by doubling the basis set so that it contains a second set of terms with the same powers but different scale factors α and β. A complete optimization with respect to the α's and β's then leads to a natural partition of the basis set into two sectors with one representing the asymptotic form of the wave function and the other representing complex inner correlation effects. The results obtained by this method are essentially exact for all practical purposes. The nonrelativistic energies are known to better than one part in 10^{16} for the entire singly excited spectrum. A sample of results for the low-lying states is given in Table 2. The indicated convergence was obtained by progressively increasing Ω up to 17, corresponding to about 1700 terms in the doubled basis set. Table 3 shows an example of the convergence for the ground state. [For the case of S states, the basis-set sizes are smaller than indicated by Eq. (124) because terms with $i > j$ can be omitted by symmetry.] The ratios of successive differences in the last column provide a convenient method to monitor the convergence of the eigen-value. They show that the differences themselves decrease in a fairly smooth and uniform fashion with increasing Ω. These high-precision results provide a benchmark for comparison with other less accurate methods of

Tab. 2 Nonrelativistic energies for several states of helium in the limit of infinite nuclear mass

State	Energy (a.u.)
$1s^2\ {}^1S$	−2.903 724 377 034 119 60(2)
$1s2s\ {}^1S$	−2.145 974 046 054 419(2)
$1s2s\ {}^3S$	−2.175 229 378 236 791 307(6)
$1s2p\ {}^1P$	−2.123 843 086 498 101 35(5)
$1s2p\ {}^3P$	−2.133 164 190 779 283 17(3)
$1s3s\ {}^1S$	−2.061 271 989 740 911(5)
$1s3s\ {}^3S$	−2.068 689 067 472 457 192(1)
$1s3p\ {}^1P$	−2.055 146 362 091 943 33(7)
$1s3p\ {}^3P$	−2.058 081 084 274 275 3(2)
$1s3d\ {}^1D$	−2.055 620 732 852 246 51(8)
$1s3d\ {}^3D$	−2.055 636 309 453 261 34(4)

Tab. 3 Convergence study for the ground-state energy of helium (in atomic units). The numbers in the last column give the ratios of successive differences

Ω	$N_{tot}(\Omega)$	$E(\Omega)$	$R(\Omega)$[a]
4	44	−2.903 724 131 001 531 810	
5	67	−2.903 724 351 566 477 006	
6	98	−2.903 724 373 891 109 909	9.88
7	135	−2.903 724 376 548 959 510	8.40
8	182	−2.903 724 376 960 412 587	6.46
9	236	−2.903 724 377 018 168 462	7.12
10	302	−2.903 724 377 030 786 217	4.58
11	376	−2.903 724 377 033 426 037	4.78
12	464	−2.903 724 377 033 966 492	4.88
13	561	−2.903 724 377 034 076 500	4.91
14	674	−2.903 724 377 034 107 875	3.51
15	797	−2.903 724 377 034 116 019	3.85
16	938	−2.903 724 377 034 118 518	3.26
17	1090	−2.903 724 377 034 119 239	3.47
18	1262	−2.903 724 377 034 119 479	3.01
Extrapolation		−2.903 724 377 034 119 597(15)	

[a] $R(\Omega) = [E(\Omega - 1) - E(\Omega - 2)]/[E(\Omega) - E(\Omega - 1)]$.

calculation such as Hartree–Fock and configuration interaction. Results for many other states are given by Drake (1993b, 1994).

A comparison of the results in Table 2 with experiment is meaningful only after corrections for finite nuclear mass, special relativity, and quantum-electrodynamic (QED) effects (such as electron self-energy and vacuum polarization) are taken into account. A detailed discussion of these corrections can be found in Drake (1993b, 1994). When they are included, the calculated transition frequencies agree to within the estimated accuracy of the QED shift. If the measurements are expressed in terms of ionization energies for the various states, then their accuracies range from ±30 MHz ($\pm 5 \times 10^{-9}$ a.u.) for the ground state to ±0.1 MHz ($\pm 1.5 \times 10^{-11}$ a.u.) for the higher-lying P and D states. Since the nonrelativistic energies in Table 2 are much more accurate than this, the comparison with experiment is primarily a test of higher-order contributions to the QED shift (two-electron Lamb shift), which is the dominant source of uncertainty in the calculations.

As one example, the calculated ionization energy of the $1s2s\,^1S$ state is (Drake et al., 1993)

$$960\,332\,039.4 \pm 1 \text{ MHz}.$$

Of this total, -2808.5 ± 1 MHz comes from the calculated QED shift. For comparison, the two experimental values are

$$960\,332\,041.52 \pm 0.21 \text{ MHz},$$

$$960\,332\,040.87 \pm 0.15 \text{ MHz}.$$

The first is obtained from an extrapolation of the $1s2s\ ^1S$–$1snp\ ^1P$ transition frequencies to the series limit (Sansonetti and Gillaspy, 1992), and the second from

the $1s2s\,^1S$–$1snd\,^1D$ two-photon transition frequencies (Lichten et al., 1991). Although the measurements do not quite agree with each other, taken together they determine the QED shift of the $1s2s^1S$ state to an accuracy of about 100 parts per million and verify the calculated value to better than 0.1%. For the ground state, the calculated QED shift in the ionization energy has the much larger value $-(41\,233 \pm 35)$ MHz. This has recently been verified to an accuracy of ± 45 MHz ($\pm 0.1\%$) from the total $1s^2\,^1S$–$1s2p\,^1P$ transition frequency (Eikema et al., 1996).

In summary, the results in Table 2 provide a firm foundation of nonrelativistic energies upon which higher-order corrections can be built and compared with experiment. Further improvements in the QED part of the theory remain an important challenge for the future.

3.3.5
Extensions to More Complex Systems

Fully correlated variational calculations of the type described in the previous section are difficult to extend to systems more complex than helium because of the rapid increase in the number of terms required. For an atom containing K electrons, there are K single-particle radial distances r_s and $K(K-1)/2$ interparticle distances r_{st} for a total of $P = K(K+1)/2$ radial coordinates. If all combinations of powers of the r_s and r_{st} are included in the basis set such that the sum of powers is $\leq \Omega$ [cf. Eq. (123)], then the generalization of Eq. (124) for the number of terms is

$$N = \frac{(\Omega+1)(\Omega+2)\cdots(\Omega+P)}{P!}. \quad (126)$$

Since the time required to calculate a single eigenvector increases in proportion to N^3, the overall complexity of the calculation increases roughly in proportion to

$$\left[\frac{6(\Omega+P)!}{P!(\Omega+3)!}\right]^3 \quad (127)$$

relative to helium with the same Ω.

As an example, from Table 3, an accuracy of 10^{-10} a.u. for the ground state of helium requires $\Omega = 8$. A similar accuracy for lithium with $K = 3$ and $P = 6$ therefore requires about 6000 times the computer resources, and for beryllium with $K = 4$ and $P = 10$, the factor from expression (127) becomes 1.4×10^{13}.

Because of this rapid increase of complexity with the number of electrons, fully correlated calculations of spectroscopic accuracy have only been extended as far as lithium (see Yan and Drake, 1995; Yan et al., 1996; and earlier references therein). The pattern of convergence for the ground state is similar to that shown in Table 3. The results up to $\Omega = 8$ yield the extrapolated nonrelativistic eigenvalue

$$E(1s^2 2s^2\ S) = -7.478\,060\,323\,10(31)\ \text{a.u.} \quad (128)$$

The uncertainty of $\pm 3 \times 10^{-10}$ is about what one would expect from Table 3 for $\Omega = 8$.

For systems more complex than lithium, one must resort to other methods of calculation that can be extended to arbitrarily complex systems, but typically having much lower accuracy ($\pm 10^{-6}$ a.u. or more). These methods include multiconfiguration Hartree–Fock (MCHF), configuration-interaction (CI), many-body perturbation-theory, finite-element, diffusion Monte Carlo (DMC), and variational Monte Carlo (VMC) techniques. The MCHF and CI methods are similar in concept to the fully correlated variational method described in Sec. 3.3.4, except that the members of the basis

set χ_p are constructed from antisymmetrized products of one-electron orbitals corresponding to definite electronic configurations. The effect is analogous to including only the *even* powers of r_{12} as shown in Table 1, and so convergence with increasing angular momentum of the individual electrons is slow. Recently, Goldman (1994) has devised a modified CI method involving extrapolation procedures to overcome this problem, at least for simple systems. For recent work on finite-element and many-body perturbation-theory methods, see Ackermann (1995) and Plante et al. (1994), respectively.

The DMC and VMC Monte Carlo methods attempt to reduce the complexity problem for more complex systems by the use of random-sampling techniques. The DMC method takes advantage of the fact that the time-dependent Schrödinger equation is formally identical to the diffusion equation in imaginary time, and for large imaginary time, an arbitrary starting solution quickly decays to the ground state (see, e.g., Moskowitz et al., 1982; Barnett et al., 1995; and earlier references therein). A random sampling of initial configurations is then propagated forward in time to construct the wave function. The VMC method is more closely related to the standard variational techniques discussed in Sec. 3.3.4. The idea is to define a trial wave function Ψ_{tr} in terms of variational parameters, as in Sec. 3.3.4, and then to optimize them over a statistical distribution of sample points $\mathbf{r}_i$ by minimizing an expression for the variance such as

$$\frac{\sum_i (H\,\Psi_i - E_{\mathrm{ref}}\Psi_i)^2/w_i}{\sum_i \Psi_i^2/w_i}, \tag{129}$$

or the energy variance given by

$$\frac{\sum_i (H\,\Psi_i - E_{\mathrm{ref}}\Psi_i)^2\Psi_i^2/w_i^2}{\left[\sum_i \Psi_i^2/w_i\right]^2}. \tag{130}$$

Here, $\Psi_i = \Psi_{\mathrm{tr}}(\mathbf{r}_i)$ is the trial wave function evaluated at some particular set of values for the electronic coordinates collectively denoted by $\mathbf{r}_i$, and the weight function $w_i = w(\mathbf{r}_i)$ is the probability of choosing $\mathbf{r}_i$ if the sampling is nonuniform. The optimum strategy is to bias the sampling according to the value of a guiding function $g(\mathbf{r}_i)$ that resembles the actual Ψ^2 as closely as possible and to choose the reference energy E_{ref} as close as possible to the desired eigenvalue. Although the method could be applied to a direct optimization of $\langle H\rangle$, the advantage gained by optimizing the variance is that the sample space required for a given accuracy is much smaller. Several sample problems and illustrative examples are discussed by Alexander et al. (1991).

3.4 Variation–Perturbation Methods

For many problems, it is advantageous to split the Hamiltonian into two parts according to

$$H = H^{(0)} + gV, \tag{131}$$

where the eigenvalue problem for $H^{(0)}$ can be solved exactly (or to high precision), and V is a perturbation whose strength is controlled by the parameter g. If the wave functions and energies are similarly expanded,

$$\Psi = \Psi^{(0)} + g\Psi^{(1)} + g^2\Psi^{(2)} + \cdots, \tag{132}$$

$$E = E^{(0)} + gE^{(1)} + g^2E^{(2)} + \cdots, \tag{133}$$

and substituted into the Rayleigh–Ritz quotient (106), then the terms linear in g give

$$E^{(1)} = \frac{1}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}[\langle\Psi^{(0)}|V|\Psi^{(0)}\rangle + 2\langle\Psi^{(0)}|H^{(0)} - E^{(0)}|\Psi^{(1)}\rangle]. \quad (134)$$

This is stationary with respect to variations $\delta\Psi^{(0)}$ if $\Psi^{(1)}$ satisfies the first-order perturbation equation

$$(H^{(0)} - E^{(0)})|\Psi^{(1)}\rangle + (V - E^{(1)})|\Psi^{(0)}\rangle = 0. \quad (135)$$

Since by assumption

$$H^{(0)}|\Psi^{(0)}\rangle = E^{(0)}|\Psi^{(0)}\rangle, \quad (136)$$

it follows from Eq. (134) that

$$E^{(1)} = \frac{\langle\Psi^{(0)}|V|\Psi^{(0)}\rangle}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}. \quad (137)$$

The entire series of perturbation equations to all orders can be similarly generated from the Rayleigh–Ritz variational principle. Computational methods based on these results were first developed by Slater and Kirkwood (1931), and by Dalgarno and Lewis (1955, 1956) (see also Dalgarno and Stewart, 1956; Sternheimer, 1951, 1954, 1957; Schwartz, 1959). They have since been employed by numerous other authors for a wide variety of problems.

3.4.1 Variational Bounds

A particular advantage of the variational derivation of the perturbation equations is its use in establishing bounds (see, e.g., Glover and Weinhold, 1976). For example, consider the second-order energy

$$E^{(2)} = \frac{1}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}\Big[2\langle\Psi^{(0)}|V|\Psi^{(1)}\rangle + 2\langle\Psi^{(0)}|H^{(0)} - E^{(0)}|\Psi^{(2)}\rangle + \langle\Psi^{(1)}|H^{(0)} - E^{(0)}|\Psi^{(1)}\rangle\Big], \quad (138)$$

from terms quadratic in g. $E^{(2)}$ is stable with respect to variations $\delta\Psi^{(0)}$ if $\Psi^{(2)}$ satisfies the second-order perturbation equation

$$(H^{(0)} - E^{(0)})|\Psi^{(2)}\rangle + (V - E^{(1)})|\Psi^{(1)}\rangle = E^{(2)}|\Psi^{(0)}\rangle, \quad (139)$$

from which it follows that

$$E^{(2)} = \frac{\langle\Psi^{(0)}|V - E^{(1)}|\Psi^{(1)}\rangle}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}. \quad (140)$$

However, $\Psi^{(1)}$ is typically not known exactly and must be approximated in some way. From Eq. (138), $E^{(2)}$ is stable with respect to variations $\delta\Psi^{(1)}$ if $\Psi^{(1)}$ satisfies the first-order equation (135). Since the total $E = E^{(0)} + gE^{(1)} + g^2E^{(2)} + \cdots$ is an upper bound for sufficiently small g, and $E^{(0)}$ and $E^{(1)}$ are known exactly (by assumption), it follows that the value of $E^{(2)}$ calculated from Eq. (138) with some approximate $\Psi^{(1)}_{\mathrm{tr}}$ must be an upper bound.

