PRENTICE-HALL MATHEMATICS SERIES

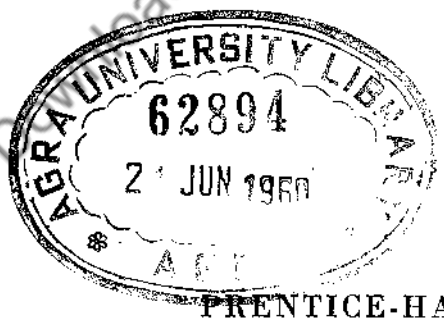
Dr. Albert A. Bennett, Editor

METHODS OF APPLIED MATHEMATICS

By

F. B. HILDEBRAND

Associate Professor of Mathematics
Massachusetts Institute of Technology



PRENTICE-HALL, INC.
Englewood Cliffs, N. J.

Library of Congress Catalog Card Number: 52-9880.

First Printing .....September, 1952
Second Printing .....August, 1954
Third Printing .....February, 1956
Fourth Printing .....January, 1958
Fifth Printing .....December, 1958

PRINTED IN THE UNITED STATES OF AMERICA

57922

Preface

The principal aim of this volume is to place at the disposal of the engineer or physicist the basis of an intelligent working knowledge of a number of facts and techniques relevant to four fields of mathematics which usually are not treated in courses of the "Advanced Calculus" type, but which are useful in varied fields of application. The text includes the result of a series of revisions of material originally prepared in mimeographed form for use at the Massachusetts Institute of Technology.

Account is taken of the fact that most students in the fields of application have neither the time nor the inclination for the study of elaborate treatments of each of these topics from the classical point of view. At the same time it is realized that efficient use of facts or techniques depends strongly upon a substantial understanding of the basic underlying principles. For this reason, care has been taken throughout the text either to provide rigorous proofs, when it is believed that those proofs can be readily comprehended by a wide class of readers, or to state the desired results as precisely as possible and indicate why those results might have been *formally* anticipated.

In each chapter, the treatment consists in showing how typical problems may arise, in establishing those parts of the relevant theory which are of principal practical significance, and in developing techniques for analytical and numerical analysis and problem solving.

Whereas experience gained from a course on the Advanced Calculus level is presumed, the treatments are almost completely self-contained, so that the nature of this preliminary course is not of great importance.

In order to increase the usefulness of the volume as a basic or supplementary text, and as a reference volume, an attempt has been made to organize the material so that there is very little essential interdependence among the chapters, and so that considerable flexibility exists with regard to the omission of topics within chapters. In addition, a large amount of supplementary material is included in annotated problems which complement numerous exercises, of varying difficulty, which are arranged in correspondence with successive sections of the text at the ends of the chapters. Answers to all problems either are incorporated into their statement or are listed at the end of the book.

The first chapter deals principally with *linear algebraic equations*, *quadratic* and *Hermitian forms*, and operations with *vectors* and *matrices*, with special emphasis on the concept of characteristic values. A brief summary of corresponding results in *function space* is included for comparison, and for convenient reference. Whereas a considerable amount of material is presented, particular care was taken here to order and even to overlap the demonstrations in such a way that maximum flexibility in selection of topics is present.

The first portion of the second chapter deals carefully with the variational notation and derives the Euler equations relevant to a large class of problems in the *calculus of variations*. More than usual emphasis is placed on the significance of natural boundary conditions. Generalized coordinates, Hamilton's principle, and Lagrange's equations are treated and illustrated within the framework of this theory. The chapter concludes with a discussion of the formulation of minimal principles of more general type, and with the application of direct and semidirect methods of the calculus of variations to the exact and approximate solution of practical problems.

The third chapter combines the presentation of available methods for solving the simpler types of *difference equations* with a description of the application of *finite-difference methods* to the approximate solution of problems governed by partial differential equations, and includes consideration of the troublesome problems of convergence and stability. Much of this material, the importance of which has increased greatly with modern developments in the field of numerical analysis, has not appeared previously in integrated form.

The concluding chapter deals with the formulation and theory of linear *integral equations*, and with exact and approximate methods for obtaining their solutions, particular emphasis being placed on the several equivalent interpretations of the relevant Green's function. Considerable supplementary material is provided in annotated problems.

Many compromises between mathematical elegance and practical significance were found to be necessary. It is hoped that the present volume will serve to ease the way of the engineer or physicist into the more advanced areas of applicable mathematics, for which his need is steadily increasing, without obscuring from him the existence of certain *difficulties* often implied by the phrase "It can be shown," and without failing to warn him of certain *dangers* involved in formal application of techniques beyond the limits inside which their validity has been well established.

The author is indebted to colleagues and students in various fields for help in selecting and revising the content and presentation, and particularly to Professor A. A. Bennett for many valuable criticisms and suggestions.

F. B. HILDEBRAND

Contents

CHAPTER 1

Matrices, Determinants, and Linear Equations

CHAPTER 2

Calculus of Variations and Applications

CHAPTER 3

Difference Equations

CHAPTER 4

Integral Equations

METHODS OF

APPLIED MATHEMATICS

CHAPTER ONE

Matrices, Determinants, and Linear Equations

1.1. Introduction. In many fields of analysis we find it necessary to deal with an *ordered set* of elements, which may be numbers or functions. In particular, we may deal with an ordinary *sequence* of the form $a_1, a_2, \ldots, a_n$, or with a two-dimensional *array* such as the rectangular arrangement

$$\begin{array}{llll} a_{11}, & a_{12}, & \cdots, & a_{1n} \\ a_{21}, & a_{22}, & \cdots, & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}, & a_{m2}, & \cdots, & a_{mn}, \end{array}$$

consisting of m rows and n columns.

When suitable laws of equality, addition and subtraction, and multiplication are associated with sets of such rectangular arrays, the arrays are called *matrices*, and are then designated by a special symbolism. The laws of combination are specified in such a way that the matrices so defined are of frequent usefulness in both practical and theoretical considerations.

Since matrices are perhaps most intimately associated with sets of *linear algebraic equations*, it is desirable to investigate the general nature of the solutions of such sets of equations by elementary methods, and hence to provide a basis for certain definitions and investigations which follow.

1.2. Linear equations. The Gauss-Jordan reduction. We deal first with the problem of attempting to obtain solutions of a set of m linear equations in n unknown variables $x_1, x_2, \ldots, x_n$, of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m \end{aligned}\right\}, \tag{1}$$

by direct calculation.

Under the assumption that (1) does indeed possess a solution, the *Gauss-Jordan reduction* proceeds as follows:

First Step. Suppose that $a_{11} \neq 0$. (Otherwise, renumber the equations or variables so that this is so.) Divide both sides of the first equation by a_{11}, so that the resultant equivalent equation is of the form

$$x_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n = c'_1. \tag{2}$$

Multiply both sides of (2) successively by $a_{21}, a_{31}, \ldots, a_{m1}$, and subtract the respective resultant equations from the second, third, $\ldots$, mth equations of (1), to reduce (1) to the form

$$\left.\begin{aligned} x_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n &= c'_1, \\ a'_{22}x_2 + \cdots + a'_{2n}x_n &= c'_2, \\ \cdots\cdots\cdots\cdots\cdots & \\ a'_{m2}x_2 + \cdots + a'_{mn}x_n &= c'_m \end{aligned}\right\}. \tag{3}$$

Second Step. Suppose that $a'_{22} \neq 0$. (Otherwise, renumber the equations or variables so that this is so.) Divide both sides of the second equation in (3) by a'_{22}, so that this equation takes the form

$$x_2 + a''_{23}x_3 + \cdots + a''_{2n}x_n = c''_2, \tag{4}$$

and use this equation, as in the first step, to eliminate the coefficient of x_2 in *all other equations* in (3), so that the set of equations becomes

$$\left.\begin{aligned} x_1 \quad + a''_{13}x_3 + \cdots + a''_{1n}x_n &= c''_1, \\ x_2 + a''_{23}x_3 + \cdots + a''_{2n}x_n &= c''_2, \\ a''_{33}x_3 + \cdots + a''_{3n}x_n &= c''_3, \\ \cdots\cdots\cdots\cdots\cdots & \\ a''_{m3}x_3 + \cdots + a''_{mn}x_n &= c''_m \end{aligned}\right\}. \tag{5}$$

Remaining Steps. Continue the above process r times until it terminates, that is, until $r = m$ *or* until all coefficients of the x's are *zero* in all equations following the rth equation. We shall speak of these $m - r$ equations as the *residual equations.*

There then exist two alternatives. *First,* it may happen that one or more of the residual equations has a nonzero right-hand member, and hence is of the form $0 = c_k^{(r)}$ (where in fact $c_k^{(r)} \neq 0$). In this case, the assumption that a solution of (1) exists leads to a *contradiction,* and hence *no solution exists.* The set (1) is then said to be *inconsistent* or *incompatible.*

Otherwise, no contradiction exists, and the set (1) of m equations is reduced to an equivalent set of r equations which, after a transposition, can be written in the form

$$\left.\begin{aligned} x_1 &= \gamma_1 + \alpha_{11}x_{r+1} + \cdots + \alpha_{1,n-r}x_n, \\ x_2 &= \gamma_2 + \alpha_{21}x_{r+1} + \cdots + \alpha_{2,n-r}x_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_r &= \gamma_r + \alpha_{r1}x_{r+1} + \cdots + \alpha_{r,n-r}x_n \end{aligned}\right\}, \tag{6}$$

where the γ's and α's are definite constants related to the coefficients in (1). Hence, in this case the most general solution of (1) expresses each of the r variables $x_1, x_2, \ldots, x_r$ as a definite constant plus a definite linear combination of the remaining $n - r$ variables, each of which can be assigned *arbitrarily.*

If $r = n$, a *unique* solution is obtained. Otherwise, we say that an $(n - r)$*-fold infinity* of solutions exists. The number $n - r = d$ may be called the *defect* of the system (1). We notice that if the system (1) is consistent and r is less than m, then $m - r$ of the equations (namely, those which correspond to the residual equations) are actually ignorable, since they are implied by the remaining r equations.