In particular, if $\Psi^{(1)}_{\mathrm{tr}}$ is expanded in a finite basis set of functions χ_p, as in Eq. (114), then the variational condition $\partial E^{(2)}/\partial b_p = 0$ for the expansion coefficients b_p yields the set of N linear algebraic equations [cf. Eq. (115)]

$$(\mathbf{H}^{(0)} - E^{(0)}\mathbf{O})\mathbf{b} + \mathbf{V}^E = 0, \quad (141)$$

where $\mathbf{V}^E$ is a column vector with elements $V^E_p = \langle\chi_p|V - E^{(1)}|\Psi^{(0)}\rangle$. [This works correctly if $\langle\chi_p|\Psi^{(0)}\rangle = 0$ by symmetry. However, if $\Psi^{(0)}$ can be expressed as a linear combination of the χ_p, then these equations are singular. In that case, it is

sufficient just to delete one of the equations to obtain a nonsingular set.] It then follows that

$$\langle\Psi^{(1)}|(H^{(0)} - E^{(0)})|\Psi^{(1)}\rangle = -\langle\Psi^{(1)}|(V - E^{(1)})|\Psi^{(0)}\rangle \quad (142)$$

within the finite basis set, and Eq. (138) for $E^{(2)}$ reduces to Eq. (140).

3.4.2 Spectral Representations and Pseudostates

Although computationally less efficient, it is instructive and sometimes useful to transform the χ_p basis set to a new basis $\tilde{\chi}_p$ that diagonalizes $\mathbf{H}$. The matrix elements in the transformed basis are then

$$\langle\tilde{\chi}_p|H|\tilde{\chi}_q\rangle = \tilde{E}_p\delta_{p,q}, \quad (143)$$

$$\langle\tilde{\chi}_p|\tilde{\chi}_q\rangle = \delta_{p,q}. \quad (144)$$

The eigenvectors $\tilde{\chi}_p$ with eigenvalues $\tilde{E}_p$ are called *pseudostates* that form a discrete variational representation of the actual spectrum of the system (including the continuous spectrum of $H^{(0)}$; see Fig. 3). Although the diagonalization step is computationally slow, the advantage gained is that the system of equations (141) is brought into diagonal form with the immediate solutions

$$\tilde{b}_p = \frac{\langle\tilde{\chi}_p|V - E^{(1)}|\Psi^{(0)}\rangle}{E^{(0)} - \tilde{E}_p}, \; p = 1, 2, \ldots, N. \quad (145)$$

A term with $\tilde{E}_p = E^{(0)}$, if present, is simply omitted. Without loss of generality, one can then assume that $\langle\tilde{\chi}_p|\Psi^{(0)}\rangle = 0$ for the remaining $\tilde{\chi}_p$. Equation (140) then becomes

$$E^{(2)} = \sum_{p=1}^{N}{}' \frac{|\langle\tilde{\chi}_p|V - E^{(1)}|\Psi^{(0)}\rangle|^2}{E^{(0)} - \tilde{E}_p}, \quad (146)$$

where the prime denotes that terms with $\tilde{E}_p = E^{(0)}$ are to be omitted. This expression for $E^{(2)}$ is formally identical to the standard second-order perturbation expression, except that the summation over the actual spectrum of $H^{(0)}$ (including an integration over the continuum) is replaced by a summation over the discrete variational pseudostates. The results are completely equivalent to those obtained by solving Eqs. (141) directly. A similar formal identity can be extended to all orders of perturbation theory.

From a more general point of view, the above results correspond to a discrete variational representation for the Green's function of complex variable z, defined by

$$(H^{(0)} - z)G^{(0)}(\mathbf{r}, \mathbf{r}') = \delta(\mathbf{r} - \mathbf{r}'). \quad (147)$$

Then

$$G^{(0)}(\mathbf{r}, \mathbf{r}') \simeq \sum_{p=1}^{N} \frac{|\tilde{\chi}(\mathbf{r})\rangle\langle\tilde{\chi}(\mathbf{r}')|}{\tilde{E}_p - z}. \quad (148)$$

The great advantage is that often a small number of appropriately chosen pseudostates can replace the infinite summation over bound states plus an integration over the continuous spectrum of $H^{(0)}$.

The pseudostate method has important applications in scattering theory as an alternative way to represent the scattering continuum. For example, the convergent close-coupling method of Bray and Stelbovics (1992) has yielded essentially exact solutions to the electron–hydrogen scattering problem.

3.4.3 Example: The Polarizability of Hydrogen

Consider the problem of a hydrogen atom in its 1s ground state subjected to a static electric field of strength F pointing in the

z direction. The total Hamiltonian in polar coordinates is then

$$H = H^{(0)} + eFr\cos\theta. \tag{149}$$

With $V = eFr\cos\theta$ as the perturbation, the first-order equation (135) can be solved analytically with the result

$$\Psi^{(1)} = -\left(\frac{1}{\sqrt{3}}\right)(2r + r^2)e^{-r}Y_1^0(\hat{\mathbf{r}}). \tag{150}$$

Since V is of odd parity, $E^{(1)} = 0$, and, from Eq. (140), $E^{(2)} = -\frac{9}{4}a_0^3$, where $E^{(2)}$ is the coefficient of F^2 in the energy expansion. By definition, the dipole polarizability is

$$\alpha_{\mathrm{d}} \equiv -2E^{(2)} = \tfrac{9}{2}a_0^3. \tag{151}$$

If $E^{(2)}$ is written in the form of Eq. (146), summed over the actual spectrum of hydrogen, then nearly half of α_{d} comes from virtual transitions to the continuum.

Suppose now that a variational solution is constructed of the form

$$\Psi_{\mathrm{tr}}^{(1)} = -\left(\frac{1}{\sqrt{3}}\right)(b_1 r + b_2 r^2)e^{-\lambda r}Y_1^0(\hat{\mathbf{r}}), \tag{152}$$

where b_1 and b_2 are linear variational parameters, and λ is an additional nonlinear variational parameter. This provides a two-dimensional basis set, with the exact solution being recovered for the case $\lambda = 1$. For $\lambda \neq 1$, the basis set provides the best variational representation of $\Psi^{(1)}$. After solving Eq. (141) for b_1 and b_2, the expression for α_{d} as a function of λ becomes

$$\alpha_{\mathrm{d}}(\lambda) = 6\lambda^5\left(\frac{2}{\lambda+1}\right)^{12} \times \frac{9\lambda^4 - 12\lambda^3 + 14\lambda^2 - 10\lambda + 5}{5\lambda^4 - 10\lambda^3 + 18\lambda^2 - 10\lambda + 5}. \tag{153}$$

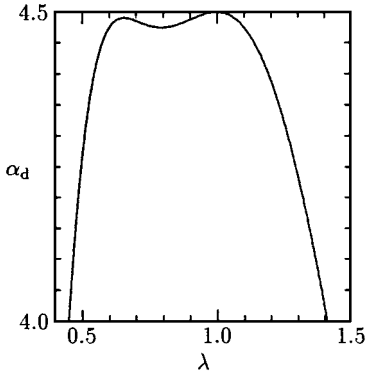


Fig. 4 Variational polarizability α_d of hydrogen, in units of a_0^3. The exact value at $\lambda = 1$ is $\alpha_d = 4.5a_0^3$

A graph of this function near its peak is displayed in Fig. 4. Since $E^{(2)}$ is an upper bound, α_{d} is a lower bound for any λ. It is clear that the exact value of $4.5a_0^3$ is recovered at $\lambda = 1$, where there is an absolute maximum. But what is also significant is the broad region of stability for $0.6 \leq \lambda \leq 1.1$, with a second local maximum near $\lambda = 0.657$. For any λ in this region, α_{d} is in error by no more than 0.6%, even with a basis set containing only two terms. In other words, the entire spectrum of hydrogen is being well represented by just two pseudostates, neither of which corresponds to a physical state of hydrogen. In fact, the accuracy is considerably worse at $\lambda = \frac{1}{2}$, corresponding to the physical $2p$ state. As higher powers of r are added to the basis set, the region of stability rapidly becomes broader and flatter.

3.4.4
Time-Dependent Perturbations

Similar methods can be used to treat the case where the perturbation V is time dependent. For example, with the replacement $E^{(0)} \to i\hbar\partial/\partial t$, the variational

condition

$$\delta\left[\langle\Psi^{(1)}(t)\left|H^{(0)}-i\hbar\frac{\partial}{\partial t}\right|\Psi^{(1)}(t)\rangle\right.$$
$$\left.+2\langle\Psi^{(1)}(t)|V(t)|\Psi^{(0)}(t)\rangle\right]=0 \quad (154)$$

with respect to $\Psi^{(1)}(t)$ leads to the first-order time-dependent perturbation equation

$$\left(H^{(0)}-i\hbar\frac{\partial}{\partial t}\right)|\Psi^{(1)}(t)\rangle$$
$$+V(t)|\Psi^{(0)}(t)\rangle=0. \quad (155)$$

This can be solved by Dirac's method of variation of constants. Many other techniques have been developed, but these will not be further pursued here. See, for example, Dalgarno (1966).

4 The General Sturm–Liouville Problem

Many of the variational techniques discussed in Sec. 3 were developed long before the invention of quantum mechanics, in connection with boundary-value problems in classical physics such as vibrating membranes. Any linear second-order differential equation (of which the radial Schrödinger equation is just one example) can be written in the Sturm – Liouville form

$$\frac{d}{dx}\left[K(x)\frac{dy}{dx}\right]-G(x)y=0, \quad (156)$$

defined over some closed interval $[a,b]$, together with suitable boundary conditions. By application of the Euler – Lagrange equation with fixed end points, this equation follows from the variational condition

$$\delta\int_a^b\left[K(x)\left(\frac{dy}{dx}\right)^2+G(x)y^2\right]dx, \quad (157)$$

and so all the techniques discussed thus far can be applied. With the choice

$$G(x)=-\lambda g(x)+l(x), \quad (158)$$

the Sturm – Liouville problem becomes an eigenvalue problem with λ adjusted to satisfy the boundary conditions.

4.1 The Oscillation Theorem

A great many theorems have been proven concerning the solutions to Sturm – Liouville problems (see, e.g., Ince, 1956). Of particular importance for physical applications is the oscillation theorem. Suppose that $K(x), g(x)$, and $l(x)$ are all continuous, real, positive, monotonic decreasing functions of x in the interval $[a,b]$. It can then be proved that the two-point eigenvalue problem

$$\frac{d}{dx}\left[K(x)\frac{dy}{dx}\right]+[\lambda g(x)-l(x)]y=0 \quad (159)$$

has an infinite sequence of increasing eigenvalues $\lambda_1, \lambda_2, \ldots$, with corresponding eigenvectors $y_1(x), y_2(x), \ldots$ such that $y_m(x)$ has exactly $m-1$ zeros in the open interval $[a,b]$. The eigenvalues are entirely discrete. If $g(x)$ changes sign in the interval, then the sequence of eigenvalues becomes doubly infinite with both an increasing ($\lambda_m^{(+)}$) and a decreasing ($\lambda_m^{(-)}$) set. In either case, the solutions are orthogonal with respect to the weight function $g(x)$.

The importance of this theorem (and its extensions) is that the $y_m(x)$ form the basis for a generalized Fourier series in terms of which an arbitrary function $f(x)$ can be

expanded in the form

$$f(x) = \sum_{m=1}^{\infty} c_m y_m(x) \tag{160}$$

with

$$c_m = \int_a^b f(x) g(x)\, dx. \tag{161}$$

Since the eigenvalues are entirely discrete, there is no integration over a continuum in Eq. (160). Such a basis is called a *Sturmian* basis set. Most of the mathematical apparatus developed for Fourier analysis can be carried over directly. In fact, a Fourier series just corresponds to the choices $K(x) = 1, g(x) = 1, l(x) = 0$.

4.2 Example: The Coulomb Problem

Consider the radial Schrödinger equation

$$\left[-\frac{1}{2r^2}\frac{d}{dr}r^2\frac{d}{dr} + \frac{l(l+1)}{2r^2} - \frac{Z}{r}\right] \times R(r) = ER(r) \tag{162}$$

for an electron moving in the field of a nucleus with charge Z. The quantum-mechanical eigenvalue problem is solved on the interval $[0, \infty]$ by holding Z fixed and varying E such that the boundary conditions

$$\lim_{r\to 0} rR(r) = 0, \lim_{r\to\infty} R(r) = 0, \tag{163}$$

are satisfied for the infinity of bound states with $E < 0$. There is in addition a continuum of scattering solutions with $E > 0$.

The Sturmian eigenvalue problem differs in that E is held fixed at some negative value $-\varepsilon$ with $\varepsilon > 0$, and Z is varied so as to satisfy the boundary conditions. Since the eigenvalues to the Coulomb problem are

$$E_n(Z) = -\frac{Z^2}{2n^2}, \tag{164}$$

it is clear that the Sturmian eigenvalue condition $E_n(Z) = -\varepsilon$ can be satisfied infinitely many times by progressively increasing both n and Z. As Z increases, one eigenvalue after another from the original problem is pulled down through the value $-\varepsilon$. The Sturmian eigenvalues are thus $Z_n = n\sqrt{\varepsilon}$ and the corresponding eigenfunctions are

$$R_{nl}(r) = \frac{1}{(2l+1)!}\left(\frac{(n+l)!}{(n-l-1)!2n}\right)^{\frac{1}{2}} \times (2\alpha)^{\frac{3}{2}}(2\alpha r)^l e^{-\alpha r} \times F(-n+l+1, 2l+2; 2\alpha r), \tag{165}$$

where $\alpha = \sqrt{2\varepsilon}$ and $F(a, b; z)$ is a confluent hypergeometric function. The $R_{nl}(r)$ form a complete set of finite Sturmian polynomials for $n \geq l+1$ that are orthogonal with respect to the potential $1/r$; i.e.

$$\int_0^\infty R_{n'l}(r)\frac{1}{r}R_{nl}(r)r^2 dr = \varepsilon\delta_{n',n}. \tag{166}$$

The Sturmian functions $R_{nl}(r)$ closely resemble the bound-state Coulomb wave functions. The main distinguishing feature is that α is a constant in the exponential factor instead of decreasing as $1/n$. The first N of them differ only by a transformation of the basis set from the functions used to construct a finite variational representation of $\Psi_{\text{tr}}^{(1)}$ in Sec. 3.4.3. The theory of Sturmian functions therefore provides a rigorous foundation for the choice of basis functions in variational calculations, and their property of completeness ensures convergence to the correct answer as N increases.

5
Applications to Electrodynamics

Consider an electromagnetic field propagating through a medium with a charge density $\rho(\mathbf{r})$ moving with velocity $\mathbf{v}(\mathbf{r})$. It follows from Maxwell's equations that the scalar and vector potentials introduced in Sec. 2.4.1 satisfy the equations

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -4\pi\rho\frac{\mathbf{v}}{\mathbf{c}}, \quad (167a)$$

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -4\pi\rho, \quad (167b)$$

provided that the Lorentz gauge condition

$$\nabla\cdot\mathbf{A} + \frac{1}{c}\frac{\partial\phi}{\partial t} = 0 \quad (168)$$

is imposed. Many other gauge choices can be made that leave the physical fields $\mathbf{E}$ and $\mathbf{B}$ invariant, but this one is simplest for the present discussion.

Unlike problems involving point particles, we are now dealing with fields that vary continuously in space. Equations (167a) and (167b) can be derived from a variational principle if a Lagrangian density $\mathscr{L}$ is first defined such that

$$L = \int \mathscr{L}\,dxdydz. \quad (169)$$

The action integral in Hamilton's principle then assumes the four-dimensional form

$$J = \int \mathscr{L}\,dxdydzdt. \quad (170)$$

The condition $\delta J = 0$ is obtained in a manner similar to that leading to Laplace's equation (31). The present case is an application of Eq. (23) with $f = \mathscr{L}$ and four independent variables $t_1 = x$, $t_2 = y$, $t_3 = z$, $t_4 = t$. There will be four equations corresponding to $q_1 = A_1$, $q_2 = A_2$, $q_3 = A_3$, and $q_4 = \phi$. The choice of $\mathscr{L}$ is severely restricted for fields *in vacuo* by the requirement that it be quadratic in the field components (since the field equations are linear), and relativistically invariant. The only quantity satisfying both requirements is a term proportional to $E^2 - B^2$. The inhomogeneous interaction terms on the right-hand sides of Eqs. (167a) are also included with the definition

$$\mathscr{L} = \frac{1}{8\pi}\left[\left(-\nabla\phi - \frac{1}{c}\frac{\partial\mathbf{A}}{\partial t}\right)^2 - (\nabla\times\mathbf{A})^2 - \left(\nabla\cdot\mathbf{A} + \frac{1}{c}\frac{\partial\phi}{\partial t}\right)^2\right] + \rho\left(\mathbf{A}\cdot\frac{\mathbf{v}}{c} - \phi\right), \quad (171)$$

where the first two squared terms correspond to E^2 and $-B^2$, respectively, and the third term generates the Lorentz gauge condition (168). The last term is the interaction term. A straightforward application of Eq. (23) with the terms identified as described following Eq. (170) then yields Eqs. (167a) and (167b).

The Lagrangian density $\mathscr{L}$ gives the equations of motion for the fields in the presence of a predefined matter distribution $\rho(\mathbf{r}, t)$. For comparison, the Lagrangian L defined by Eq. (96) gives the converse equations of motion for particles in the presence of predefined fields $\mathbf{A}$ and ϕ. The remarkable point emerging from a comparison of $\mathscr{L}$ and L is that the matter-field interaction term $\rho(\mathbf{A}\cdot\mathbf{v}/c - \phi)$ in $\mathscr{L}$ is very similar in form to the second and third terms in L. In fact, the terms become identical with the choice $\rho(\mathbf{r}) = q\delta(\mathbf{r} - \mathbf{r}')$, corresponding to a point particle of charge q at position $\mathbf{r}'$. This suggests that the two Lagrangians can be combined into a single Lagrangian

$$L_{\text{tot}} = L_0 + \int \mathscr{L}_0\,dxdydz + L_{\text{int}}, \quad (172)$$

where L_0 and $\mathscr{L}_0$ are the Lagrangians for free particles and free fields, respectively, and L_{int} is the remaining interaction term common to both L and $\int \mathscr{L} d\mathbf{r}$. Hamilton's principle and the Euler–Lagrange equations then give the equations of motion for the combined system of interacting particles and fields.

The above is of course not a proof that the resulting equations of motion provide an exact description of nature. The derivation is based on the supposition that something like Hamilton's principle remains valid for the combined system of interacting particles and fields, and it ignores the quantum nature of both matter and fields. However, Eq. (172) provides a basis for combining a quantized description of matter fields and electromagnetic fields into a single theory called *quantum electrodynamics*, whose predictions have been verified to an extremely high degree of precision (see, e.g., Kinoshita and Yennie, 1990). It can safely be described as the most successful theory ever invented. However, further discussion of this topic would take us beyond the scope of this article (see Further Reading).

6 Feynman's Path Integral

This article would not be complete without at least a passing reference to Feynman's path integral because of the way in which it provides an underlying coherent formalism unifying classical mechanics, quantum mechanics, and optics.

Consider for simplicity the x coordinate of a particle moving in a potential. The aim is to construct a path integral giving the quantum-mechanical transition amplitude for the particle to move from position x_0 at time t_0 to position x_f at time t_f. Let the state vector corresponding to a particle at position x be denoted by $|x\rangle$. In the coordinate representation

$$\langle x'|x\rangle = \delta(x - x'), \tag{173}$$

and by closure,

$$\int |x\rangle\langle x| = \mathbf{1}, \tag{174}$$

where $\mathbf{1}$ is the identity operator. In the momentum representation,

$$\langle p|x\rangle = (2\pi\hbar)^{-\frac{1}{2}} e^{-\frac{px}{\hbar}} = \langle x|p\rangle^*. \tag{175}$$

The remaining ingredient is the time-evolution operator $e^{-iHt/\hbar}$. Its inverse defines states $|x, t\rangle$ in the Heisenberg representation such that

$$|x, t\rangle = e^{\frac{iHt}{\hbar}} |x\rangle. \tag{176}$$

With these preliminaries, an initial expression for the desired transition amplitude is

$$\begin{aligned} K(x_f, t_f; x_0, t_0) &= \langle x_f, t_f|x_0, t_0\rangle \\ &= \langle x_f| \exp\left[-iH\frac{(t_f - t_0)}{\hbar}\right] |x_0\rangle. \end{aligned} \tag{177}$$

The key idea in constructing a path integral is to suppose that the system passes through a large number N of intermediate states $|x_k, t_k\rangle$ in making the above transition from $|x_0, t_0\rangle$ to $|x_f, t_f\rangle$ such that $t_{k+1} = t_k + \varepsilon$, with $\varepsilon = (t_f - t_0)/(N + 1)$. This can be formally achieved by making repeated use of the closure relation (174) to write

$$\begin{aligned} &K(x_f, t_f; x_0, t_0) \\ &\quad = \int dx_1 \cdots dx_N \langle x_f, t_f|x_N, t_N\rangle \\ &\quad \times \langle x_N, t_N|x_{N-1}, t_{N-1}\rangle \cdots \langle x_1, t_1|x_0, t_0\rangle. \end{aligned} \tag{178}$$

Then for each intermediate step, the inner product is

$$\begin{aligned}\langle x_k, t_k | x_{k-1}, t_{k-1}\rangle &= \langle x_k | \exp\left[-i(t_k - t_{k-1})\frac{H}{\hbar}\right] | x_{k-1}\rangle \\ &= \langle x_k | e^{-i\frac{\varepsilon H}{\hbar}} | x_{k-1}\rangle. \end{aligned} \tag{179}$$

Using Eq. (175), this can be evaluated in the momentum representation to obtain (with symmetric or Weyl operator ordering)

$$\langle x_k, t_k | x_{k-1}, t_{k-1}\rangle = \int \frac{dp_k}{2\pi\hbar} \times \exp\left\{i\frac{[p_k \Delta x_k - i\varepsilon H(\bar{x}_k, p_k)]}{\hbar}\right\}, \tag{180}$$

where $\Delta x_k = x_k - x_{k-1}$ and $\bar{x}_k = (x_k + x_{k-1})/2$. Substitution of this form into Eq. (178) then yields

$$K(x_f, t_f; x_0, t_0) = \int dx_1 \cdots dx_N \frac{dp_1}{2\pi\hbar} \cdots \frac{dp_{N+1}}{2\pi\hbar} e^{iS_N}, \tag{181}$$

where

$$S_N = \frac{\varepsilon}{\hbar}\sum_{k=1}^{N+1}\left[\frac{p_k \Delta x_k}{\varepsilon} - H(\bar{x}_k, p_k)\right]. \tag{182}$$

Consider now the limit $N \to \infty$, $\varepsilon \to 0$. Although each x_k separately ranges over all possible values due to the integrations in Eq. (181) (see Fig. 5), the quantity $\Delta x_k/\varepsilon \equiv (x_k - x_{k-1})/\varepsilon$ contributes a large and rapidly varying phase that averages to zero unless $x_k \approx x_{k-1}$. We can therefore identify $\Delta x_k/\varepsilon$ with $\dot{x}$ in the limit $\varepsilon \to 0$ and write

$$\lim_{\varepsilon\to 0} S_N = \frac{1}{\hbar}\int_{t_0}^{t_f} [p\dot{x} - H(x, p)]dt$$

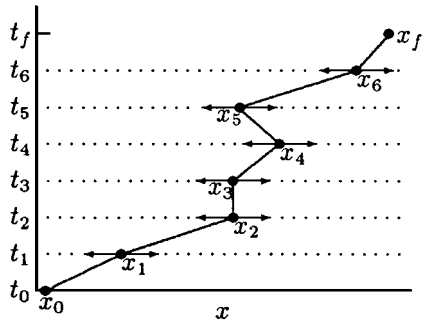


Fig. 5 A possible path for the Feynman integral with six intermediate states. Each of x_1 through x_6 varies independently over all possible values

$$= \frac{1}{\hbar}\int_{t_0}^{t_f} L\, dt. \tag{183}$$

This result provides a version of the Feynman path integral with the path integrated over all possibilities (including discontinuous ones) in phase space. However, the result can be carried a step further by performing the momentum integrations in Eq. (181). For example, if H has the form

$$H(x, p) = \frac{p^2}{2m} + V(x), \tag{184}$$

then a typical momentum integral has the Gaussian form

$$\int_{-\infty}^{\infty} \frac{dp_k}{2\pi\hbar} \exp\left\{-\frac{i\varepsilon}{\hbar}\left[\frac{p_k^2}{2m} - \frac{p_k \Delta x_k}{\varepsilon}\right]\right\} = \left(\frac{m}{2\pi\hbar i\varepsilon}\right)^{\frac{1}{2}} \exp\left[\frac{im\varepsilon}{2\hbar}\left(\frac{\Delta x_k}{\varepsilon}\right)^2\right]. \tag{185}$$

Using this in Eq. (181) gives

$$K(x_f, t_f; x_0, t_0) = \left(\frac{m}{2\pi\hbar i\varepsilon}\right)^{\frac{(N+1)}{2}} \int dx_1 \cdots dx_N e^{iS_N}, \tag{186}$$

where now

$$S_N = \frac{\varepsilon}{\hbar} \sum_{k=1}^{N+1} \left[\frac{m\Delta x_k}{2\varepsilon} - V(\bar{x}_k) \right]. \quad (187)$$

Taking the limit $N \to \infty, \varepsilon \to 0$ as in Eq. (183) leads to the final result

$$K(x_f, t_f; x_0, t_0) = A \int \mathscr{D}x \exp\left(\frac{i}{\hbar} \int_{t_0}^{t_f} L\, dt \right), \quad (188)$$

where A is a constant independent of the path, and $\mathscr{D}x$ is a shorthand notation for the infinitely nested integrals in Eq. (186). Since this structure is mathematically not well defined, any practical calculation must be done with a finite subdivision and the limit $\varepsilon \to 0$ taken at the end.