The reduction may be illustrated by considering the four simultaneous equations

$$\left.\begin{aligned} x_1 + 2x_2 - x_3 - 2x_4 &= -1, \\ 2x_1 + x_2 + x_3 - x_4 &= 4, \\ x_1 - x_2 + 2x_3 + x_4 &= 5, \\ x_1 + 3x_2 - 2x_3 - 3x_4 &= -3 \end{aligned}\right\}. \tag{7}$$

It is easily verified that after two steps in the reduction one obtains the equivalent set

$$\left.\begin{aligned} x_1 \quad + x_3 \qquad &= 3, \\ x_2 - x_3 - x_4 &= -2, \\ 0 &= 0, \\ 0 &= 0 \end{aligned}\right\}.$$

Hence the system is of defect *two*. If we write $x_3 = c_1$ and $x_4 = c_2$, it follows that the general solution can be expressed in the form

$$x_1 = 3 - c_1, \quad x_2 = -2 + c_1 + c_2, \quad x_3 = c_1, \quad x_4 = c_2, \tag{8a}$$

where c_1 and c_2 are arbitrary constants. This two-parameter family of solutions can also be written in the symbolic form

$$\{x_1, x_2, x_3, x_4\} = \{3, -2, 0, 0\} + c_1\{-1, 1, 1, 0\} + c_2\{0, 1, 0, 1\}. \tag{8b}$$

It follows also that the third and fourth equations of (7) must be consequences of the first two equations. Indeed, the third equation is obtained by subtracting the first from the second, and the fourth by subtracting one-third of the second from five-thirds of the first.

The Gauss-Jordan reduction is useful in actually obtaining numerical solutions of sets of linear equations,* and it has been presented here also for the purpose of motivating certain definitions and terminologies which follow.

1.3. Matrices. The set of equations (1) can be visualized as a *linear transformation* in which the set of n numbers $\{x_1, x_2, \ldots,$

* In place of eliminating x_k from *all* equations except the kth, in the kth step, one may eliminate x_k only in those equations *following* the kth equation. When the process terminates, after r steps, the rth unknown is given explicitly by the rth equation. The $(r-1)$th unknown is then determined by substitution in the $(r-1)$th equation, and the solution is completed by working back in this way to the first equation. The method just outlined is associated with the name of *Gauss*. In order that the "round-off" errors be as small as possible, it is desirable that the sequence of eliminations be ordered such that the coefficient of x_k in the equation used to eliminate x_k is as large as possible in absolute value, relative to the remaining coefficients in that equation.

A modification of this method, due to Crout (Reference 7), which is particularly well adapted to the use of desk computing machines, is described in an appendix.

$x_n\}$ is transformed into the set of m numbers $\{c_1, c_2, \ldots, c_m\}$. The transformation is clearly specified by the coefficients a_{ij}.

The rectangular array of these coefficients, usually enclosed in square brackets,

$$\mathbf{a} \equiv [a_{ij}] \equiv \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{9}$$

which consists of m *rows* and n *columns* of *elements*, is called an $m \times n$-*matrix* when certain laws of combination, yet to be specified, are laid down. In the symbol a_{ij}, representing a typical element, the *first* subscript (here i) denotes the row and the *second* subscript (here j) the column occupied by the element.

The sets of quantities x_i $(i = 1, 2, \ldots, n)$ and c_i $(i = 1, 2, \ldots, m)$ are conventionally represented as matrices of *one column* each. In order to emphasize the fact that a matrix consists of only one column, it is convenient to indicate it by braces, rather than brackets, and so to write

$$\mathbf{x} \equiv \{x_i\} \equiv \begin{Bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{Bmatrix}, \qquad \mathbf{c} \equiv \{c_i\} \equiv \begin{Bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_m \end{Bmatrix}. \tag{10a,b}$$

For convenience in writing, the elements of a one-column matrix are frequently arranged horizontally, the use of braces then serving to indicate the transposition.

If we visualize (1) as stating that the matrix $\mathbf{a}$ transforms the one-column matrix $\mathbf{x}$ into the one-column matrix $\mathbf{c}$, it is natural to write the transformation in the form

$$\mathbf{a}\,\mathbf{x} = \mathbf{c}, \tag{11}$$

where $\mathbf{a} = [a_{ij}]$, $\mathbf{x} = \{x_i\}$, and $\mathbf{c} = \{c_i\}$.

On the other hand, the set of equations (1) can be written in the form

$$\sum_{k=1}^{n} a_{ik}x_k = c_i \qquad (i = 1, 2, \cdots, m), \tag{12}$$

which leads to the matrix equation

$$\left\{\sum_{k=1}^{n} a_{ik}x_k\right\} = \{c_i\}. \tag{12a}$$

Hence, if (11) and (12a) are to be equivalent, we are led to the *definition*

$$\mathbf{a}\,\mathbf{x} = [a_{ik}]\{x_k\} \equiv \left\{\sum_{k=1}^{n} a_{ik}x_k\right\}. \tag{13}$$

Formally, we merely replace the *column* subscript in the general term of the *first* factor by a new *dummy index* k, and replace the *row* subscript in the general term of the *second* factor by the same dummy index, and sum over that index.

The definition is clearly applicable only when the number of *columns* in the *first* factor is equal to the number of *rows* (elements) in the *second* factor. Unless this condition is satisfied, the product is undefined.

We notice that a_{ik} is the element in the ith row and kth column of $\mathbf{a}$, and that x_k is the kth element in the one-column matrix $\mathbf{x}$. Since i ranges from 1 to m in a_{ij}, the definition (13) states that the product of an $m \times n$-matrix into an $n \times 1$-matrix is an $m \times 1$-matrix (m elements in one column). The ith element in the product is obtained from the ith row of the first factor and the single column of the second factor, by multiplying together the first elements, second elements, and so forth, and adding these products together algebraically.

Thus, for example, the definition leads to the result

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \\ -1 & 2 \end{bmatrix} \cdot \begin{Bmatrix} 1 \\ 2 \end{Bmatrix} = \begin{Bmatrix} 1 \cdot 1 + 0 \cdot 2 \\ 2 \cdot 1 + 1 \cdot 2 \\ -1 \cdot 1 + 2 \cdot 2 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 4 \\ 3 \end{Bmatrix}.$$

Now suppose that the n variables $x_1, \ldots, x_n$ are expressed as linear combinations of s new variables $y_1, \ldots, y_s$, that is, that a set of relations holds of the form

$$x_i = \sum_{k=1}^{s} b_{ik}y_k \qquad (i = 1, 2, \cdots, n). \tag{14}$$

If the original variables satisfy (12), the equations satisfied by the new variables are obtained by introducing (14) into (12). In addi-

tion to replacing i by k in (14), for this introduction, we must clearly replace k in (14) by a *new* dummy index, say l, to avoid ambiguity of notation. The result of the substitution then takes the form

$$\sum_{k=1}^{n} a_{ik}\left(\sum_{l=1}^{s} b_{kl}y_l\right) = c_i \qquad (i = 1, 2, \cdots, m), \tag{15a}$$

or, since the order in which the finite sums are formed is immaterial,

$$\sum_{l=1}^{s}\left(\sum_{k=1}^{n} a_{ik}b_{kl}\right) y_l = c_i \qquad (i = 1, 2, \cdots, m). \tag{15b}$$

In matrix notation, the transformation (14) takes the form

$$\mathbf{x} = \mathbf{b}\,\mathbf{y} \tag{16}$$

and, corresponding to (15a), the introduction of (16) into (11) gives

$$\mathbf{a}(\mathbf{b}\,\mathbf{y}) = \mathbf{c}. \tag{17}$$

But if we write

$$p_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} \qquad \begin{pmatrix} i = 1, 2, \cdots, m \\ j = 1, 2, \cdots, s \end{pmatrix} \tag{18}$$

equation (15b) takes the form

$$\sum_{l=1}^{s} p_{il}y_l = c_i \qquad (i = 1, 2, \cdots, m),$$

and hence, in accordance with (12) and (13), the matrix form of the transformation (15b) is

$$\mathbf{p}\,\mathbf{y} = \mathbf{c}. \tag{19}$$

Thus it follows that the result of operating on $\mathbf{y}$ by $\mathbf{b}$, and on the product by $\mathbf{a}$ [given by the left-hand member of (17)], is the same as the result of operating on $\mathbf{y}$ directly by the matrix $\mathbf{p}$. We accordingly *define* this matrix to be the product $\mathbf{a}\,\mathbf{b}$,

$$\mathbf{a}\,\mathbf{b} = [a_{ik}][b_{kj}] \equiv \left[\sum_{k=1}^{n} a_{ik}b_{kj}\right]. \tag{20}$$

Recalling that the first subscript in each case is the row index and the second the column index, we see that if the first factor has

m rows and n columns, and the second n rows and s columns, the index i in the *right-hand* member may vary from 1 to m while the index j in that member may vary from 1 to s. Hence, the *product of an $m \times n$-matrix into an $n \times s$-matrix is an $m \times s$-matrix.* The element p_{ij} in the ith row and jth column of the product is formed by multiplying together corresponding elements of the ith *row* of the *first* factor and the jth *column* of the *second* factor, and adding the results algebraically.

Thus, for example, we have

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} (1\cdot 1 + 0\cdot 1 + 1\cdot 2)(1\cdot 2 + 0\cdot 0 + 1\cdot 1)(1\cdot 1 + 0\cdot 1 + 1\cdot 0) \\ (1\cdot 1 - 2\cdot 1 + 1\cdot 2)(1\cdot 2 - 2\cdot 0 + 1\cdot 1)(1\cdot 1 - 2\cdot 1 + 1\cdot 0) \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 3 & 1 \\ 1 & 3 & -1 \end{bmatrix}.$$

We notice that **a b** is defined only if the number of *columns* in **a** is equal to the number of *rows* in **b**. In this case, the two matrices are said to be *conformable* in the order stated.

If **a** is an $m \times n$-matrix and **b** an $n \times m$-matrix, then **a** and **b** are conformable in either order, the product **a b** then being a *square* matrix of order m and the product **b a** a square matrix of order n. Even in the case when **a** and **b** are square matrices of the same order the products **a b** and **b a** are not generally equal. For example, in the case of two square matrices of order two we have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix},$$

and also

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{21}b_{12} & a_{12}b_{11} + a_{22}b_{12} \\ a_{11}b_{21} + a_{21}b_{22} & a_{12}b_{21} + a_{22}b_{22} \end{bmatrix}.$$

Thus, in multiplying **b** by **a** in such cases, we must carefully distinguish *pre*multiplication (**a b**) from *post*multiplication (**b a**).