An explicit calculation of K is possible only for certain special cases such as the free particle or harmonic oscillator. In the free-particle case ($V = 0$), K becomes

$$K(x_f, t_f; x_0, t_0) = \left(\frac{m}{2\pi i\hbar\Delta t} \right)^{\frac{1}{2}} \times \exp\left[-\frac{m(\Delta x)^2}{2i\hbar\Delta t} \right], \quad (189)$$

where $\Delta t = t_f - t_0$ and $\Delta x = x_f - x_0$. This is the known nonrelativistic kernel or Green's-function propagator for a free particle (see, e.g., Merzbacher, 1970). It controls the time evolution of the wave function according to

$$\Psi(x_f, t_f) = \int_{-\infty}^{\infty} K(x_f, t_f; x, t_0)\Psi(x, t_0)\, dx. \quad (190)$$

6.1 Relation to Classical Dynamics

As it stands, the Feynman path integral is not strictly a variational principle. However, in the limit $\hbar \to 0$, the exponential term

$$\frac{i}{\hbar} \int_{t_0}^{t_f} L\, dt \quad (191)$$

in general gives a large phase that varies rapidly with slight variations in the path, except very near the particular path where the integral is stationary; i.e., by Hamilton's principle, the classical trajectory. For any other path, the rapid phase variations cause the contributions to the integral to cancel. All possible paths contribute with equal weight, but the phase cancellation singles out the classical path. It similarly singles out the path of a light ray in geometrical optics by Fermat's principle. The connection between variational principles in classical and quantum mechanics is discussed from a different point of view by Gray et al. (1996), who show that the reciprocal form of the Maupertuis principle of least action (see Sec. 2.3.1) can be obtained from the classical limit of the Schrödinger variational principle without the use of the Feynman path integral.

On the other hand, the path integral embodies quantum mechanics, and, with the use of the Lagrangian density, electrodynamics. It therefore provides a powerful underlying formalism that unifies all of these diverse branches of physics. It provides a general starting point for the construction of quantum field theories and the description of new phenomena. These topics may be further pursued through some of the books in the reading list.

Acknowledgments

The author is grateful to the Natural Sciences and Engineering Research Council of Canada for financial support,

and to Mark Cassar and Alice Kolakowska for a careful reading of the manuscript.

Glossary

Brachistochrone Problem: The problem of finding the curve giving the shortest travel time for a bead constrained to slide on a wire of arbitrary shape.

Calculus of Variations: A mathematical technique for finding the extrema (maxima, minima, or stationary points) with respect to variations in the function defining the path for an integration.

Canonical Momentum: The momentum p_i conjugate to a generalized coordinate q_i, defined by $p_i = \partial L/\partial \dot{q}_i$, where $\dot{q}_i = dq/dt$ and L is the Lagrangian.

Canonical Transformation: A transformation from one set of generalized coordinates and momenta to another such that the form of Hamilton's equations of motion is preserved.

Conservative System: A mechanical system for which the Hamiltonian is a constant of the motion.

Constraint: An equation that must be satisfied by the solution to a variational problem.

Correlation Energy: The difference between the exact and Hartree–Fock energies of a many-particle quantum-mechanical system.

Electrodynamics: The study of electromagnetic fields and their interactions with charged particles, treated as a single dynamical system.

Euler–Lagrange Equation: A second-order differential equation whose solutions determine the function that makes a path integral an extremum.

Fermat's Principle: A principle from geometrical optics stating that the path of a light beam through a medium of varying index of refraction is such that the travel time is a minimum.

Feynman Path Integral: An integral over all possible paths (both continuous and discontinuous ones) connecting the initial and final states of a system, with each path weighted by a phase factor of the form $\exp(iS/\hbar)$, where S is the path integral over the Lagrangian appearing in Hamilton's principle, and $\hbar$ is Planck's constant.

Functional: A function $F[y(x)]$ that depends on the functional form chosen for $y(x)$. For example, $y(x)$ might determine the path of an integration between fixed end points in a plane.

Generalized Coordinate: Any function of the Cartesian coordinates used to describe the evolution of a mechanical system, or its dynamical analogues involving continuous fields.

Generalized Force: The component of the work done by the actual force acting on a particle due to a change of a generalized coordinate.

Generalized Momentum: The momentum conjugate to a generalized coordinate – see Canonical Momentum.

Hamiltonian: A constant of the motion for time-independent systems, which for ordinary conservative systems becomes the sum $T + V$ of the kinetic energy T and the potential energy V.

Hamilton's Characteristic Function: The time-independent part of Hamilton's principal function.

Hamilton's Equations of Motion: A set of $2N$ first-order differential equations linking directly the N generalized coordinates and N generalized momenta through their time derivatives.

Hamilton's Principal Function: The function represented formally by the indefinite integral of the Lagrangian with respect to time. It corresponds to the general solution to the Hamilton–Jacobi equation.

Hamilton's Principle: A variational condition, involving a path integral over the Lagrangian with respect to time, that determines the equations of motion for the system under consideration.

Hartree–Fock Approximation: The best possible approximation to the wave function for a many-particle quantum-mechanical system that can be written in the form of a separable product of independent functions, with a separate function for each particle. The coupled differential equations to be solved follow from the Rayleigh–Schrödinger variational principle.

Holonomic Constraint: An equation of constraint involving the generalized coordinates that can be expressed in integrated form, as opposed to a relation among differential quantities.

Hylleraas–Undheim–MacDonald Theorem: A theorem establishing variational upper bounds for the energies of excited states of a quantum-mechanical system as well as for the ground state.

Lagrange Undetermined Multiplier: A mathematical technique for the introduction of constraints into variational problems.

Lagrange's Equations of Motion: A system of N second-order partial differential equations that determine the equations of motion for the generalized coordinates $q_i = q_i(t)$.

Lagrangian: For ordinary conservative systems, the Lagrangian L is the quantity $T - V$, expressed as a function of the N generalized coordinates q_i and their time derivatives $\dot{q}_i$.

Lagrangian Density: A generalization of the Lagrangian to continuous systems and fields in which the continuous system is first conceived as consisting of discrete elements.

Oscillation Theorem: A theorem applicable to two-point boundary-value problems of the Sturm–Liouville type. The theorem establishes that there is an infinite sequence of eigenvalues whose eigenfunctions have progressively more zeros between the boundary points, and hence progressively more oscillations.

Path Integral: The integral of a function $f(y(x))$ between given points in an xy plane (or its higher-dimensional generalizations) along a path specified by the function $y(x)$.

Perturbation Theory: A technique for the progressive approximation of more difficult problems involving the solution of differential equations (for example), starting from an exactly soluble simpler one. The difference between the two equations is called the perturbation term.

Principle of Least Action: A variational principle in classical dynamics, closely related

to Hamilton's principle, which establishes a direct connection with Fermat's principle in geometrical optics.

Pseudostate: A member of a set of states obtained by diagonalization of the Hamiltonian matrix in a discrete variational basis set.

Quantum Electrodynamics: A quantized field theory describing the dynamical interactions of charged particles with electromagnetic fields.

Rayleigh–Ritz Variational Method: A method for the construction of an approximate wave function Ψ by expansion in a finite basis set of functions with expansion coefficients determined by the Rayleigh–Schrödinger variational principle.

Rayleigh–Schrödinger Variational Principle: A principle stating that the ratio $\langle\Psi|H|\Psi\rangle/\langle\Psi|\Psi\rangle$ is an upper bound to the lowest eigenvalue of H for any arbitrary (but normalizable) choice for the wave function Ψ.

Schrödinger Equation: A second-order partial differential wave equation that forms the basis of nonrelativistic quantum mechanics.

Sturmian Basis Set: The set of discrete eigenvalues and corresponding eigenfunctions obtained by solving a two-point eigenvalue problem of the Sturm–Liouville type.

Sturm–Liouville Problem: A class of second-order differential equations of the form $(d/dx)[K(x)dy/dx] - G(x)y = 0$, together with suitable boundary conditions.

Spectral Representation: A representation of the Green's function, or the terms in a perturbation series, in terms of explicit summations over the eigenvalue spectrum of the unperturbed problem.

Temple Bound: A method for constructing variational lower bounds to the energy based on the square of the Hamiltonian.

Variational Bound: An upper or lower bound on the energy or some other quantity obtained by means of a trial solution to the underlying differential equation, typically with parameters in the trial solution that can be adjusted to obtain the best possible solution.

List of Works Cited

Ackermann, J. (1995), *Phys. Rev. A* **52**, 1968–1975, and earlier references therein.

Alexander, S. A., Coldwell, R. L., Monkhorst, H. J., Morgan, J. D., III (1991), *J. Chem. Phys.* **95**, 6622–6633.

Barnett, R. N., Johnson, E. M., Lester, W. A., Jr. (1995), *Phys. Rev. A* **51**, 2049–2052.

Bray, I., Stelbovics, A. T. (1992), *Phys. Rev. A* **46**, 6995–7011, and earlier references therein.

Cohen, M. R., Drabkin, I. E. (1958), *A Source Book in Greek Science*, Cambridge, MA: Harvard Univ. Press.

Courant, R., Hilbert, D. (1966), *Methods of Mathematical Physics*, New York: Interscience, p. 190.

Dalgarno, A. (1966), in: C. H. Wilcox (Ed.), *Perturbation Theory and Its Applications in Quantum Mechanics*, New York: Wiley, pp. 145–183.

Dalgarno, A., Lewis, J. T. (1955), *Proc. R. Soc. (London) Ser. A* **233**, 70–74.

Dalgarno, A., Lewis, J. T. (1956), *Proc. Phys. Soc. (London) A* **69**, 628–630.

Dalgarno, A., Stewart, A. L. (1956), *Proc. R. Soc. (London) Ser. A* **238**, 269–275.

Drake, G. W. F. (1993a), in: F. S. Levin, D. A. Micha (Eds.), *Long-Range Casimir Forces: Theory and Recent Experiments on Atomic Systems*, New York: Plenum.

Drake, G. W. F. (1993b), *Adv. At. Mol. Opt. Phys.* **31**, 1–62.

Drake, G. W. F. (1994), *Adv. At. Mol. Opt. Phys.* **32**, 93–116.
Drake, G. W. F. (1996), in: G. W. F. Drake (Ed.) *Atomic, Molecular, and Optical Physics Handbook*, New York: American Institute of Physics, pp. 154–171.
Drake, G. W. F., Goldman, S. P. (1988), *Adv. At. Mol. Phys.* **25**, 393–416.
Drake, G. W. F., Khriplovich, I. B., Milstein, A. I., Yelkovski, A. S. (1993), *Phys. Rev. A* **48**, R15–R17.
Eikema, K. S. E., Ubachs, W., Vassen, W., Hogorvorst, W. (1996), *Phys. Rev. A* **55**, 1866–1880.
Glover, R. M., Weinhold, F. (1976), *J. Chem. Phys.* **65**, 4913–4926, and other references therein.
Goldman, S. P. (1997), *Phys. Rev. Lett.* **78**, 2325–2328.
Grant, I. P. (1996), in: G. W. F. Drake (Ed.), *Atomic, Molecular, and Optical Physics Handbook*, New York: American Institute of Physics, pp. 258–286.
Gray, C. G., Karl, G., Novikov, V. A. (1996), *Ann. Phys. (N.Y.)* **251**, 1–25.
Hildebrand, S., Tromba, A. (1985), *Mathematics and Optimal Form*, San Francisco: W. H. Freeman (Scientific American Library).
Hylleraas, E. A. (1928), *Z. Phys.* **48**, 469–494.
Hylleraas, E. A. (1929), *Z. Phys.* **54**, 347–366.
Hylleraas, E. A., Undheim, B. (1930), *Z. Phys.* **65**, 759–772.
Ince, E. L. (1956), *Ordinary Differential Equations*, New York: Dover.
Jeffreys, H., Jeffreys, B. S. (1972), *Methods of Mathematical Physics*, Cambridge, U.K.: Cambridge Univ. Press, p. 315.
Kinoshita, T., Yennie, D. R. (1990), in: T. Kinoshita (Ed.), *Quantum Electrodynamics*, Singapore: World Scientific, and other articles contained therein.
Klahn, B., Bingel, W. A. (1977), *Theoret. Chim. Acta* **44**, 27–43. See also Klahn, B., Bingel, W. A. (1978), *Int. J. Quantum Chem.* **11**, 943–957.
Lichten, W., Shiner, D., Zhou, Z.-X. (1991), *Phys. Rev. A* **43**, 1663–1665.
MacDonald, J. K. L. (1933), *Phys. Rev.* **43**, 830–833.
Merzbacher, E. (1970), *Quantum Mechanics*, New York: Wiley, p. 163.
Moskowitz, J. W., Schmidt, K. E., Lee, M. A., Kalos, M. H. (1982), *J. Chem. Phys.* **77**, 349–355.
Plante, D. R., Johnson, W. R., Sapirstein, J. (1994), *Phys. Rev. A* **49**, 3519–3530, and earlier references therein.
Sansonetti, C. J., Gillaspy, J. D. (1992), *Phys. Rev. A* **45**, R1–R3.
Schwartz, C. (1959), *Ann. Phys. (N.Y.)* **2**, 156–169, 170–177.
Slater, J. C., Kirkwood, J. G. (1931), *Phys. Rev.* **37**, 682–697.
Sternheimer, R. (1951), *Phys. Rev.* **84**, 244–253.
Sternheimer, R. (1954), *Phys. Rev.* **96**, 951–968.
Sternheimer, R. (1957), *Phys. Rev.* **105**, 158–169.
Temple, G. (1928), *Proc. R. Soc. (London) Ser. A* **119**, 276.
Yan, Z.-C., Drake, G. W. F. (1995), *Phys. Rev. A* **52**, 3711–3717, R4316–R4319.
Yan, Z.-C., McKenzie, D. K., Drake, G. W. F. (1996), *Phys. Rev. A* **54**, 1322–1327.

Further Reading

Techniques of the calculus of variations are covered in great detail by Courant and Hilbert (1966), and by Morse, P. M., and Feshbach, H. (1953), *Methods of Theoretical Physics*, New York: McGraw-Hill, along with most other books on the techniques of theoretical physics. The book by Yourgrau, W., and Mandelstam, S. (1968), *Variational Principles in Dynamics and Quantum Theory*, 3rd ed., Philadelphia: Saunders (Dover reprint 1979) provides an interesting and informative historical perspective.