The *sum* of two $m \times n$-matrices $[a_{ij}]$ and $[b_{ij}]$ is defined to be the matrix $[a_{ij} + b_{ij}]$. Further, the product of a number k and a

matrix $[a_{ij}]$ is defined to be the matrix $[k\,a_{ij}]$, in which *each* element of the original matrix is multiplied by k.

Two $m \times n$-matrices are said to be *equal* if and only if corresponding elements in the two matrices are equal.

From the preceding definitions, it is easily shown that, if $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are each $m \times n$-matrices, *addition* is *commutative* and *associative:*

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}, \qquad \mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}. \tag{21}$$

Also, if the relevant products are defined, *multiplication* of matrices is *associative,*

$$\mathbf{a}(\mathbf{b}\,\mathbf{c}) = (\mathbf{a}\,\mathbf{b})\mathbf{c}, \tag{22}$$

and *distributive,*

$$\mathbf{a}(\mathbf{b} + \mathbf{c}) = \mathbf{a}\,\mathbf{b} + \mathbf{a}\,\mathbf{c}, \qquad (\mathbf{b} + \mathbf{c})\mathbf{a} = \mathbf{b}\,\mathbf{a} + \mathbf{c}\,\mathbf{a}, \tag{23}$$

but, in general, *not commutative.*

It is consistent with these definitions to divide a given matrix into smaller submatrices, the process being known as the *partitioning* of a matrix. Thus, we may partition a square matrix $\mathbf{a}$ of order three *symmetrically* as follows:

$$\mathbf{a} = \left[\begin{array}{cc:c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hdashline a_{31} & a_{32} & a_{33} \end{array}\right] \equiv \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} \\ \mathbf{b}_{21} & \mathbf{b}_{22} \end{bmatrix}$$

where the elements of the partitioned form are the *matrices*

$$\mathbf{b}_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \mathbf{b}_{12} = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix},$$

$$\mathbf{b}_{21} = [a_{31} \quad a_{32}], \qquad \mathbf{b}_{22} = [a_{33}].$$

If a second square matrix of order three is similarly partitioned, the submatrices can be treated as single elements and the usual laws of matrix multiplication and addition can be applied to the two matrices so partitioned, as is easily verified.

More generally, if two conformable matrices in a product are partitioned, necessary and sufficient conditions that this statement apply are that to each vertical partition line separating *columns* r and $r + 1$ in the *first* factor there correspond a horizontal partition

line separating *rows* r and $r+1$ in the *second* factor, and that no additional *horizontal* partition lines be present in the *second* factor.

1.4. Determinants. Cramer's rule. In this section we review certain properties of *determinants*. Associated with any *square* matrix $[a_{ij}]$ of order n we define the *determinant* $|\mathbf{a}| = |a_{ij}|$,

$$|\mathbf{a}| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix},$$

as a *number* obtained as the sum of all possible products in each of which there appears one and only one element from each row and each column, each such product being assigned a plus or minus sign according to the following rule: *Let the elements involved in a given product be joined in pairs by line segments. If the total number of such segments sloping upward to the right is even, prefix a plus sign to the product. Otherwise, prefix a negative sign.**

From this definition, the following properties of determinants, which greatly simplify their actual evaluation, are easily established:

1. If all elements of any row or column of a square matrix are zeros, its determinant is zero.

2. The value of the determinant is unchanged if the rows and columns of the matrix are interchanged.

3. If two rows (or two columns) of a square matrix are interchanged, the sign of its determinant is changed.

4. If all elements of one row (or one column) of a square matrix are multiplied by a number k, the determinant is multiplied by k.

5. If corresponding elements of two rows (or two columns) are equal or in a constant ratio, then the determinant is zero.

6. If each element in one row (or one column) is expressed as the sum of two terms, then the determinant is equal to the sum of two determinants, in each of which one of the two terms is deleted in each element of that row (or column).

* This statement of the rule of signs is equivalent to the statement which involves *inversions* of subscripts. It possesses the advantage of being readily applicable in actual cases when the elements are numbers (or functions) and are not provided with explicit subscripts. Also, with this statement of the rule, the proofs of the properties which follow are in general simplified.

7. If to the elements of any row (column) are added k times the corresponding elements of any other row (column), the determinant is unchanged.*

If the row and column containing an element a_{ij} in a square matrix $\mathbf{a}$ are deleted, the determinant of the remaining square array is called the *minor* of a_{ij}, and is denoted here by M_{ij}. The *cofactor* of a_{ij}, denoted here by A_{ij}, is then defined by the relation

$$A_{ij} = (-1)^{i+j} M_{ij}. \tag{24}$$

Thus if the sum of the row and column indices of an element is *even*, the cofactor and the minor of that element are identical; otherwise they differ in sign.

It is a consequence of the definition of a determinant that *the cofactor of a_{ij} is the coefficient of a_{ij} in the expansion of* $|\mathbf{a}|$. This fact leads to the important *Laplace expansion* formula:

$$|\mathbf{a}| = \sum_{k=1}^{n} a_{ik} A_{ik} \qquad \text{or} \qquad |\mathbf{a}| = \sum_{k=1}^{n} a_{kj} A_{kj}, \tag{25a,b}$$

for any relevant value of i or j. This formula states that *a determinant is equal to the sum of the products of the elements of any single row or column by their cofactors.*

If a_{ik} is replaced by a_{rk} in (25a), the result $\Sigma\, a_{rk} A_{ik}$ must accordingly be the determinant of a new matrix in which the elements of the ith row are replaced by the corresponding elements of the rth row, and hence must vanish if $r \neq i$ in virtue of Property 5. An analogous result follows if a_{kj} is replaced by a_{ks} in (25b), when $s \neq j$. Thus, in addition to (25), we have the relations

$$\sum_{k=1}^{n} a_{rk} A_{ik} = 0 \quad (r \neq i), \qquad \sum_{k=1}^{n} a_{ks} A_{kj} = 0 \quad (s \neq j). \tag{26a,b}$$

These results lead directly to *Cramer's rule* for solving a set of n linear equations in n unknown quantities, of the form

$$\sum_{k=1}^{n} a_{ik} x_k = c_i \qquad (i = 1, 2, \cdots, n), \tag{27}$$

* It can be shown that if we impose the condition that Properties 4 and 7 hold, and in addition impose the requirement that the determinant be unity when the diagonal elements are unity and all other elements are zero, then these conditions imply all other properties of determinants, and may serve as the definition of a determinant.

in the case when the determinant of the matrix of coefficients is not zero,

$$|\, a_{ij} \,| \neq 0. \tag{28}$$

For if we multiply both sides of (27) by A_{ir}, where r is any integer between 1 and n, and sum the results with respect to i, there follows (after an interchange of order of summation)

$$\sum_{k=1}^{n} \left(\sum_{i=1}^{n} a_{ik} A_{ir} \right) x_k = \sum_{i=1}^{n} c_i A_{ir} \qquad (r = 1, 2, \cdots, n). \tag{29}$$

In virtue of (25b) and (26b), the inner sum in (29) vanishes unless $k = r$ and is equal to $|\, \mathbf{a} \,|$ in that case. Hence (29) takes the form

$$|\, \mathbf{a} \,| \, x_r = \sum_{i=1}^{n} A_{ir} c_i \qquad (r = 1, 2, \cdots, n). \tag{30}$$

The expansion on the right in (30) differs from the right-hand member of the expansion

$$|\, \mathbf{a} \,| = \sum_{i=1}^{n} A_{ir} a_{ir}$$

only in the fact that the column $\{c_i\}$ replaces the column $\{a_{ir}\}$ of the coefficients of x_r in $\mathbf{a}$. Thus, if $|\, \mathbf{a} \,| \neq 0$, we deduce *Cramer's rule*, which can be stated as follows:

When the determinant $|\, \mathbf{a} \,|$ *of the matrix of coefficients in a set of n linear algebraic equations in n unknowns* $x_1, \cdots, x_n$ *is not zero, that set of equations has a unique solution. The expression for any* x_r *is the ratio of two determinants, the denominator being the determinant of the matrix of coefficients, and the numerator being the determinant of the matrix obtained by replacing the column of the coefficients of* x_r *in the coefficient matrix by the column of the right-hand members.**

In the case when all right-hand members c_i are zero, the equations are said to be *homogeneous*. In this case, one solution is clearly the *trivial* one $x_1 = x_2 = \cdots = x_n = 0$. The preceding result then states that this is the only possible solution if $|\, \mathbf{a} \,| \neq 0$, so that *a set of n linear homogeneous equations in n unknowns can-*

* The proof given here shows only that if there is a solution, then it is given by Cramer's rule. That the expressions given do indeed satisfy (27) can be shown by direct substitution.

not possess a nontrivial solution unless the determinant of the coefficient matrix vanishes.

We postpone the treatment of the case when $|\mathbf{a}| = 0$, as well as the case when the number of equations differs from the number of unknowns, until Sections 1.8 and 1.10.

It can be shown that the *product* of two determinants $|a_{ij}|$ and $|b_{ij}|$ of the same order n can be calculated by a rule completely analogous to that corresponding to matrix multiplication:

$$|a_{ij}| \cdot |b_{ij}| = \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|. \tag{31}$$

Consequently, it follows that the *determinant of the product* of two square matrices of the same order is equal to the *product of the determinants*:

$$|\mathbf{a}| \cdot |\mathbf{b}| = |\mathbf{a}\,\mathbf{b}|. \tag{32}$$

A square matrix whose determinant vanishes is called a *singular* matrix. From (32) it follows that *the product of two nonsingular matrices is also nonsingular.*

It follows from the definitions that the determinant of the negative of a square matrix is *not* necessarily the negative of the determinant, but that one has the relationship

$$|-\mathbf{a}| = (-1)^n |\mathbf{a}|,$$

where n is the order of the matrix $\mathbf{a}$.

1.5. Special matrices. In this section we define certain matrices which are of special importance, and investigate some of their properties.