Applications of variational principles to classical mechanics are covered in a very thorough, detailed, and readable manner by Goldstein, H. (1980), *Classical Mechanics*, 2nd ed., Reading, MA: Addison-Wesley. This book also contains a good pedagogical introduction to variational principles for continuous systems and fields. See also Lanczos, C. (1970), *The Variational Principles of Mechanics*, Toronto: University of Toronto Press (Dover reprint 1986).

The development of field theory from variational principles is covered in many recent books such as Ramond, P. (1981), *Field Theory, a Modern Primer*, Menlo Park, CA: Benjamin/Cummings; and Itzykson, C., and Zuber, J.-B. (1980), *Quantum Field Theory*, New York: McGraw-Hill. A good introduction to the Feynman path integral approach is given by Das, A. (1993), *Field Theory, a Path Integral Approach*, Singapore: World Scientific, and more detailed developments are

contained in Rivers, R. J. (1987), *Path Integral Methods in Quantum Field Theory*, Cambridge, U.K.: Cambridge Univ. Press. The original development in Feynman, R. P., and Hibbs, A. R. (1965), *Quantum Mechanics and Path Integrals*, New York: McGraw-Hill, remains an authoritative source.

A wide variety of applications of variational principles in quantum-mechanical calculations can be found in numerous articles throughout Drake, G. W. F. (Ed.) (1996), *Atomic, Molecular, and Optical Physics Handbook*, New York: American Institute of Physics. In addition, the Kohn and Schwinger variational methods for scattering problems are covered in most books on scattering theory, such as Taylor, J. R. (1972), *Scattering Theory: The Quantum Theory on Nonrelativistic Collisions*, New York: Wiley. All of the above contain numerous other references to the literature and are intended only as a guide.

Index

Abel problem 35, 70
Abel transform 85, 87 ff, 100
Abelian groups 191, 207
– algebraic methods 3 ff, 6 ff, 26 ff
– symmetries 570
– topology 601, 604
Abrikosov–Gorkov–Dzyaloshinski–Fradkin theorem 183
abscissae 129, 133, 338
abstract groups 191
abstract quantum logic 455, 466
acceptance frequency 260
accidental attributes 453
accumulation point 593
accuracy, Monte Carlo methods 265
ACM, numerical software 378
Adams formulas 353
adaptive methods 354
adaptive quadrature 343
addition formulas 3,131
adjoint operators 478 f
– differential geometry 142 ff
– Green's functions 185
– quantum logic 441, 466
– stochastic processes 518
advanced Green's functions 165
affine Lie algebras 25
affine varieties 154
after-effect function 536, 543
Airy differential equations 509
algebraic geometry 154
algebraic methods 1–32
algebraic semantics 455, 464 ff
algebraic systems
– numerical methods 290, 304, 320
– variational methods 636
algebraic topology 596 f, 613
algebras
– definitions 21 ff
– symmetries 574 f
algorithm formulation 259
alias 106
aliasing 237
alphabet (totality of letters) 227
alternative direction implicit (ADI) 308
Ampére–Maxwell law 145
amplitude functions 454
analysis, group representations 206
analytic functions 37
analytic geometry 128 ff
analytical methods 33–82
AND operations 420
angular momentum 136
– Green's functions 176
– group theory 209 ff
angular standard form 399
annihilation operators 178, 572
anticommutation 178
antisymmetric tensors 139 ff
antisymmetries 10
applicates 133
applications
– boundary–element method 172 ff
– fractal geometry 116 f
– functional analysis 70 ff
– mathematical modeling 240 f
– transforms 94 ff
– variational methods 634
– group theory 209 ff
approximation, numerical 328
Argand diagram 36
arguments, principal 36
Arnold's diffusion 400
Arnold's unfolding 412
Arnoldi method 320
arrow-head matrix 300
associated one-form 140
associative operations 3 , 191
asymptotic perturbation methods 390, 405, 412
atlas 137
atomic-scale computing 419
atomic sentences 455
atomic shells 209
attempt frequency 260
attractors 216
attributes 453, 466
attrition problem 256
automorphism 11
autoregressive/moving average (ARMA) 227, 236
averaging
– Monte Carlo methods 256
– numerical methods 305
– perturbation methods 397 ff, 412
– stochastic processes 518 f

Mathematical Tools for Physicists. Edited by George L. Trigg

© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ISBN: 3-527-40548-8

Avogadro's number 528
AXIOM software 218
axiomatizability, quantum logic 455, 459, 465 f

Bäcklund transforms 219
backsubstitution 290 ff, 376
backward differentiation formulas 353
balance principle 277
balancing *see:* scaling
balls 590
Banach space 596
band-structured Green's functions 177
banded matrices 298, 379
baryons 574 f, 583 f
basis, changes 6
Bashforth–Adams formulas 353
basic postulate, quantum logic 452
basics 592 ff
see also: definitions, fundamentals
basin of attraction 216
basis, changes 6
Bayesian approach 238
Bennet's scheme 421, 427
Bernoulli, brachistochrone problem 622, 652
Bernoulli's equation 49
Bessel functions 476 f, 499 ff, 504 ff
– Abel transforms 94
– Green's functions 176
beta ray processes 210
Betti number 149, 604, 613
Bianchi identity 147
biased sampling 256
bifurcations 221
– El Niño 241
– perturbation methods 389, 412
binomial theorem 18 f
Birkhoff group
– quantum logic 441
– topology 590, 600, 607
bisection methods 321
black box 161
BLAS, numerical software 378
block diagonal matrix 379
block tridiagonal matrix 364
Bohr model 441, 503
Boltzmann constant 182, 523
Boltzmann statistics
– Monte Carlo methods 253
– stochastic processes 536, 540
Bolzano–Weierstrauss property 596
Bonnet's theorem 140
Boolean algebra
– quantum computation 420
– quantum logic 440, 449, 467
– rings 16, 218
Borel set 445, 463
Born probability 463
Born series 176
Bose field 183
Bose particles 540
bosons 178, 575, 584
boundary conditions 478 ff, 498 ff
– analytical methods 47
– Laplace equation 61
– Monte Carlo methods 265 f
– ocean heat transport 243
– perturbation methods 386
boundary-element method (BEM) 171 f, 373
boundary layers 404, 407, 412
boundary operators 603
boundary point 593
boundary-value problems (BVPs)
– Green's functions 160, 164 f, 169 ff, 185
– ordinary differential equations 346, 354 ff
– partial differential equations 357 ff, 365 ff
bounds 637, 643
brachistochrone problem 73, 622, 652
Bragg position 276
Bravais lattice 567
Brillouin zone 276
broken symmetries 575, 582 f
Brouwer–Zadeh logics 461 ff, 467
Brownian motion 262, 528 ff
Brownian noise 549, 559
Brownian process 117, 121 ff
Broyden method 323
Broyden–Fletcher–Goldfarb–Shanno (BFGS) methods 328
building blocks, algebraic 22
bundles 606 f
Burgers vector 174

calculus of variations 620, 652
calculus theorem, fundamental 598
Campbell theorem 531 f
canard 407
canonical ensemble 263

canonical Hamilton mechanics 76
canonical mapping 14, 453
canonical matrices 10 ff
canonical null-root 27
canonical plane trigonometry 134
canonical transformations 400, 631 ff, 652
Cantor fractals 110, 220
carrier fluctuations 536 f
Cartan algebra 24, 141
Cartesian coordinates 62, 128 ff, 597
Cartesian product 13, 451, 590
catastrophic cancellation 288
Cauchy condition 478 f
Cauchy sequence 596
Cauchy theorem 35–43, 57 ff, 162
Cauchy–Riemann conditions 37
causal Green's functions 165
Cayley tables 196
cellular automation 223, 244
chain groups 603
chaotic systems 216, 227, 242 ff
Chapman–Kolmogorov relation 515, 545 f, 559
characteristic functions
– fields 4 ff
– matrices 315
– numerical methods 379
– perturbation methods 410
– stochastic processes 530 ff, 559
– variational methods 632
characterization, fractals 112, 115
charge conjugation 572, 584
Chebyshev polynomials 477 f, 492 f
Chern classes 149
Chern–Simons bundles 609
Cholesky factorization 293 f, 297 ff, 308 ff
chord Newton's method 322
Christoffel symbols 150
circles 130 ff
circular functions 131 f
circulation model 245
classical dynamics 651
classical mechanics 625
classical universal machines 420
classification, algebraic geometry 155
classification–regression trees (CART) 238
Clifford algebra 23
climate feedback control 232
closed groups 3 f
closed-loop control 232
closure point 593
cloud simulation 215 f
cluster algorithms 260 ff, 273
cluster–cluster aggregation 111
coarse grid residual 374
coarse-grained surface 118, 227
coclosed fields 142
coefficients, differential 50 ff
coherent superposition 447, 452
cohomologic group 149
cohomology 599, 605 f, 610 f
collinearity group 11
collocation method 367
colloidal flocculation 111
communication theory 239
commutative groups 3 f
see also: Abelian groups
commutativity, quantum logic 449
commutators
– Green's functions 178
– quantum mechanics 78
– symmetries 569 ff
compact groups 197
compact topologies 594
companion matrix 324
compatibility 443 f, 449, 467
complement 462
completeness
– eigenfunctions 480
– quantum logic 455, 459, 465 ff
complex boundaries 171 f
complex manifolds 147
complex numbers 2, 20, 200
complex projective space 588
complex systems 217
complex variables 35 ff, 80
complexity 418 f, 425, 429, 435
composite quadrature rules 341
composite solutions 412
computer simulations 217
concentration fluctuations, semiconductors 540
concentration, topology 610
condensed matter 273
condition number 301
conductivity fluctuations 535 f
confluent hypergeometric functions 476 f, 499 ff, 510
conformal transformation 44
congruent groups 11
conic sections 130 f
conjugate direction method 327
conjugate gradient methods 304, 309 f, 328
conjugate groups 193

conjugates 38, 481
– algebraic methods 11
– analytical methods 36
– Monte Carlo methods 272
– symmetries 571
conjunction 449, 453, 462, 467
conjunctive groups 11
connected groups 197
connected sum 589, 594
connections, differential 143
consequences, quantum logic 455
conservation laws 76, 565–586
conservative forces 79
conservative oscillator 396
conservative systems 626, 632, 652
– chaotic 228
consistence, quantum logic 449, 458
constant splines 332
constraints 623. 652
construction
– Green's functions 164
– group representations 201
contact transformations 631
continuity 594
continuous groups 191, 196
continuous media and fields, Lagrangian 77
continuous norm 336
continuous time series 236
continuous transformations 85, 569
contour integrals 38
contradictory 459
contravariance 7, 12, 138
control theory 231
controlling chaos 243
convection 175, 227
convergence 92 f
– boundary conditions 360
– numerical methods 305 f, 321, 327
convolution
– algebraic methods 23
– stochastic processes 522
– transformations 104 ff
Cooley–Tukey algorithm 96
Cooper pairs 575
coordinate perturbations 389, 402
coordinates method 128 ff, 137
correlation functions
– fractal geometry 112, 120
– Monte Carlo methods 265
– stochastic processes 517 ff
cosets 193
cosines law 131, 134
cotangent space 138
Coulomb forces 210
Coulomb potential 168, 179
Coulomb problem 647
countable additive probability 446
covariance 7, 13, 139
covers 594
Coxeter–Dynkin diagrams (CDD) 24
Coxeter–Killing transformations 26
Cramer's rule 8
Cray computers 217
creation operators 17, 572
critical numbers 612
critical points 325
critical slowing down 260, 265, 277
criticality 246
cross product 144
cryptography 418, 430, 435
crystal field theory 211
crystal lattice domain 94
crystal symmetries 567 f
crystallography 190
cubic splines 334
cumulants 529
curl 142, 610
curvatures 143, 606
curvilinear coordinates 130 ff
cut and paste techniques 588
cutoff lengths 111 ff, 120 ff
cyclic groups 191, 207
cyclic normal form 604
cyclic reduction 375 f
cylinders 134, 588
cylindrical coordinates 135, 503
cypher text 430

damping
– analytical methods 49 f
– nonlinear oscillator 388
– ocean heat transport 242
– stochastic processes 552
data assimilation models 216, 229
data encryption standard (DES) 430
DATAPLOT, numerical software 378
Davidon–Fletcher–Powell (DFP) method 328
De Morgan's law 457, 464
de Rham cohomology 149, 599, 606, 610 f
Debye functions 509
decay processes 552

decomposition theorem 149
decryption key 430
deep-ocean constant, El Niño 240
definite integrals 71
deformations 174, 588
degenerate kernels 68
degrees of freedom (DOF) 360, 433
degrees, algebraic 5, 9, 18 ff
delta function 161, 515
denormalized numbers 285
density operator 446, 467
density probability 236
DERIVE software 218
Desargue's theorem 151
Descartes–Euler characteristics 590
determinants 8, 199
deterministic model 221, 245
diagonal matrix 297, 379
diagonal sum 203
see also: trace
diffeomorphism 608
differential equations 33 ff
see also: ordinary, partial etc.
differential geometry 136
differential operators 161 ff
differential topology 606 ff, 613
diffraction
– Fourier transforms 94, 100
– Green's functions 177
– symmetries 568
diffusion
– Monte Carlo methods 275, 641
– ordinary differential equations 57
– stochastic processes 517 f, 537 ff, 541 f
diffusion-limited aggregation (DLA) 111, 251, 258 f
digital computing 217
dilation 110, 198
dimensions 123, 358, 362
dipole moment 201
Dirac matrices 23
Dirac–Feynman approach 78
direct models 216
direct product 195
direct search method 325
direct solutions 289
direct sum 13
directrix 130
Dirichlet condition 478 f
– ordinary differential equations 57
– Green's functions 163 f, 168 ff, 185
– partial differential equations 359
discrete Fourier transforms 85 f
discrete groups 196
discrete norm 336
discrete symmetries (T,C,P) 572
discrete time series 236
discrete topology 592
discrete transformations 106, 569
discretization 226, 245
– chaotic systems 228, 245
– Poisson equation 301
discriminants 130
disjunction 449, 467
dislocations 174
disordered series 391
dispersion relations 43 f
displacements 69, 118
dissipation
– chaotic systems 216, 228
– quantum computation 419
– solitons 244
distributed-memory passing machines 375
distribution law 3, 442
divergence 142, 598
divided differences 330
division, synthetic 324
Dolbeault operators 148
dominant matrix 379
dopant impurities 419
double precision floating points 285
doublet states 212
downwelling, El Niño 241
drift 517, 541
dual modes 94
dual space 12 f
dynamic models 216
dynamic quantum logic 447, 452, 467 f
dynamical law 452, 468
dynamical systems 245
– perturbation methods 392
– stochastic processes 515 f
dynamics 221
– classical 651
– Monte Carlo methods 261
Dynkin–Coxeter diagrams 24
Dyson's equation 179, 184 f