That matrix which is obtained from $\mathbf{a} = [a_{ij}]$ by interchanging rows and columns is called the *transpose* of $\mathbf{a}$, and is here indicated by $\mathbf{a}^T$:

$$\mathbf{a}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{33}$$

Thus the transpose of an $m \times n$-matrix is an $n \times m$-matrix. If the element in row r and column s of $\mathbf{a}$ is a_{rs}, where r may vary from 1 to m and s from 1 to n, then the element a'_{rs} in row r and column s of

$\mathbf{a}^T$ is given by $a'_{rs} = a_{sr}$, where now r may vary from 1 to n and s from 1 to m.

If $\mathbf{a}$ is an $m \times l$-matrix and $\mathbf{b}$ is an $l \times n$-matrix, then both the products $\mathbf{a}\,\mathbf{b}$ and $\mathbf{b}^T\,\mathbf{a}^T$ exist, the former being an $m \times n$-matrix, and the latter an $n \times m$-matrix. We show next that the latter matrix is the transpose of the former. Since the element in row r and column s of the product $\mathbf{a}\,\mathbf{b} \equiv \mathbf{c}$ is given by

$$\sum_{k=1}^{l} a_{rk} b_{ks} \equiv c_{rs},$$

where r may vary from 1 to m and s from 1 to n, whereas the element c'_{rs} in row r and column s of the product $\mathbf{b}^T\,\mathbf{a}^T$ is given by

$$c'_{rs} \equiv \sum_{k=1}^{l} b'_{rk} a'_{ks} = \sum_{k=1}^{l} b_{kr} a_{sk} = c_{sr},$$

where now r may vary from 1 to n and s from 1 to m, it follows that $\mathbf{b}^T\,\mathbf{a}^T$ is indeed the transpose of $\mathbf{a}\,\mathbf{b}$.

Thus, we have shown that *the transpose of* $\mathbf{a}\,\mathbf{b}$ *is the product of the transposes in reverse order*:

$$(\mathbf{a}\,\mathbf{b})^T = \mathbf{b}^T\,\mathbf{a}^T. \tag{34}$$

This result will be of frequent usefulness.

When $\mathbf{a}$ is a *square* matrix, the matrix obtained from $\mathbf{a}$ by replacing each element by its cofactor and then interchanging rows and columns is called the *adjoint* of $\mathbf{a}$:

$$\text{Adj } \mathbf{a} = \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} = [A_{ji}]. \tag{35}$$

The adjoint of a product is found to be equal to the product of the adjoints *in the reverse order*.

The *unit matrix* I of order n is the *square* matrix having ones in its *principal diagonal* and zeros elsewhere,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{36}$$

while the *zero matrix* **0** has zeros for *all* its elements. It is readily verified that for any square matrix **a** there follow

$$\mathbf{a}\,\mathbf{I} = \mathbf{I}\,\mathbf{a} = \mathbf{a} \tag{37}$$

and

$$\mathbf{a}\,\mathbf{0} = \mathbf{0}\,\mathbf{a} = \mathbf{0}. \tag{38}$$

The notation of the so-called *Kronecker delta*,

$$\delta_{pq} = \begin{cases} 0 \text{ when } p \neq q, \\ 1 \text{ when } p = q, \end{cases} \tag{39}$$

is frequently useful. With this notation, the general term of the unit matrix is merely δ_{ij}; that is, we can write

$$\mathbf{I} = [\delta_{ij}]. \tag{40}$$

More generally, if all elements of a square matrix except those in the principal diagonal are zeros, the matrix is said to be a *diagonal matrix*. A diagonal matrix can thus be written in the form

$$\mathbf{d} = [d_i\,\delta_{ij}] = [d_j\,\delta_{ij}]$$

where the diagonal elements, for which $i = j$, are $d_1, d_2, \ldots, d_n$. *Pre*multiplication of a matrix **a** by **d** multiples the ith *row* of **a** by d_i; *post*multiplication multiplies the jth *column* by d_j. This result follows from the calculations

$$\mathbf{d}\,\mathbf{a} = [d_i\,\delta_{ik}][a_{kj}] = \Big[\sum_{k=1}^{n} d_i\,\delta_{ik}a_{kj}\Big] = [d_i a_{ij}]$$

and

$$\mathbf{a}\,\mathbf{d} = [a_{ik}][d_k\,\delta_{kj}] = \Big[\sum_{k=1}^{n} a_{ik}d_k\,\delta_{kj}\Big] = [a_{ij}d_j] = [d_j a_{ij}].$$

A diagonal matrix whose diagonal elements are all *equal* is called a *scalar matrix*. Thus, a scalar matrix must be of the form $k\,\mathbf{I} = [k\,\delta_{ij}]$.

1.6. The inverse matrix. With the notation of (39), the two equations (25a) and (26a) can be combined in the form

$$\sum_{k=1}^{n} a_{ik}A_{jk} = |\,\mathbf{a}\,|\,\delta_{ij}, \tag{41a}$$

while (25b) and (26b) lead to the relation

$$\sum_{k=1}^{n} a_{kj}A_{ki} = |\,\mathbf{a}\,|\,\delta_{ij}. \tag{41b}$$

If we write temporarily

$$\alpha_{ij} = \frac{A_{ji}}{|\,\mathbf{a}\,|}, \tag{42}$$

under the assumption that $|\,\mathbf{a}\,| \neq 0$, these equations become

$$\sum_{k=1}^{n} a_{ik}\alpha_{kj} = \delta_{ij}, \qquad \sum_{k=1}^{n} \alpha_{ik}a_{kj} = \delta_{ij}. \tag{43a,b}$$

Hence, reviewing the definition (20) of the matrix product, we see that these equations imply the matrix equations

$$[a_{ik}][\alpha_{kj}] = \mathbf{I}, \qquad [\alpha_{ik}][a_{kj}] = \mathbf{I}. \tag{44}$$

That is, the matrix $\boldsymbol{\alpha} = [\alpha_{ij}]$ has the property that

$$\mathbf{a}\,\boldsymbol{\alpha} = \boldsymbol{\alpha}\,\mathbf{a} = \mathbf{I}, \tag{45}$$

where $\mathbf{I}$ is the unit matrix. It is natural to define this matrix to be the *inverse* or *reciprocal* of $\mathbf{a}$, and to write $\boldsymbol{\alpha} = \mathbf{a}^{-1}$. We notice that *a singular matrix does not possess an inverse.*

Further, *a matrix can have only one inverse.* To prove this statement, we assume that the contrary is true and show that a contradiction follows. That is, we suppose that $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$ is such that

$$\mathbf{a}\,\boldsymbol{\beta} = \mathbf{I}.$$

If we premultiply both sides of this equation by $\boldsymbol{\alpha}$ and use (45) and (37), there follows

$$(\boldsymbol{\alpha}\,\mathbf{a})\boldsymbol{\beta} = \boldsymbol{\alpha}\,\mathbf{I} \qquad \text{or} \qquad \boldsymbol{\beta} = \boldsymbol{\alpha},$$

contradicting the assumption that $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$.

We conclude that *if the square matrix* $\mathbf{a} = [a_{ij}]$ *is nonsingular, it possesses a unique inverse* $\mathbf{a}^{-1}$ *such that*

$$\mathbf{a}^{-1}\,\mathbf{a} = \mathbf{a}\,\mathbf{a}^{-1} = \mathbf{I}, \tag{46}$$

and that inverse is of the form

$$\mathbf{a}^{-1} = [\alpha_{ij}] \qquad \textit{where} \qquad \alpha_{ij} = \frac{A_{ji}}{|\,\mathbf{a}\,|}. \tag{47}$$

Thus, to obtain the inverse of a nonsingular square matrix $[a_{ij}]$, we may first replace a_{ij} by its cofactor $A_{ij} = (-1)^{i+j}M_{ij}$, then interchange rows and columns and divide each element by the determinant $|\,a_{ij}\,|$. In the terminology of Section 1.5, the inverse of $\mathbf{a}$ is the adjoint of $\mathbf{a}$ divided by the determinant of $\mathbf{a}$:

$$\mathbf{a}^{-1} = \frac{1}{|\,\mathbf{a}\,|}\,\text{Adj}\,\mathbf{a}. \tag{48}$$

This equation can also be written in the useful form

$$\mathbf{a}\,\text{Adj}\,\mathbf{a} = |\,\mathbf{a}\,|\,\mathbf{I}. \tag{48a}$$

It may be noticed that equation (48a) also follows directly from (41a), and hence is valid even when $|\,\mathbf{a}\,| = 0$.

To determine the inverse of a *product* of nonsingular square matrices, we write

$$\mathbf{a}\,\mathbf{b} = \mathbf{c}.$$

If we premultiply both sides of this equation successively by $\mathbf{a}^{-1}$ and $\mathbf{b}^{-1}$, there follows

$$\mathbf{I} = \mathbf{b}^{-1}\,\mathbf{a}^{-1}\,\mathbf{c}$$

and hence, postmultiplying both sides of this equation by $\mathbf{c}^{-1}$ and replacing $\mathbf{c}$ by $\mathbf{a}\,\mathbf{b}$, we obtain the rule

$$(\mathbf{a}\,\mathbf{b})^{-1} = \mathbf{b}^{-1}\,\mathbf{a}^{-1}. \tag{49}$$

To illustrate the use of the inverse matrix, we consider again the problem of solving the set of linear equations (27) under the assumption (28). In matrix notation we have

$$\mathbf{a}\,\mathbf{x} = \mathbf{c},$$

and hence, by premultiplying both sides by $\mathbf{a}^{-1}$, there follows

$$\mathbf{x} = \mathbf{a}^{-1}\,\mathbf{c} \tag{50a}$$

or

$$\begin{Bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{Bmatrix} = \frac{1}{|\,\mathbf{a}\,|}\begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ & & & \\ \cdots & \cdots & \cdots & \cdots \\ & & & \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}\begin{Bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{Bmatrix} \tag{50b}$$

or

$$x_i = \frac{1}{|\mathbf{a}|}(A_{1i}c_1 + A_{2i}c_2 + \cdots + A_{ni}c_n) \qquad (i = 1, 2, \cdots, n), \tag{50c}$$

in accordance with the expanded form of Cramer's rule.

1.7. Elementary operations. Associated with a set of m linear equations in n unknowns,

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m \end{aligned}\right\}, \tag{51}$$

we consider two matrices: the $m \times n$-matrix of coefficients $[a_{ij}]$, and the $m \times (n+1)$-matrix formed by joining to the columns of $[a_{ij}]$ the column of constants $\{c_i\}$. We refer to the former matrix as the *coefficient matrix* and to the second as the *augmented matrix*.

As was shown in Section 1.2, we may use the Gauss-Jordan reduction (renumbering certain equations or variables, if necessary) to replace (51) by an equivalent set of equations of the form

$$\left.\begin{aligned} x_1 \qquad - \alpha_{11}x_{r+1} - \cdots - \alpha_{1,n-r}x_n &= \gamma_1, \\ x_2 \qquad - \alpha_{21}x_{r+1} - \cdots - \alpha_{2,n-r}x_n &= \gamma_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ x_r - \alpha_{r1}x_{r+1} - \cdots - \alpha_{r,n-r}x_n &= \gamma_r, \\ 0 &= \gamma_{r+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ 0 &= \gamma_m \end{aligned}\right\}, \tag{52}$$

by a process which involves, in addition to possible renumbering, only the addition of equal quantities to equal quantities. Accordingly, the augmented matrix of (51) is transformed to the augmented matrix of (52), which is of the form

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & -\alpha_{11} & -\alpha_{12} & \cdots & -\alpha_{1,n-r} & \gamma_1 \\ 0 & 1 & \cdots & 0 & -\alpha_{21} & -\alpha_{22} & \cdots & -\alpha_{2,n-r} & \gamma_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & -\alpha_{r1} & -\alpha_{r2} & \cdots & -\alpha_{r,n-r} & \gamma_r \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \gamma_{r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \gamma_m \end{bmatrix}, \quad (53)$$

and, at the same time, the coefficient matrix of (51) transforms into the result of deleting the extreme right-hand column of the matrix (53).

The steps in the reduction involve only the following so-called *elementary operations:*

1. The interchange of two rows or of two columns.

2. The multiplication of the elements of a row by a number other than zero.

3. The addition, to the elements of a row, of k times the corresponding elements of another row.

We define the *rank* of a matrix as *the order of the largest square array in that matrix* (formed by deleting certain rows and columns) *whose determinant does not vanish.* It is clear that the transformed *coefficient* matrix in the above case is of rank r, whereas if one or more of the numbers $\gamma_{r+1}, \gamma_{r+2}, \ldots, \gamma_m$ is not zero, the rank of the transformed *augmented* matrix is $r + 1$. If $\gamma_{r+1} = \gamma_{r+2} = \ldots = \gamma_m = 0$, *both* transformed matrices are of rank r.

It is next shown that *the ranks of the two matrices associated with* (51) *are the same as the ranks of the corresponding transformed matrices,* that is, that *the rank of a matrix is not changed by the elementary operations* listed above.

Suppose that a matrix $[a_{ij}]$ is of rank r, that is, that all determinants of order greater than r vanish, but at least one square array $\mathbf{A}$ of order r possesses a nonvanishing determinant. *Operation 1* is equivalent to renumbering rows or columns, and obviously cannot affect over-all vanishing or nonvanishing of determinants. Similarly, *operation 2* can only multiply certain determinants by a nonzero constant.

According to Property 7 of determinants (page 11), *operation 3* does not change the value of any determinant which involves either *both* or *neither* of the two rows concerned. We need show here only

that a nonvanishing determinant of *largest* order r is not reduced to zero, and that no nonvanishing determinant of higher order is *introduced* by this operation. To simplify the notation, we suppose that the square array $\mathbf{A}$ of order r in the upper left corner of the original matrix has a nonvanishing determinant, and consider the following matrix:

$$\mathbf{M} = \begin{bmatrix} a_{11} & \cdots & a_{1r} & a_{1s} \\ \cdots & \cdots & \cdots & \cdots \\ a_{r1} & \cdots & a_{rr} & a_{rs} \\ a_{q1} & \cdots & a_{qr} & a_{qs} \end{bmatrix} \equiv \left[\begin{array}{ccc|c} & & & a_{1s} \\ & & & \cdots \\ & \mathbf{A} & & \cdots \\ & & & a_{rs} \\ \hline a_{q1} & \cdots & a_{qr} & a_{qs} \end{array}\right] \tag{54}$$

where $s > r$ and $q > r$. Then, if the original matrix is of rank r, the determinant of this square matrix must vanish for all such s and q.

Now it is possible to determine constants $\lambda_1, \lambda_2, \cdots, \lambda_r$ such that the equations

$$\left.\begin{aligned} \lambda_1 a_{11} + \lambda_2 a_{21} + \cdots + \lambda_r a_{r1} &= a_{q1}, \\ \lambda_1 a_{12} + \lambda_2 a_{22} + \cdots + \lambda_r a_{r2} &= a_{q2}, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ \lambda_1 a_{1r} + \lambda_2 a_{2r} + \cdots + \lambda_r a_{rr} &= a_{qr} \end{aligned}\right\} \tag{55}$$

are satisfied, since the coefficient determinant $|\mathbf{A}|$ assuredly does not vanish and Cramer's rule applies. Hence, with these constants of combination, we can determine a row of elements which is a linear combination of the first r rows of (54), and which will have its first r elements identical with the first r elements of the last row. Let the last element of that combination be denoted by a'_{qs}. In evaluating the *determinant* of $\mathbf{M}$, we may subtract this linear combination of the first r rows from the last row without changing the value of the determinant, to obtain the result

$$|\mathbf{M}| = \begin{vmatrix} a_{11} & \cdots & a_{1r} & a_{1s} \\ a_{21} & \cdots & a_{2r} & a_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ a_{r1} & \cdots & a_{rr} & a_{rs} \\ 0 & \cdots & 0 & a_{qs} - a'_{qs} \end{vmatrix} \tag{56}$$

But since $|\mathbf{M}|$ is equal, by the Laplace expansion, to the product of $(a_{qs} - a'_{qs})$ and the determinant $|\mathbf{A}|$ which does *not* vanish, by assumption, and since $|\mathbf{M}| = 0$, it follows that $a'_{qs} = a_{qs}$. Hence we see that *the last row of* (54) *is a linear combination of the first r rows*. Since this is true for *any* q and s greater than r, the result can be stated as follows:

If a matrix is of rank r, and a set of r rows containing a nonvanishing determinant of order r is selected, then any other row in the matrix is a linear combination of these r rows.

The same statement is easily seen to be true, by a similar argument, if the word "row" is replaced by "column" throughout.

This result now shows that adding k times the qth row to the ith row, where $i \leqq r$, cannot *reduce* the rank of $\mathbf{a}$. For either the first r elements of the qth row are combinations of corresponding elements of rows of $\mathbf{A}$ *excluding* the ith, and $|\mathbf{A}|$ is unchanged, or the $r \times r$ array which is obtained by deleting the ith row of $\mathbf{A}$ and joining the row of the first r elements of the qth row, and which is unaffected, is nonsingular. Conversely, operation 3 cannot *increase* the rank since the reversed operation (which would be of the same type) would then *reduce* the rank of the *new* matrix.

Since, by an argument analogous to this one, we see that if operations 2 or 3 were effected also on *columns*, rather than *rows*, the same result would follow, we may deduce also that *the elementary operations, applied to rows or to columns, do not change the rank of a matrix.*

1.8. Solvability of sets of linear equations. If we notice that, in the Gauss-Jordan reduction, no *column* operations are involved, except perhaps a renumbering of certain columns of the *coefficient* matrix, we conclude both that the augmented matrices of (51) and (52) are of equal rank, and also that the same is true of the coefficient matrices.

If and only if one or more of the numbers $\gamma_{r+1}, \gamma_{r+2}, \ldots, \gamma_m$ in (52) is not zero, the given set of equations possesses no solution. But if and only if this is so, the rank of the augmented matrix is greater than the rank of the coefficient matrix. Thus we deduce the following basic result:

A set of linear equations possesses a solution if and only if the rank of the augmented matrix is equal to the rank of the coefficient matrix.

If the two ranks *are* both equal to r, and if a set of r equations containing a nonvanishing determinant of order r is selected, then all other equations are implied by these equations (since their coefficients are linear combinations of the coefficients of the r basic equations), and hence may be disregarded. The $n - r$ unknowns whose coefficients are *not* involved in this determinant can be assigned arbitrary values, after which the remaining r unknowns can be determined in terms of them (by Cramer's rule or otherwise).*

In particular, *if $r = n$ the n unknowns are determined uniquely.* Otherwise, if $n - r = d > 0$, the most general solution involves d independent arbitrary constants.

In the *homogeneous* case, when the right-hand members of (51) are all zeros, the coefficient matrix and the augmented matrix are automatically of equal rank, and a solution *always* exists. But this fact is obvious, since such a set of equations is always satisfied by the *trivial solution* $x_1 = x_2 = \cdots = x_n = 0$. If the rank r of the coefficient matrix is equal to the number n of unknowns, then this is the *only* solution, in accordance with the special results of Section 1.4. However, if $r < n$ (in particular, if the number of equations is less than the number of unknowns) infinitely many solutions exist, the number of independent arbitrary constants involved being given by the difference $n - r$.

We notice that, in consequence of the *linearity* of the relevant equations, the general solution of a *nonhomogeneous* set of equations is the sum of any one *particular* solution of that set and the most general solution of the associated homogeneous set.

A case of particular interest is that of a set of n homogeneous equations in n unknowns, in which the coefficient matrix is of rank $n - 1$; that is, a set of the form

$$\sum_{k=1}^{n} a_{ik}x_k = 0 \qquad (i = 1, 2, \cdots, n) \tag{57}$$

where

$$|\, a_{ij} \,| = 0 \tag{58}$$

but at least one determinant of order $n - 1$ formed from the elements of a does not vanish. In consequence of equations (41a) and

* In actual numerical cases, a procedure such as that of the Gauss or Gauss-Jordan reduction avoids the necessity of perhaps evaluating a large number of determinants. However, the results obtained here are of great importance in more general considerations.

(58), these equations are satisfied by the expressions

$$x_i = C\,A_{si} \qquad (i = 1, 2, \cdots, n) \tag{59}$$

where C is an arbitrary constant and s may take on any value from 1 to n. Since here $d = n - r = 1$, and since (59) contains one arbitrary constant, (59) must represent the most general solution of (57) unless, for the particular value of s chosen, all cofactors A_{si} happen to vanish. (This exception cannot exist for *all* values of s if the rank of $\mathbf{a}$ is $n - 1$.) With this reservation, the result obtained is equivalent to the statement that, in the case under consideration, *the unknowns are proportional to the cofactors of their coefficients in any row of the matrix* $[a_{ij}]$.