E numbers 23
eccentricity 130
Eddington theory 23
edge dislocations 174
efficiency 361

eigenvalues
– algebraic methods 8
– group theory 201
– matrices 315
– perturbation methods 386
– quantum logic 476 ff
eigenvectors 478
– algebraic methods 8
– matrices 315
– numerical methods 379
– perturbation methods 386
Einstein functions 509
Einstein relation 517, 527 ff, 537 ff
Einstein's general relativity theory 12, 150
Einstein's summation convention 2, 139
Einstein–Podolski–Rosen paradox 454
EISPACK, numerical software 378
El Niño (ENSO) 216, 245
elasticity 172
electrodynamics 648, 652
electromagnetism
– boundary–element method 172
– differential geometry 144
– variational methods 633
electron band structure 177
electrons 541
elementary group theory 190
elementary quantum notation 422
ellipses/ellipsoids 130–136
elliptic differential equations 58, 65, 358
embedding space 123
empirical orthogonal function (EOF) 234
energy levels 209
energy–momentum tensor 572
enrichment technique 257
environment, quantum computation 433, 435
equilibration *see:* scaling
equilibrium fluctuations 262
equilibrium states 516
equipartition 524
equivalence relation 10, 14
equivalent group representations 203
equivalent spaces 594
ERASE gates 420
ergodic systems 239, 263, 277
Erlangen program 152
errors
– Monte Carlo methods 265 ff
– numerical methods 301
– partial differential equations 374
– polynomial interpolation 331
– quadrature 343
– quantum computation 418, 432, 435
– roundoff 287, 302
essential boundary conditions 359
Euclidian algebra 24
Euclidian elements 590
Euclidian groups 199
Euclidian metrics 140
Euclidian norm 379
Euclidian space 135 ff, 608
Euler bundles 609
Euler classes 149
Euler equation
– analytical methods 36
– functional analysis 72
– ordinary differential equations 54, 346
Euler series 83
Euler–Lagrange equation 75, 621, 652
Euler–Poincaré characteristics 604, 612
exact differential equations 48
excited states 637
expanding interval 395
experimental dynamics 222
experimental theoretical physics (ETP) 217
explicit formula 347 f
exponents, floating points 285
exterior algebra 17
see also: Grassman algebra
exterior derivatives 141
extrema 71

faces 602
factorial function 491
factorial n 19
factorization
– numerical methods 292–315
– quantum computation 429 ff
– special functions 476 ff, 484 ff
FANOUT gates 420
Faraday's law 145
fast-Fourier transforms (FFT) 96 ff, 106, 237
fast-Hartley algorithm 97
fat fractals 219
fault-tolerant quantum computing 433 ff
feedback control 232
Fenichel perturbation methods 407
Fermat's principle 620, 629, 652
Fermi–Dirac statistics 270, 536, 540
fermions
– algebraic methods 18

– Green's functions 178
– group theory 210
– Monte Carlo methods 270
– symmetries 580
Feynman diagrams 163 f, 180 ff, 184 f
Feynman path integrals
– quantum mechanics 36
– variational methods 621, 649 f, 652
Feynman quantum mechanics 78
FFTPACK, numerical software 378
FHA cycle, (Fourier–Hankel– Abel) transforms 102
fiber bundles 142 f
Fibonacci series 253, 325 f
field operators 178
fields, algebraic 3 ff
filters 230 f, 521 f
finite-difference method
– ordinary differential equations 54 f
– partial differential equations 65 f, 359 ff, 369 ff
finite-element method 228, 245
– partial differential equations 359 ff, 365 ff
– variational 641
finite groups 191, 195 f
finite Lie algebras 25
finite-model property 460
finite regions 168
finite-termination property 311
finite-size fractals 219
finite-size problems 265, 277
first-order ordinary differential equations 35, 47 f, 56 f
fixed-point iteration 320
flavors 272
floating points 272, 284 ff
flocculation 111
flop (floating point operation) 292, 379
Flory–Huggins theory 274
flow 245
fluctuations 262, 516 ff, 535 ff
fluid mechanics
– ^{4}He momentum distribution 269
– boundary–element method 172
– fractal geometry 118
– perturbation methods 409
flux quantum 175
FNLIB, numerical software 378
focus 130
Fokker–Planck equation 516 f, 534–559
foliations 403
formal dynamics 222
FORTRAN software 377
forward elimination 290, 293
forward Euler method 346
Fourier series 481
– perturbation methods 396
– topology 590
Fourier transform 83–108
– algebraic methods 23
– analytical methods 35, 57 ff, 62 ff
– Green's functions 166 f, 179
– integral equations 68
– numerical modeling 226
– quantum computation 427
– stochastic processes 515 f, 519, 543
– wavelets 230
fractal geometry 109–126
fractal orbit structures 216
fractals 219 ff, 233
fractional power series 410
Fredholm integral equation 67 f, 171 f, 185
Fredkin–Toffoli scheme 421
freedom degrees 360, 433
Frenet–Serret formulas 140
frequencies
– partial differential equations 374
– perturbation methods 397 ff
– stochastic processes 525 ff
– transformations 94, 106
Frobenius algebra 4, 23 f
Frobenius–Fuchs power series 476 f, 490 f, 496, 501 f
Frobenius–Perron theorem 10
function domain 94
functionals
– analytical methods 35, 70 ff, 80
– transformations 84 f
– variational methods 621, 652
functions, analytic 37
fundamental theorem of calculus 598
fundamental theorem of algebra 19 ff
fuzzy complement 456 f, 462

Galerkin methods 366
Galois groups 4, 20
Gamma function 491, 510
GAMS, numerical software 378
gates 420 f, 425
gauges
– differential geometry 145 f
– Monte Carlo methods 271 f, 277

– perturbation methods 389, 412
– symmetries 578 ff, 583
– topology 609
Gauss–Bonnet theorem 149
Gauss–Codazzi equations 141
Gauss differential equations 490 ff
see also: hypergeometric functions
Gauss distribution 256, 529
Gauss elimination 290, 299
Gauss ensemble 265
Gauss law 20 ff, 144
Gauss–Legendre quadrature rules 338 ff
Gauss model 536, 546
Gauss–Seidel iteration 305 ff, 323
Gear formulas 353
Gegenbauer polynomials 477 f, 492 f
general boundary conditions 359
general circulation models (GCMs) 215, 229, 245
general quantum logic 448
general relativity theory (GRT) 12, 150
generalized Hamilton mechanics 74
generalized series 389, 413
generating function 483, 510, 631
generation–recombination processes 515, 533 ff, 537
genus 604
geometric scaling 111
geometric singular perturbation methods 407
geometrical methods 127–158
Gibbs ensemble 265
Ginzburg–Landau equation 182
Glashow–Weinberg–Salam model 580
glide reflection plane symmetries 567
gluons 584
golden section-search method 325
Goldstone bosons 575, 584
gradients 142 ff, 325 ff
Gram–Schmid orthogonalization 314, 337
grand canonical ensemble 263
Grassmann algebra 17, 24
gravity stabilized invasion percolation 119
Green' theorem 598
Green's functions 159–188
– analytical methods 35, 53 ff, 57 ff
– potential theory 64
– stochastic processes 543
Green's operator 175
grid representation
– fractal geometry 113
– Laplace equation 65
– partial differential equations 362, 374
group theory 189–212
– algebraic 2 ff, 23
– symmetries 566, 584
– topology 600
Grover's data-base-searching algorithm 434
growth factor 302
growth fractals 219
gyration radius 110 ff, 124

Haar measure 23
Hadamard transform 103
hadrons 272
Hamilton formalism
– algebraic methods 9
– mechanics 36, 74 f
– quaternions 15
– variational methods 626 ff, 630 ff, 653
Hamiltonian operator
– atomic structure 209
– Green's functions 175 ff
– group theory 200
– Monte Carlo methods 258, 268
– quantum logic 448
– symmetries 580
Hamilton–Jacobi theory 74, 627, 632 f
Hankel functions 175, 504
Hankel transforms 87 ff, 100
hardware, digital computing 217
harmonic functions 38
harmonic oscillators 486
– analytical methods 36
– perturbation methods 396
– stochastic processes 526
Hartley transforms 87 ff, 95 ff
Hartree–Fock configuration 640. 653
Hausdorff space
– chaotic systems 228
– differential geometry 137
– fractals 219
– topology 155, 594
heat-bath method 260, 277
heat dissipation 419
heat transport
– chaotic 242
– Green's functions 166 f, 170 f
– ordinary differential equations 57 ff, 67
– topology 610
– transforms 94

Heaviside step-function 106, 165, 182
Heisenberg model
– Green's functions 178, 182 f
– quantum logic 447
– quantum mechanics 78
helium 638
Helmholtz equation 476 f, 494 ff, 500 ff
– analytical methods 62
– boundary–element method 172
– Green's functions 165, 175
– numerical methods 359
– topology 597
Hénon maps 222
Hermitian conjugate 272
Hermitian interpolation 331
Hermitian kernels 68
Hermitian matrices 297, 300 f, 305 f, 379
Hermitian metrics 148
Hermitian operator 480, 518, 569 ff
Hermitian polynomials 334 ff, 476 f, 499 ff
Hermitian symmetry 105
Hermitian tensor 11
Hessenberg matrices 316 ff, 379
Hessian operator 325, 328, 611
Higgs mechanism 581
Hilbert space
– algebraic methods 22
– quantum computation 423 f, 432 f
– quantum logic 441, 444, 463 f
– quantum mechanics 78
– symmetries 569 ff
– topology 596
Hilbert transform 43, 85 ff, 99 ff
Hilbert's fifth problem 197
Hilbert–Bernays style 459
Hilbertian quantum logic 458
histogram method 265
Hodge decomposition theorem 149
Hodge-star map 141, 610
holes 115, 541
holonomic constraints 624
homeomorphic spaces 594
homogeneous boundary conditions 168
homogeneous coordinates 152
homogeneous differential equations 50
homogeneous space 14
homogeneous systems 172
homology 601
homomorphism 7, 22, 195
homotopy 599
Hopfield type neural networks 235
Horner's rule 324
Householder transformations 313
Hurst exponent 117, 121
Huygen's principle 166, 173
hydrogen polarizability 644
Hylleraas–Undheim–MacDonald (HUM) theorem 637, 653
hyperbolas 130 f, 136
hyperbolic differential equations 58, 67
hyperbolic functions 132
hyperbolic partial differential equations 358, 371
hyperboloids 134
hypergeometric functions 476 f, 490 ff, 499 ff, 510
hyperplanes 135, 252
hypersurfaces 135, 599
hypervolumes 598

identity element 3, 191
identity transformations 566
IEEE standard floating points 284 f
Ikeda maps 222
imaginary number i 36
impedance 44, 526
imperfect bifurcations 389
implicit ordinary differential equations 347 f
importance sampling 245, 253, 258 f, 277
IMSL, numerical software 378
incomplete factorization 308
independent vectors 6
indetermined truth value 462
indicial equation 52
induced rounding 265
inertia law 11
Infeld–Hull factorization 476 ff, 484 ff, 489 ff
infinite groups 191
infinite integrals 254, 344
infinitesimal groups 17
information theory 233 f, 239
initial value problems (IVPs)
– analytical methods 47
– Green's functions 160, 169 ff
– ordinary differential equations 346 ff
– partial differential equations 358 ff, 369 ff
– perturbation methods 404, 413
inner product 336, 593
inner solutions 404, 413
integers 2, 15

integrals
– analytical methods 38, 67 ff, 80
– Monte Carlo methods 254
– quadrature 340
– stochastic processes 548
– variational methods 621, 626, 649 f, 652
interfaces 116 f, 266
interference pattern 418
intermediate truth value 462
intermediate variable 406
internal symmetries 573
interpartition smoothing criteria 227
interpolation 329 ff, 334 ff, 338
invariance
– differential geometry 144
– fractals 220
– group theory 194, 199, 204
– projective geometry 153
invasion percolation model 119
inverse iteration 319
inverse models 216, 245
inverse scattering theory (IST) 244 f
inverse transformation 45, 106, 191
inversely restricted sampling 256
involution 10, 444, 449, 468 ff
irreducibility 202 ff, 206
irregular grids 362
irregular surfaces 116
Ising model 184, 266, 274
island method 123
isolated point 593
isomorphic groups 194
isospins 211, 573
isothermal-isobaric ensemble 2634
isotropic systems 172
iteration
– mapping 245
– numerical methods 290, 303 ff
– ordinary differential equations 55
Ito definitions 517
Ito–Stratonovic controversy 548, 555

Jacobi identity 17
Jacobi iteration 305, 319
Jacobi polynomials 477 f, 492 f
Jacobi–Hamilton theory 74, 627, 632 f
j–j coupling 211
Johnson noise 523 ff
joins 443, 468
Jordan blocks 9 ff, 386
Jordan–Brouwer separation 599
jump rate 541
Jung notation 553

Kac–Moody algebras 10, 24
Kähler metrics 148
Kalman filters 230 f
Kaluza–Klein theories 151
Kelvin's solution 173 ff
Kepler's law 135, 386, 400
kernels
– algebraic methods 7 ff, 10 f, 22, 28
– Green's functions 161
– integral equations 67
Killing groups 22 ff
Killing–Coxeter transformation 26
Killing–Weyl groups 26
kinetic energy 136, 200
Kirchhoff loop 49, 95, 173
Klein bottle 588, 608
Klein–Gordon equation 167 f, 171, 184 f
Kolmogoroff–Arnold–Moser (KAM) theorem 386, 399 f
Kolmogoroff–Sinai entropy 228, 239
Königsberg bridge problem 590
Korringa–Kohn–Rostoker eigenvalues 177
Kortweg–de Vries equation 244
Kramer's model 553, 556
Kripkean semantics 455, 468
Kronecker delta
– algebraic methods 24
– boundary–element method 173
– group theory 206, 210
– Hamilton mechanics 77
– symmetries 570
Kronig–Kramers relation 43
Krylov–Bogoliubov–Mitropolski (KEM) method 397
Kubo theory 164, 182
Kummer functions 499

La Niña 245
ladder approximation 179
Lagrange theorem
– analytical methods 53
– functional analysis 74 ff
– group theory 193
– numerical methods 330 f
Lagrange undetermined multiplier 623, 653

Lagrangian operator
– Green's functions 163, 183
– symmetries 566 ff, 571 ff, 578 ff
– variational methods 630, 653
lags p,q 253
Laguerre polynomials 476 f, 486 ff, 499 ff
lambda matrices 18
Lanczos method 320
Landau–Lifshitz description 566 f
Langevin processes 531, 544 ff, 559
LAPACK, numerical software 378
Laplace equation 476 f, 497 f
– analytical methods 38, 57 ff, 61 ff, 65 f
– conformal transformations 45
– Green's functions 162, 168 ff
– Mellin transform 98
– numerical methods 358
– topology 591, 597, 610
– variational methods 625
Laplace transforms 87 f, 91 ff
– algebraic methods 23
– integral equations 68
Laplace–Beltrami operator 610
Laplacian operator
– analytical methods 35, 62 f
– differential geometry 142
– Green's functions 162
– Helmholtz equation 494
– topology 610
lattices
– Bravais 567
– gauge theory 271 f, 276 f
– group theory 211
– Monte Carlo methods 254 ff
– orthocomplemented 441, 469
– quantum logic 446
latus rectum 132
Laurent polynomials 19, 35, 39
LDL factorization 297 f
least action principle 629, 653
least square fitting 121, 336
Legendre polynomials 476 f, 486 ff, 492 f, 495 f
Legendre transformations 631
Leibnitz formula 483, 502
leptons 580 f, 583 f
letter totality (alphabet) 227
Liapunov exponent 227 f
Lie groups 197, 477 ff, 484 ff
– algebraic methods 16, 22 ff
– differential geometry 143, 146
– digital computing 219
– linear representation 200
– perturbation methods 400, 413
– symmetries 566 ff, 569 ff, 584
– topology 590 f
limit point 593
Lindenbaum property 460, 468
Lindstedt's method 394 ff, 413
linear differential equations 35 ff
linear groups 198
linear multiplicate algorithm 252
linear multistep formulas 352 f
linear operators 478
linear spaces 5 ff
linear splines 333 ff, 366
linear stationary method (1st degree) 304
linear systems 312
linear transformations 106
– algebraic methods 11
– perturbation methods 386
linear variation parameters 636
linear vector space 191
lines method 372
LINPACK, numerical software 378
Liouville operator 478 f
Liouville theorem 37
Lipschitz condition 93
logic gates 420 f, 425 f
logical reversibility 419
logician's approach 454
Lorentz boosts 147, 570 ff
Lorentz equation 227
Lorentz groups 199, 568 ff
Lorentzian metrics 142, 147
LU factorization 292, 296 f

machine epsilon (macheps) 286
Mackey's formulation 443
MacNeille completion 463, 368
MACSYMA software 218
magnetization, spontaneous 265
magnification 45
manifolds
– differential geometry 137
– perturbation methods 403
– topology 607 f
mantissa 285
manual space 454
many-body processes 163, 178
many-valued possible-word semantics 456
MAPLE software 218, 229

mapping 245
– algebraic methods 28
– canonical 14
– functional analysis 71
– projective geometry 152
– quantum logic 453
– tensors 139 f
Marchenko equation 244
Markov processes 221, 245
– Monte Carlo methods 261
– stochastic 515 f, 544 ff, 559
mass–length scaling 112
matching 405, 413
material science 273
MATHEMATICA software 218
mathematical modeling 213–248
Mathieu functions 509
MATLAB, numerical software 218, 378
matrices
– perturbation methods 386, 410
– properties 2
– transposition 222
matrix–vector products 316
Maupertius least action principle 629, 653
maxima 71
maximization 324
Maxwell equations
– differential geometry 144
– Green's functions 176
– topology 598, 610
mechanical wave equation 57, 476 ff
mechanics
– classical 625
– Hamilton principle 74
– Monte Carlo methods 273
mechanistic models 215, 245
meet 443, 468
Meissner effect 175
Mellin transforms 87 ff, 98 ff
Melnikov function 403
memory complexity 361
mesons 273, 574
metallurgy 273
metrics 140 ff, 592
Metropolis importance sampling 258 f
Metropolis–Hastins procedure 225
Michelson interferometer 95
midpoint quadrature rule 338, 350
minima 71
minimal polynomials 9
minimization 324
minimum curvature 335
minimum functional theorem 358
Minkowskian metric 140, 145
minus-sign problem 271
mixed boundary conditions 243, 359
Möbius strip 588
modal logics 440
model quantum computer 426
model-theoretic consequence 455
modular law 442
moduli 28, 156
Moivre's theorem 37
molecular dynamics 217, 278
momentum
– conjugate 571, 574
– fluid ^{4}He 269
– variational methods 632
monomials 18, 330 f
Monte Carlo methods 217, 224, 249–280, 345
Moody–Kac algebras 10, 24
morphisms 28
Morse theory 149, 611 ff
motion equations 476 ff
– analytical methods 50 ff
– differential geometry 147
– group theory 190
– variational methods 630 ff, 653
see also: Hamilton-, Newton- etc.
multicanonical ensemble 265
multiconfigurations 641
multidimensional transforms 99 ff
multifractals 219
multifrequencies systems 399
multigrid method 373 f
multinomial theorem 19
multiple instruction, multiple data machine (MIMD) 375
multiple-scale method 400, 404, 413
multiple scattering effect 177
multiple shooting 355
multiplication 3, 324
multiplier 623, 653
multivalued logics 440
multivariate time series 235
mutual information 239

n factorial 19
NAG, numerical software 378
Nambu–Goldstone bosons 576
natural boundary conditions 359
natural curvature 335

natural numbers 2
negative frequency 106
neighborhood 593
Nekhoroshev theorem 386, 399 f
nested dissection ordering 301
nested multiplication 324
netlib, numerical software 378
Neumann condition 478 f
– Green's functions 163 ff, 185
– ordinary differential equations 57
– partial differential equations 359
– quantum logic 441
Neumann functions 504
Neumann integrals 70
neural networks 235, 245
neurons 235
neutral element 3
neutrinos 581
neutrons 577
Newton equations 625
Newton gravitation law 386, 399
Newton iteration matrix 350, 356
Newton mechanics 74
Newton method 320 ff, 328 ff
Newton physics 16
Newton–Cotes quadrature rules 338 ff
nilpotency index 9 ff
nodes 235, 338
Noether's theorem 570
noise 216, 236
– quantum computation 433
– shot 528 ff, 533 ff, 545 f
– stochastic processes 515–559
– thermal 523 ff
nonhomogeneous differential equations 51 f
nonlinear algebraic equations 320
nonlinear dynamics 221
nonlinear oscillator 388, 392, 396
nonstiffness 346 f
normal form 312, 389, 402
normalized numbers 285
normed space 593
norms 336
North Atlantic Ocean circulation 216
Norton's theorem 524
NOT operations 420
not-a-number (NaNs) 285
notation
– Einstein 13
– general 2
– group theory 192
– quantum computation 422
– stochastic processes 553
NP complete problem 434
NTV ensemble 263
nuclear structure 210
nucleation sites 113
nucleon states 211
nucleus *see:* kernels
nullity 8
null-root, canonical 27
numerical algorithms group (NAG) 218
numerical methods 281–384
– integration 338
– modeling 226
– ordinary differential equations 54
– partial differential equations 64 f
– software 377
Nyquist's theorem 523 ff, 527 ff

observables 445
occupation
– atomic shells 211
– lattice sites 254
ocean heat transport 242
odd–even reduction 376
Ohmic current 182
one-bit gate 423
one-forms 137, 140 f
one-step methods 356
Onsager's regression 516
open loop control 232
operations 3 f
operators 478 ff
– Green's functions 161 ff, 178 ff, 185
– partial differential equations 358
– quantum logic 441, 444
– ranks 8, 11
– symmetries 566 ff
– topology 610
optical character recognition (OCR) 235
optical field phase experiment 425
optimization methods 324
OR operations 420
ordered series 391
orders
– group theory 191
– matrices 301
– ordinary differential equations 350
– partial differential equations 358, 362
– quantum logic 462

ordinary differential equations (ODEs) 476 ff
– analytical methods 35 ff, 46 ff, 80
– numerical methods 346
– perturbation methods 386
ordinates 129, 133
orthoalgebras 441, 448 f, 469
orthoarguesian law 458
orthocomplementation 441, 449, 469
orthodox quantum logic 457, 469
orthogonal eigenfunctions 480
orthogonal groups 11, 198
orthogonal matrix 379
orthogonal polynomials 482 ff
orthogonal transformations 313
orthogonalization 337
orthomodular law 442, 469
orthopair models 462
oscillation theorem 646, 653
oscillators
– harmonic 486, 526
– perturbation methods 388, 413
– quantum 479
– self-sustained 557
– stochastic processes 516
oscillatory interpolation 331
outer solutions 405, 413
overall efficiency 361
overdetermined linear systems 312 ff
overlap domain matching 406

p forms 141
Padé approximation 224, 245
paleoclimatology 245
Pappus' theorem 151
parabolas 130, 136
parabolic cylinder functions 509
parabolic differential equations 58, 67, 358, 370
paraboloids 134
paraconsistent quantum logic 460, 469
parallel computation 375 ff
parallel processors 218
parallel shooting 355
parallelism 427
parametrization 135, 216, 245
– perturbation methods 387, 413
parity transformation 572
partial differential equations (PDEs) 476 ff
– analytical methods 35, 57 ff, 80
– cellular automation 223
– Green's functions 160
– numerical methods 304 f, 357 ff
– parabolic 358, 370
– perturbation methods 386
partial quantum logic 464
particle path 257
partition
– algebraic methods 10
– Monte Carlo methods 271
– statistical interference 238
– topology 596
Pascal's theorem 151
path integrals
– Feynman 79
– Monte Carlo methods (PIMC) 268
– variational methods 621, 649 f, 652
paths topology 595, 600
Paul trap 425
Pauli matrices 5, 16
Pauli principle 210
Peaceman–Rachford method 309
penetration depth 175
perceptrons 235
percolation model 119, 124
periodic functions 85
periodic standard form 397
permutation groups 192, 208
permutation matrix 296, 379
permutations 141
Perron–Frobenius theorem 10
perturbation theory 385–418
– Green's functions 163, 175
– numerical methods 301
– variational methods 641 ff, 654
phase factor 447
phase shift 425
phase space 221
– differential geometry 150
– Monte Carlo methods 262
– quantum logic 440
phase transitions 265
physical reversibility 419
physical systems, quantum logic 451
piecewise polynomials 332 ff, 360, 365
pivoting 291 f
Planck constant 200, 568
plane analytic geometry 129 f
plane trigonometry 131 ff
plane waves 176
PLU factorization 296 f
Pochhammer symbol 491

Poincaré groups
– algebraic methods 29
– perturbation methods 389, 394, 413
– symmetries 570
– topology 590, 607
Poincaré lemma 611
Poincaré section 599
Poincaré–Hopf index 591 f, 606, 609 f
point groups 206, 566
point response functions 161
point-set topology 591 f, 613
point sources 173
point transformations 631
Poisson brackets 77
Poisson equation
– Green's functions 169 f
– numerical methods 359
– ordinary differential equations 57
– stochastic processes 528 ff
– topology 610
Poisson ratio 173
polar coordinates 132, 397
polarizability, hydrogen 644
pole–zero approach 237
polygonal grids 362
polymer science 274
polynomials
– algebraic methods 5, 18 f
– interpolations 329
– numerical methods 321, 379
polyspectra 237
Pontryagin bundles 609
Pontryagin classes 149
Pontryagin–Andronov–Vitt (PAV) model 553
population fluctuations 538 f
PORT, numerical software 378
posets, orthomodular 444, 449, 468
position vectors 138
positive definite matrices 297, 380
positive definite operators 358
possible-word semantics 455
potential energy 136
potential theory 64
Powell's method 327
power law distribution 115
power series 476 ff
– numerical methods 316
– perturbation methods 389, 410
power spectra 98, 245
practical fractal geometry 123
preconditioning 374
predicate calculi 440
prefractals 123
preserved foliations 403
primitive unit cell 567
principal arguments 36
principal attribute 454
principal components analysis (PCA) 233
probabilities 236
– Monte Carlo methods 253 f, 277
– quantum logic 450, 469
– stochastic processes 534 ff
products 191, 593
projection operators 441 444, 469
projection–slice problem 101
projective geometry 151
projective transformations 152
projective varieties 155
proof-theoretic consequence 455, 459
propagation 235
propagators 163, 174 f, 177
protons 210, 577
pseudorandom numbers 224, 246
– Monte Carlo methods 250 f
– quantum computation 430
pseudostates 644, 653
public-key cryptography 418, 435, 430
Pythagorean theorem 131