1.9. Linear vector space. The preceding results have interesting and instructive interpretations in terms of so-called "vector space," which is briefly discussed in this section.

It is conventional to speak of a one-column matrix $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ or of its transpose, the one-row matrix $\mathbf{x}^T = [x_1, x_2, \ldots, x_n]$, as a *vector*. In *two-dimensional* space, the *elements* of the vector $\{x_1, x_2\}$ can be considered as the *components* of $\mathbf{x}$ in the directions of the rectangular coordinate (x_1- and x_2-) axes. The square of the *length* of this vector is given by $l^2 = x_1^2 + x_2^2 = \mathbf{x}^T\mathbf{x}$.* Also, if $\mathbf{u}$ and $\mathbf{v}$ are two vectors in two-dimensional space, the *scalar product* of $\mathbf{u}$ and $\mathbf{v}$ is defined to be $u_1v_1 + u_2v_2 = \mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$. It is seen that the scalar product $\mathbf{u}^T\mathbf{v}$ is the equivalent, in matrix notation, of the "dot product" $\mathbf{u} \cdot \mathbf{v}$ in vector analysis. We recall that the vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal (perpendicular) if and only if this scalar product vanishes. The vectors $\mathbf{i}_1 = \{1, 0\}$ and $\mathbf{i}_2 = \{0, 1\}$ are the orthogonal *unit vectors* ordinarily denoted by $\mathbf{i}$ and $\mathbf{j}$, respectively, in vector analysis.

The above terminology is extended by analogy to the *general case of n dimensions*. When $n > 3$, it is impossible to visualize the vectors geometrically. However, we use the language associated with space of two or three dimensions, and say that an n-dimensional coordinate system comprises n mutually orthogonal axes, that a point has n corresponding coordinates, and that a vector has n components along these axes. The *scalar product* of two vectors $\mathbf{u}$

* A product of the form $\mathbf{x}^T\mathbf{x}$, which is truly a one-element matrix, is conventionally treated as a scalar.

and $\mathbf{v}$ is defined to be

$$\mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n \tag{60}$$

and the *square of the length* of a vector $\mathbf{u}$ is defined to be

$$l^2 = \mathbf{u}^T \mathbf{u} = u_1^2 + u_2^2 + \cdots + u_n^2. \tag{61}$$

It is convenient to denote the scalar product by the abbreviation $(\mathbf{u}, \mathbf{v})$,

$$(\mathbf{u}, \mathbf{v}) \equiv \mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u}. \tag{62a}$$

Two vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be *orthogonal* if their scalar product vanishes, $(\mathbf{u}, \mathbf{v}) = 0$. A *zero vector* is thus orthogonal to *all* vectors. A vector is said to be a *unit vector* if its length l is unity, so that $(\mathbf{u}, \mathbf{u}) = 1$. It is convenient to use the abbreviation

$$\mathbf{u}^2 \equiv (\mathbf{u}, \mathbf{u}). \tag{62b}$$

When the components of the vectors are *complex* numbers, the definitions (60) to (62) are inconvenient, and are conventionally modified. We denote by $\bar{\mathbf{a}}$ the matrix obtained from *any* matrix $\mathbf{a}$ by replacing all complex elements by their complex conjugates, and call this matrix the *complex conjugate matrix*.

The *Hermitian scalar product* of a vector $\mathbf{u}$ into a vector $\mathbf{v}$ is then defined as

$$(\bar{\mathbf{u}}, \mathbf{v}) = \bar{\mathbf{u}}^T \mathbf{v} = \bar{u}_1 v_1 + \bar{u}_2 v_2 + \cdots + \bar{u}_n v_n = (\mathbf{v}, \bar{\mathbf{u}}), \tag{63a}$$

and is, in general, complex and *not* equal to $(\mathbf{u}, \bar{\mathbf{v}})$. The square of the *absolute length* of a vector $\mathbf{u}$ with complex components is defined to be the *real* quantity

$$l^2 = \mathbf{u}^2 = (\bar{\mathbf{u}}, \mathbf{u}) = \bar{\mathbf{u}}^T \mathbf{u} = \bar{u}_1 u_1 + \bar{u}_2 u_2 + \cdots + \bar{u}_n u_n. \tag{63b}$$

In case the elements involved are real, they are equal to their complex conjugates, and it is seen that (63a,b) reduce to (62a,b), as would be required for consistency.

A set of m vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ is said to be *linearly independent* if *no* set of constants $c_1, c_2, \ldots, c_m$ (at least one of which is not zero) exists such that

$$c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \cdots + c_m \mathbf{v}_m = \mathbf{0}. \tag{64}$$

In two-dimensional space, the existence of c_1 and c_2 such that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \mathbf{0}$$

would imply that the two-dimensional vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are scalar multiples of each other. Hence any two vectors which are *not multiples of each other* (parallel to a line) are linearly independent in two-dimensional space. Further, geometrical considerations indicate that any *three* vectors which are not *parallel to a plane* are linearly independent in three-dimensional space.

To obtain an analytical criterion for linear dependence of a set of vectors with *real* components, we suppose that c's *do* exist, at least one of which is not zero, such that (64) is satisfied. Then, by successively forming the scalar products of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ into both sides of (64), we find that the constants c_i must also satisfy the equations

$$\begin{aligned} c_1\mathbf{v}_1^2 + c_2(\mathbf{v}_1, \mathbf{v}_2) + \cdots + c_m(\mathbf{v}_1, \mathbf{v}_m) &= 0, \\ c_1(\mathbf{v}_2, \mathbf{v}_1) + c_2\mathbf{v}_2^2 + \cdots + c_m(\mathbf{v}_2, \mathbf{v}_m) &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ c_1(\mathbf{v}_m, \mathbf{v}_1) + c_2(\mathbf{v}_m, \mathbf{v}_2) + \cdots + c_m\mathbf{v}_m^2 &= 0. \end{aligned}$$

These conditions clearly require merely that the left-hand member of (64) be simultaneously orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$. But, according to Cramer's rule, this set of m equations in the m constants c_i cannot possess a nontrivial solution unless the determinant of the matrix of coefficients vanishes:

$$G \equiv \begin{vmatrix} \mathbf{v}_1^2 & (\mathbf{v}_1, \mathbf{v}_2) & \cdots & (\mathbf{v}_1, \mathbf{v}_m) \\ (\mathbf{v}_2, \mathbf{v}_1) & \mathbf{v}_2^2 & \cdots & (\mathbf{v}_2, \mathbf{v}_m) \\ \cdots & \cdots & \cdots & \cdots \\ (\mathbf{v}_m, \mathbf{v}_1) & (\mathbf{v}_m, \mathbf{v}_2) & \cdots & \mathbf{v}_m^2 \end{vmatrix} = 0. \qquad (65)$$

This determinant is called the Gram determinant or *Gramian* of $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Thus, if the vectors are linearly dependent the Gramian must vanish. The converse can also be shown to be true (see Problem 23). Hence it follows that *a set of vectors is linearly dependent if and only if its Gramian vanishes.**

* For a vector with *complex* components, this theorem is still true if the scalar products in the definition of the Gramian are replaced by *Hermitian* scalar products.

The set of all vectors $\mathbf{v}$ which can be expressed in the form

$$\mathbf{v} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_m\mathbf{a}_m, \tag{66}$$

where the $\mathbf{a}$'s are vectors, is called the *vector space* generated by the $\mathbf{a}$'s. If r and only r of the $\mathbf{a}$'s are linearly independent, the set is said to be of *rank* r. When $r < m$, we see that $m - r$ of the $\mathbf{a}$'s can be expressed as linear combinations of the r independent $\mathbf{a}$'s, and (66) can accordingly be expressed equivalently as a combination of only r *linearly independent* vectors, so that only r independent constants of combination are indeed available in such a case.

In a space of n dimensions, *any* vector $\mathbf{v}$ can be generated by any vector set of rank n, in the form

$$\mathbf{v} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n, \tag{67}$$

where the n $\mathbf{a}$'s are therefore linearly independent. For, in the n equations which equate the n components of the two members of (67), the matrix of coefficients of the c's has the property that no column is a linear combination of the others, and hence the determinant of the coefficient matrix cannot vanish.

To determine the constants more directly, we may merely form the scalar product of each $\mathbf{a}$ into the equal members of (67). The resultant set of n scalar equations can always be solved for the c's, since the determinant of the relevant coefficient matrix is the Gramian of the $\mathbf{a}$'s, and hence does not vanish. In particular, it follows that if $\mathbf{v}$ is orthogonal to all the $\mathbf{a}$'s, then $\mathbf{v}$ must be the *zero* vector.

We say that such a set of n linearly independent vectors *spans* the space, and that a set of rank r is of *defect* (or *nullity*) $d = n - r$ in a space of n dimensions.

A set of n linearly independent vectors in n-space is called a *basis* in that space. Thus, a basis is a set of vectors which spans the space, so that any vector in that space can be expressed as a linear combination of the members of a basis. Clearly, any set of n nonzero *mutually orthogonal* vectors is a basis in n-space, since its Gramian is the determinant of a matrix with zeros in all positions outside the principal diagonal and positive numbers in that diagonal, and hence cannot vanish. An especially convenient basis consists of the particular orthogonal *unit vectors*

$$\mathbf{i}_1 = \{1, 0, 0, \cdots, 0\}, \quad \mathbf{i}_2 = \{0, 1, 0, \cdots, 0\}, \quad \cdots,$$
$$\mathbf{i}_n = \{0, 0, 0, \cdots, 1\}. \tag{68}$$

1.10. Linear equations and vector space. We now apply the preceding considerations to a set of m homogeneous linear equations in n variables, of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned}\right\}. \tag{69}$$

Each of these equations can be interpreted as requiring that the vector $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ be orthogonal to a *vector* $\mathbf{a}_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$; that is, the set of equations can be written in the form

$$(\mathbf{a}_i, \mathbf{x}) = 0 \qquad (i = 1, 2, \cdots, m), \tag{70}$$

where

$$\mathbf{a}_1 = \{a_{11}, a_{12}, \cdots, a_{1n}\}, \quad \cdots, \quad \mathbf{a}_m = \{a_{m1}, a_{m2}, \cdots, a_{mn}\}. \tag{71}$$

Thus we may consider the successive elements of the ith *row* of the rectangular matrix

$$\mathbf{a} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

as comprising the components of the vector $\mathbf{a}_i$, and the matrix equation

$$\mathbf{a}\,\mathbf{x} = \mathbf{0} \tag{72}$$

corresponding to (69) then requires that $\mathbf{x}$ be orthogonal to *each* vector $\mathbf{a}_i$ in the vector space of n dimensions.

If the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ span that space, this situation is clearly impossible unless the vector $\mathbf{x}$ is a zero vector. Hence, in this case the only solution of (69) is the trivial solution $x_1 = x_2 = \ldots = x_n = 0$.

However, if the m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ form a set of rank $r < n$, it is possible to find $d = n - r$ linearly independent vectors, say

$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$, which are orthogonal to all the $\mathbf{a}$'s. Thus, any vector $\mathbf{x}$ which is a linear combination of these vectors,

$$\mathbf{x} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_d\mathbf{u}_d, \tag{73}$$

will satisfy the equation (72), and its components will satisfy (69).

The analogy between these results and the results of Section 1.8 suggests that *the rank of the vector set* $\mathbf{a}_1, \cdots, \mathbf{a}_m$ *is equal to the rank of the matrix* $[a_{ij}]$ *made up of the components of these vectors.* That this is indeed the case follows directly from the fact, established in Section 1.7, that if $[a_{ij}]$ is of rank r, then no linear combination of a certain set of r rows can vanish and, in addition, all other rows are linear combinations of these r rows.

In order to display the general solution of (51) or, equivalently, (52) in the form (73) when the right-hand members vanish, we may write $x_{r+1} = C_1, \; x_{r+2} = C_2, \ldots, x_n = C_{n-r}$, where the C's are arbitrary. The solution can then be written in the vector form

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_r \\ x_{r+1} \\ x_{r+2} \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = C_1 \begin{pmatrix} \alpha_{11} \\ \alpha_{21} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{r1} \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} + C_2 \begin{pmatrix} \alpha_{12} \\ \alpha_{22} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{r2} \\ 0 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} + \cdots + C_{n-r} \begin{pmatrix} \alpha_{1,n-r} \\ \alpha_{2,n-r} \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{r,n-r} \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}. \tag{73a}$$

It is clear from the form of the $n - r$ solution vectors that these vectors are indeed linearly independent.

In the more general case of a *nonhomogeneous* set of linear equations, of the form (51), the requirements are that the scalar products of $\mathbf{x}$ and the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ each take on prescribed values. The most general vector $\mathbf{x}$ having this property is expressible as the sum of any *particular* vector having this property (if such exist) and an arbitrary linear combination of all vectors which are *orthogonal* to all the $\mathbf{a}$'s (if such exist).

In the frequently occurring case when $m = n$, so that we have n equations in n unknowns, a further interpretation is useful. Here, corresponding to a given set of equations $\mathbf{a}\,\mathbf{x} = \mathbf{c}$, we can consider also the *transposed homogeneous set* $\mathbf{a}^T\,\mathbf{x}' = \mathbf{0}$, in which the successive components of the vector $\mathbf{a}_i$ now become the coefficients of x'_i in successive equations:

$$\left.\begin{array}{l} a_{11}x'_1 + a_{21}x'_2 + \cdots + a_{n1}x'_n = 0, \\ a_{12}x'_1 + a_{22}x'_2 + \cdots + a_{n2}x'_n = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{1n}x'_1 + a_{2n}x'_2 + \cdots + a_{nn}x'_n = 0 \end{array}\right\}. \tag{74}$$

If we consider the coefficients in the ith row of (74) as the components of a vector $\mathbf{a}'_i$, this set of equations takes the form

$$\mathbf{a}^T\,\mathbf{x}' = \mathbf{0}: \qquad (\mathbf{a}'_i, \mathbf{x}') = 0 \qquad (i = 1, 2, \cdots, n). \tag{74a}$$

But also, since the elements of the vectors $\mathbf{a}'_i$ comprise the *columns* of $\mathbf{a}$, the *original* set of equations $\mathbf{a}\,\mathbf{x} = \mathbf{c}$ can be written in the form

$$x_1 \begin{Bmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1} \end{Bmatrix} + x_2 \begin{Bmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{n2} \end{Bmatrix} + \cdots + x_n \begin{Bmatrix} a_{1n} \\ a_{2n} \\ \cdot \\ \cdot \\ \cdot \\ a_{nn} \end{Bmatrix} = \begin{Bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{Bmatrix}$$

or

$$\mathbf{a}\,\mathbf{x} = \mathbf{c}: \qquad x_1\mathbf{a}'_1 + x_2\mathbf{a}'_2 + \cdots + x_n\mathbf{a}'_n = \mathbf{c}. \tag{75}$$

Thus (75) possesses a solution if and only if $\mathbf{c}$ is a *linear combination* of the vectors $\mathbf{a}'_i$. On the other hand, (74a) states that all these vectors are orthogonal to all solutions of (74). Hence we obtain the following useful result:

The nonhomogeneous set $\mathbf{a}\,\mathbf{x} = \mathbf{c}$, *of n equations in n unknowns, possesses a solution if and only if* $\mathbf{c}$ *is orthogonal to all vector solutions of the transposed homogeneous set* $\mathbf{a}^T\,\mathbf{x}' = \mathbf{0}$.

1.11. Characteristic-value problems. Of frequent occurrence in many fields is the problem of determining those values of a constant λ for which nontrivial solutions exist to the homogeneous set of equations

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= \lambda x_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= \lambda x_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= \lambda x_n \end{aligned}\right\}. \tag{76}$$

Such a problem is known as a characteristic-value problem; values of λ for which nontrivial solutions exist are called *characteristic values* (also *eigenvalues* or *latent roots*) of the problem or *of the matrix* $\mathbf{a}$, and corresponding vector solutions are known as the *characteristic vectors* (also *eigenvectors*) of the problem or *of the matrix* $\mathbf{a}$. A column made up of the elements of a characteristic vector is often called a *modal column*.

In most practical considerations in which such problems arise, the matrix $\mathbf{a}$ is *symmetric;* that is, two elements which are symmetrically placed with respect to the principal diagonal are equal:

$$a_{ji} = a_{ij}. \tag{77}$$

More generally, when the coefficients are *complex* the most important cases are those in which symmetrically situated elements are *complex conjugates:*

$$a_{ji} = \bar{a}_{ij}. \tag{77a}$$

Matrices having the symmetry property (77a) are known as *Hermitian matrices,* and are considered in Section 1.16.

The discussion of the present section is to be restricted to *real symmetric matrices,* for which the symmetry property (77) applies.

In matrix notation, equation (76) takes the form

$$\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x} \qquad \text{or} \qquad (\mathbf{a} - \lambda\,\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{78}$$

where $\mathbf{I}$ is the unit matrix of order n. This homogeneous problem possesses nontrivial solutions if and only if the determinant of the coefficient matrix $[\mathbf{a} - \lambda\,\mathbf{I}]$ vanishes:

$$|\,\mathbf{a} - \lambda\,\mathbf{I}\,| \equiv \begin{vmatrix} (a_{11} - \lambda) & a_{12} & \cdots & a_{1n} \\ a_{21} & (a_{22} - \lambda) & \cdots & a_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ a_{n1} & a_{n2} & \cdots & (a_{nn} - \lambda) \end{vmatrix} = 0. \tag{79}$$

This condition requires that λ be a root of an algebraic equation of degree n, known as the *characteristic* (or *secular*) *equation*. The n solutions $\lambda_1, \lambda_2, \ldots, \lambda_n$, which need not all be distinct, are the characteristic numbers or latent roots of the matrix $\mathbf{a}$.

Corresponding to each such value λ_k, there exists at least one vector solution (modal column) of (76) or (78), which is determined within an arbitrary multiplicative constant.* Now let λ_1 and λ_2 be two *distinct* characteristic numbers and denote corresponding characteristic vectors by $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, so that the equations

$$\mathbf{a}\,\mathbf{x}_1 = \lambda_1 \mathbf{x}_1, \qquad \mathbf{a}\,\mathbf{x}_2 = \lambda_2 \mathbf{x}_2 \qquad (\lambda_1 \neq \lambda_2) \tag{80a,b}$$

are satisfied. If we postmultiply the transpose of (80a) by $\mathbf{x}_2$ there follows

$$(\mathbf{a}\,\mathbf{x}_1)^T \mathbf{x}_2 = \lambda_1 \mathbf{x}_1{}^T \mathbf{x}_2$$

or, using (34),

$$\mathbf{x}_1{}^T \mathbf{a}^T \mathbf{x}_2 = \lambda_1 \mathbf{x}_1{}^T \mathbf{x}_2. \tag{81a}$$

Also, by premultiplying (80b) by $\mathbf{x}_1{}^T$, we obtain

$$\mathbf{x}_1{}^T \mathbf{a}\,\mathbf{x}_2 = \lambda_2 \mathbf{x}_1{}^T \mathbf{x}_2. \tag{81b}$$

The result of subtracting (81a) from (81b), and noticing that for a *symmetric* matrix $\mathbf{a}^T = \mathbf{a}$, is then the relation

$$(\lambda_2 - \lambda_1)(\mathbf{x}_1, \mathbf{x}_2) = 0, \tag{81c}$$

and, since we have assumed that $\lambda_1 \neq \lambda_2$, it follows that $(\mathbf{x}_1, \mathbf{x}_2) = 0$. Hence we have the following important result:

Two characteristic vectors of a real symmetric matrix, corresponding to different characteristic numbers, are orthogonal,

$$(\mathbf{x}_1, \mathbf{x}_2) = 0. \tag{82}$$

A second basic result is that the characteristic numbers of such a matrix are always *real*. To establish this fact, we suppose that $\lambda_1 = \alpha + i\,\beta$ is a root of (79), where α and β are real. Then, since the coefficients of (79) are *real*, $\lambda_2 = \alpha - i\,\beta = \bar{\lambda}_1$ must also be a root. The elements of the two corresponding characteristic vectors must then also be conjugate complex quantities. Thus, if we denote