QR factorization 313, 317 f
quadratic piecewise polynomials 333 ff
quadratic splines 298, 333 ff
quadrature 338
quadrics 133
quadrilateral grids 362
quantization 183
quantum bits (qubits) 425, 435
quantum chromodynamics 226
quantum code 426
quantum computation 417–438
quantum field theory 183, 528, 576, 585
quantum Fourier transform 427, 436
quantum logic 439–474
quantum mechanics
– group theory 200
– Hamilton principle 74, 78
– variational methods 634
quantum Monte Carlo methods 268
quantum numbers 209, 573
quantum oscillator 479
quantum-chromodynamics (QCD) 271
quantum-electrodynamics (QED) 640, 654

quark model 272, 577 ff, 585
quasibiennal oscillation (QBO) 240
quasicrystals 568
quasi-equilibrium 670 states 517
quasi-Newtonian method 323, 328
quaternions 3 ff, 15 ff, 23 ff
quenched Monte Carlo approximation 272

radial Laplace equation 497
radiation 173, 257
radii 130 ff
radioactive decay 48
radiofrequency ion map 425
Radon transform 87 ff
random behavior 216
random numbers 250 ff, 278
random phase approximation 179
random walk problems 255, 514, 552 f
ranks
– group theory 203
– numerical methods 379
– operators 8, 11
– tensors 139
rational functions 5 ff, 223, 246
rational numbers 2
Rayleigh–Bénard convection 227
Rayleigh–Ritz variational method 635 f, 654
Rayleigh–Schrödinger variational principle 635 f, 654
reaction rate theory 553
real parts, complex variables 36
real projective plane/space 588
reciprocal lattice 94, 568
reciprocal transforms 87
reciprocally adequate relations 455
recombination processes 515, 533, 537
rectangle quadrature rule 338
rectangular coordinates 129
rectangular grids 362
recurrence formula 482
recursion 52, 252
reduced states 448
reducibility 203
redundancy 10
reflection 567
refraction index 44
regression theorem 517 f
regula falsi method 321
regular involution 470
regular perturbation method 393, 398, 413
regularity condition 461
relative roundoff error 287, 302
relativistic generalization 633
relativistic quantum-field theory 582
relaxation
– Monte Carlo methods 267
– numerical 267, 306
– perturbation methods 407, 413
replicas 110
representations
– complex numbers 36
– Green's functions 168 ff
– Lie group 200
rescaled variables 388, 413
residuals 302, 366, 374
residues 41
resistance 44
resistance modulation fluctuations 535 f, 539 ff
resolvent kernels 161
resonance 399, 413, 556 f
response functions 161, 182
rest points 393
retarded Green's functions 165, 174
reversibility 419 ff, 436
Rham cohomology 149, 599, 606, 610 f
Rice model 530 ff
Richadson's method 305
Riemann integrals 548
Riemann metrics 140, 148
Riemann transformations 44, 156
Riemann–Christoffel tensor 13
rigid bodies 566 f
rings 15
Rodriguez formula 482 ff, 493 f, 497 f, 511
Romberg integration 343
root multiplicity 27
rotation
– conformal transformations 45
– group theory 190, 202, 208
– symmetries 566
rough surfaces 116 ff
rounding errors 284 ff, 301 ff
Rouse diffusion 275
RSA public-key cryptography 430, 436
rubber sheet geometry 588
Ruelle–Takens–Newhouse route, El Niño 241
Runge–Kutta method 56, 350 ff, 356
Russel–Saunders coupling 210

salinity, ocean heat transport 242
sampling 245, 251 ff, 277
Sanchez–Palencia theorem 398
SAS, numerical software 378
Sasaki hook 457
scalar field 140
scalar product 5, 199, 206
scalars 5
scaling
– fractal geometry 111
– fractals 219
– Monte Carlo methods 265, 276 f
– numerical methods 293
– perturbation methods 409
scattering 176 f, 245
Schlesinger nomenclature 232
Schmidt–Hilbert integrals 68
Schrödinger equation 57, 476 ff, 486 ff, 494 ff
– Green's functions 163 f, 167 ff, 174 ff
– group theory 200 ff
– Monte Carlo methods 225, 270
– quantum mechanics 78
– variational methods 634, 654
Schrödinger model 441, 447, 610
Schur lemma 205
sea surface temperature 240
secant method 321
second-order curves 130
second-order differential equations 35, 49 f, 56 f
security, quantum computation 430
seed sites 113, 252
self-adjoint operators 358, 441, 466, 478 f
self-affine fractals 115 ff, 124
self-avoiding walk (SAW) 255 ff
self-organization 233, 246
self-similar fractals 110, 124
self-sustained oscillators 557
semantics, quantum logic 455 ff, 464 ff
semiconductors 540
semigrand canonical ensemble 263
semitransparent effect 463
sensitivity-to-initial-conditions (SIC) 216
separable differential equations 47
separable differential geometry 137
separation conditions 594
separation of variables method 476 ff
sequence periods 427
Shah function 92
Shannon's information theory 430
shared memory machines 375
Sharkovsky theorem 227
shielding problem 257
shift register generators 253
shifting 265, 318
Shockley relation 536
shooting methods 354
Shor algorithm 418 ff, 426, 431
shortest time problem
see: brachistochrone problem
shot noise 528 ff, 533 ff, 545 f, 559
signal functions 161
signal processing 237
significands 285
similarity transformation 316
simple groups 22
simple sampling 251 ff, 278
simple shooting 355
simplex 602
Simpson's quadrature rule 338
simulation models 215 f
simultaneously testing 443
sines law 131, 134
single-instruction, multiple data machine (SIMD) 375
single-frequency systems
single-qubit gate 434
singular p-simplex 602
singular spectral analysis 241
singular value decomposition (SVD) 230, 234
singularities
– analytical methods 37
– Green's functions 174
– quadrature 344
skew symmetries 10
SLATEC, numerical software 378
Slater determinants 18, 225
Slater–Koster model 176
slit island method 123
slow–fast perturbation methods 407
Smale's horseshoe map 228
Smith's canonical matrix 10
smoothing, numerical 226
software 218, 377
solid analytic geometry 133 f
solid states group theory 211
solitons 243, 246
sound waves 172
soundness 455, 459, 465, 470
source terms 357
southern oscillation index (SOI) 216, 240

space 592
space group 568
space–time curvature 13
space–time grid 66
sparse matrix 380
spatial domain 94
spatial frequency 106
special functions 475–512
spectra 410
spectral analysis 237
spectral domain 94
spectral measurements 519
spectral models 226
spectral radius 305, 380
spectral representations 43, 168
spectral theorem 445, 470
spheres 590
spherical Bessel functions 476 ff, 506 ff
spherical excess 134
spherical harmonics 229
spherical Helmholtz equation 494
spinless particles 178
spins
– group theory 209 ff
– Monte Carlo methods 276
– quantum computation 422
– symmetries 573, 578
splines
– interpolation 298
– numerical methods 227, 332 ff
– partial differential equations 366
splitting matrix 304
spontaneous magnetization 265
spontaneous symmetry breaking 577, 585
square matrices rings 15
stability, boundary conditions 360
stabilizers 14
stable manifolds 403
standard engineering definitions 519
standard p-simplex 602
standard quantum logic 444, 455, 470
standard symmetry model 580, 585
standardization (standard convention) 482 f, 493 f, 500 ff
STARPAC, numerical software 378
states, quantum logic 446, 452, 470
stationary stochastic processes 516 ff
stationary values 71
statistical ensembles 263 f
statistical errors 267
statistical interference 238
statistical mechanics 273
statistical Padé approximation (SPA) 224
statistical thermodynamics 258
statistically self-similar fractals 111
steepest descent method 327
Steifel–Whitney classes 149, 609
step sizes 362
Stieltjes integral 548
stiffness 346 f
stochastic dynamical system 246
stochastic modeling 216, 221
stochastic processes 513–564
stochastic supports 451
stochastic trajectories 262
Stokes theorem 598
Stokes damping 49
stopping criterion 320
straightforward expansion 394
Stratonovic–Ito controversy 548
string theory 16, 151
structures 13 ff, 22
Sturmian basis set 647, 654
Sturm–Liouville operator 163 ff
Sturm–Liouville problem 646, 654
Sturm–Liouville theory 477–511
SU(2)/SU(3) algebras 150, 575 f
subgroups 192 ff
successive overrelaxation (SOR) 306
sum, direct 13
sum rule 288
supercomputers 217
superconductors 175
superposition
– Green's functions 161
– ordinary differential equations 356
– quantum computation 418, 428
– quantum logic 447, 452, 470
superselection rule 448
support 451, 470
surface 116, 266
surface physics 276
susceptibility 425
Susuki–Trotter formula 269
Sylvester's inertia law 11
symbolisms *see:* notation
symmetric pivoting 298
symmetric positive definite (SPD) matrix 327
symmetric successive overrelaxation (SSOR) 306
symmetries 565–586

– algebraic methods 3, 10
– group theory 190 ff, 201, 209 ff
– operators 201, 139
symmorphic groups 196
symplectic manifolds 150
symplectic transformations 400
synergetics 231, 246
synthetic division 324
system analysis 231

tangent space 138
Taylor series 488
– analytical methods 35, 39
– functional analysis 72
– numerical methods 332
– ordinary differential equations 55, 353
– partial differential equations 362
– perturbation methods 394
– stochastic processes 534 f, 545 f
– variational methods 621
teleconnections 240
telegraphic equation 167
temperature distribution 60
temperature Green's functions 181 f
temporal frequency 106
tensors
– algebraic methods 6, 12 f
– differential geometry 139 f
– group theory 203
– products 451, 470
– symmetries 572 ff
tertium non datur principle 460
test space, quantum logic 454
thermal noise 523 ff
thermodynamics 258, 516 ff, 525, 536
thermohaline circulation (THC) 242
theta function 495
Thevenin's theorem 524
Thiele semi-invariants 529
time average 263
time complexity 360
time delay embedding 222, 239, 246
time dependence
– Schrödinger equations 167
– stochastic processes 516 f, 552 f
– variational methods 645
time series analysis 235
Toffoli gate 421 f, 433, 436
tomography 102, 106
topologies 587–618
– algebraic geometry 155
– differential geometry 137, 149
torus 588
traces 203, 446, 470
trajectories 114, 262
transcendental functions 475–512, 623
transfer Green's functions 161
transformations 190 ff
– canonical 631 ff, 652
– conformal 44
– mathematical 83–108
– numerical methods 313 ff
– perturbation methods 400
– quantum computation 423
– symmetries 566 ff, 569 ff
transistors 419
transit time effects 533
transition matrix 6
transition probability 260, 278
translations 13, 45
transport process 257
transpose matrix 380
transposition 10
trapezoidal quadrature rule 338, 343, 350
trapping 537, 541 ff
traveling length 257
trial move 260
triangular grids 362
triangular matrices 379
tridiagonal matrices 318, 375
trigonometric functions 131 f
triplet states 210 ff
trivial topology 592
truncation error 288
truth values 420, 462, 455
tuples 252
turing machines 418, 436
turning point problem 409
two slit experiment 418
Tychonoff theorem 596

uncertainty 641
uncertainty interval 326
unconstrained optimization 324
underdetermined linear systems 312 ff
underflow, floating points 285
unfolding 389, 412 f
unforced conservative oscillator 396
uniformity 390 f
unit cell 567
unitary groups 12
unitary transformations 423, 569 f

unity partition 596
univariate time series 235
universal machined quantum computation 420
universal unfolding 389
unsharp quantum logic 460, 470
upper-triangular system 290
upwelling, El Niño 241

Vallis description 240
van der Pol equation 396, 558
van Dyke matching rule 406
Vandermonde matrix 330
variance 236
variational methods 619–656
variational principle
– Hamiltonian 74
– partial differential equations 358
varieties 154 f
vector addition 5
vector fields 609
vector space 477 ff, 488 f
– analytical methods 50
– differential geometry 138
– group theory 191
vector spherical waves 186
vector states 447, 470
vectors 2, 5, 137
verification, quantum logic 456
Vernam cypher 430
vertex 131
Volterra series 237
Volterra type integral equations 67 f
vortex lines 175

w plane diagrams 37
Walsh transforms 103
water droplet freezing 265
wave equations 476 ff
– analytical methods 36
– Green's functions 173
– numerical methods 358
– ordinary differential equations 57
– topology 610
wavelets 230
Weber differential equations 500
Wedderburn theorem 23
wedge product
– differential geometry 141
– Grassmann ring 17
– topology 597
Weierstrass theorem 329
weight functions 336 f
Weinberg model 580
Weyl–Killing groups 26
white noise 236, 515 ff, 549
Whitney embedding 608
Whitney's theorem 141
Wick's theorem 179 ff, 184 f
Wiener–Hopf theory 68
Wiener–Khinchin theorem 519 ff, 530 f, 550, 559
Wigner group-theoretical approach 488
Wigner symmetry model 565 ff, 569 ff
Wigner–Moyal correlations 520
window factor 543
WKBJ method 408
Wronskian determinant 50, 53

Xnetlib, numerical software 378
XOR operations 420 ff

Young diagram 207
Young modulus 173

z-plane diagrams 37
z-transform 104
Zariski topology 155