* As was shown in Section 1.8, the components of this solution vector can be expressed as arbitrary multiples of the cofactors of the elements in a row of the matrix $\mathbf{a} - \lambda_k \mathbf{I}$ unless all those cofactors vanish.

these vectors by $\mathbf{x}_1$ and $\bar{\mathbf{x}}_1$, respectively (see page 24), we have the two relations

$$\mathbf{a}\,\mathbf{x}_1 = \lambda_1 \mathbf{x}_1, \qquad \mathbf{a}\,\bar{\mathbf{x}}_1 = \bar{\lambda}_1 \bar{\mathbf{x}}_1. \tag{83a,b}$$

By premultiplying (83a) by $\bar{\mathbf{x}}_1'^T$ and postmultiplying the transpose of (83b) by $\mathbf{x}_1$, and subtracting, we obtain the relation

$$(\lambda_1 - \bar{\lambda}_1)(\bar{\mathbf{x}}_1, \mathbf{x}_1) = \bar{\mathbf{x}}_1{}^T \mathbf{a}\,\mathbf{x}_1 - (\mathbf{a}\,\bar{\mathbf{x}}_1)^T \mathbf{x}_1 = 0. \tag{84}$$

But now, since the product $(\bar{\mathbf{x}}_1, \mathbf{x}_1)$ is a *positive* quantity [see equation (63)], it follows that $\lambda_1 - \bar{\lambda}_1 = 2i\beta$ must vanish, so that λ_1 must be real. Thus we conclude that *all characteristic numbers of a real symmetric matrix are real.*

If a characteristic number, say λ_1, *of a symmetric matrix is a multiple root of multiplicity* s, that is, if the left-hand member of (79) possesses the factor $(\lambda - \lambda_1)^s$, *then to* λ_1 *there correspond* s *linearly independent characteristic vectors.* Proof of this important fact is postponed until Section 1.21.

The preceding statement is *not* necessarily true for *nonsymmetric* matrices, as can be seen by considering the equations

$$\left.\begin{aligned} x_1 + x_2 &= \lambda x_1, \\ -x_1 - x_2 &= \lambda x_2 \end{aligned}\right\},$$

for which

$$\mathbf{a} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}.$$

Here the characteristic equation is readily found to be merely $\lambda^2 = 0$, so that $\lambda = 0$ is a characteristic number of multiplicity *two*. However, when $\lambda = 0$, the only possible solution is given by $x_1 = C_1$, $x_2 = -C_1$ or $\mathbf{x} = C_1\{1, -1\}$. Thus, here the double root $\lambda = 0$ corresponds to only *one* characteristic vector.

As is shown in the following section, it is always possible to choose the s linearly independent vectors corresponding to a characteristic number of multiplicity s in such a way that they are orthogonal *to each other*, in addition to being (automatically) orthogonal to all other characteristic vectors. Thus, if multiple roots of (79) are counted separately, we obtain always exactly n characteristic numbers, and we can determine a corresponding set of n mutually orthogonal characteristic vectors. In virtue of the results of Section 1.9, this set of vectors comprises a *basis* in n-dimensional vector

space; that is, *any vector in n-dimensional space can be expressed as some linear combination of these n vectors.*

Consider now the *nonhomogeneous* equation

$$\mathbf{a}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{c}, \tag{85}$$

where $\mathbf{a}$ is a real symmetric matrix. This equation reduces to (76) or (78) when $\mathbf{c} = \mathbf{0}$. If (85) has a solution, then that solution can be expressed as a linear combination of the characteristic vectors of $\mathbf{a}$. Suppose that n orthogonal characteristic vectors are known, and that they have each been divided by their lengths and so are *unit vectors.* If these vectors are denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, it follows that they satisfy the respective equations

$$\mathbf{a}\,\mathbf{e}_1 = \lambda_1\mathbf{e}_1, \qquad \cdots, \qquad \mathbf{a}\,\mathbf{e}_n = \lambda_n\mathbf{e}_n. \tag{86}$$

The solution of (85) can then be assumed in the form

$$\mathbf{x} = \sum_{k=1}^{n} \alpha_k\mathbf{e}_k, \tag{87}$$

where the constants α_k are to be determined. The introduction of (87) into (85), and the use of (86), then leads to the requirement

$$\sum_{k=1}^{n} (\lambda_k - \lambda)\alpha_k\mathbf{e}_k = \mathbf{c}. \tag{88}$$

From this equation, the α's are then determined by forming the scalar product of any $\mathbf{e}_i$ into both sides of (88). Remembering that

$$(\mathbf{e}_i, \mathbf{e}_k) = \delta_{ik},$$

we see that the ith coefficient α_i must then satisfy the equation

$$(\lambda_i - \lambda)\alpha_i = (\mathbf{e}_i, \mathbf{c}) \qquad (i = 1, 2, \cdots, n). \tag{88a}$$

Hence, *if λ is not a characteristic number,* the solution (87) is obtained in the form

$$\mathbf{x} = \sum_{k=1}^{n} \frac{(\mathbf{e}_k, \mathbf{c})}{\lambda_k - \lambda}\,\mathbf{e}_k. \tag{89}$$

Thus a unique solution of the nonhomogeneous problem is obtained when λ is not a characteristic number. If $\lambda = \lambda_p$, *no solution exists unless the vector* $\mathbf{c}$ *is orthogonal to the characteristic vector (or vectors) corresponding to* λ_p. In case this condition is

satisfied, equation (88a) shows that the corresponding coefficient (or coefficients) α_p may be chosen *arbitrarily*, so that *infinitely many* solutions then exist.

In particular, if $\lambda = 0$, equation (85) reduces to the equation

$$\mathbf{a}\,\mathbf{x} = \mathbf{c},$$

which was studied previously. This equation thus has a unique solution unless $\lambda = 0$ is a characteristic number of $\mathbf{a}$, that is, unless the equation $\mathbf{a}\,\mathbf{x} = \mathbf{0}$ has nontrivial solutions. In this exceptional case no solution exists unless $\mathbf{c}$ is orthogonal to the vectors which satisfy $\mathbf{a}\,\mathbf{x} = \mathbf{0}$, in which case infinitely many solutions exist. This result is in accordance with the results of the preceding section, where it was shown that the requirement for the existence of a solution in the exceptional case is that $\mathbf{c}$ be orthogonal to the vectors which satisfy the equation $\mathbf{a}^r\,\mathbf{x} = \mathbf{0}$, since in the present case we have considered only a symmetric matrix, for which $\mathbf{a}^r = \mathbf{a}$.

The existence criterion obtained here, in the *more general* case when $\lambda = \lambda_p$, is also obtainable from the last result of the preceding section, by replacing $\mathbf{a}$ by $\mathbf{a} - \lambda_p\mathbf{I}$ in that result, and noticing that the latter matrix is symmetric when $\mathbf{a}$ is symmetric.

1.12. Orthogonalization of vector sets. It is often desirable, as in the preceding section, to form from a set of s linearly independent vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_s$ an *orthogonal* set of s linear combinations of the original vectors. It is also convenient to "normalize" the vectors in such a way that each is a *unit vector*. The following procedure is a simple one, and it can be extended by analogy to other similar problems.

We first select *any one* of the original vectors, say $\mathbf{v}_1 = \mathbf{u}_1$, and divide it by its length l_1. This is the first member of the desired set:

$$\mathbf{e}_1 = \frac{\mathbf{u}_1}{l_1}. \tag{90a}$$

We next choose a second vector, say $\mathbf{u}_2$, from the original set and write $\mathbf{v}_2 = \mathbf{u}_2 - c\,\mathbf{e}_1$. The requirement that $\mathbf{v}_2$ be orthogonal to $\mathbf{e}_1$ leads to the determination

$$(\mathbf{e}_1, \mathbf{v}_2) = (\mathbf{e}_1, \mathbf{u}_2) - c(\mathbf{e}_1, \mathbf{e}_1) = 0$$

or

$$c = (\mathbf{e}_1, \mathbf{u}_2),$$

so that

$$\mathbf{v}_2 = \mathbf{u}_2 - (\mathbf{e}_1, \mathbf{u}_2)\mathbf{e}_1. \tag{90b}$$

Since $\mathbf{e}_1$ is a unit vector, the familiar geometrical interpretation of the scalar product in two or three dimensions leads us to say that $(\mathbf{e}_1, \mathbf{u}_2)$ is "the *scalar component* of $\mathbf{u}_2$ in the direction of $\mathbf{e}_1$," and hence that in (90b) we have "subtracted off the $\mathbf{e}_1$-component of $\mathbf{u}_2$."

The second member, $\mathbf{e}_2$, of the desired set of orthogonal unit vectors is obtained by dividing $\mathbf{v}_2$ by its length l_2:

$$\mathbf{e}_2 = \frac{\mathbf{v}_2}{l_2}. \tag{90c}$$

In the third step we write $\mathbf{v}_3 = \mathbf{u}_3 - c_1\mathbf{e}_1 - c_2\mathbf{e}_2$. The requirement that $\mathbf{v}_3$ be simultaneously orthogonal to $\mathbf{e}_1$ and $\mathbf{e}_2$ then determines values of c_1 and c_2 which are in accordance with the geometrical interpretation described above, and there follows

$$\mathbf{v}_3 = \mathbf{u}_3 - (\mathbf{e}_1, \mathbf{u}_3)\mathbf{e}_1 - (\mathbf{e}_2, \mathbf{u}_3)\mathbf{e}_2, \tag{90d}$$

so that the "$\mathbf{e}_1$- and $\mathbf{e}_2$-components" of $\mathbf{u}_3$ are subtracted off. The third required vector $\mathbf{e}_3$ is then given by

$$\mathbf{e}_3 = \frac{\mathbf{v}_3}{l_3} \tag{90e}$$

where l_3 is the length of $\mathbf{v}_3$.

A continuation of this process finally determines the sth member of the required set in the form

$$\mathbf{e}_s = \frac{\mathbf{v}_s}{l_s} \quad \text{where} \quad \mathbf{v}_s = \mathbf{u}_s - \sum_{k=1}^{s-1} (\mathbf{e}_k, \mathbf{u}_s)\mathbf{e}_k. \tag{91}$$

This method, which is often called the *Schmidt orthogonalization procedure*, would fail if and only if at some stage $\mathbf{v}_r = \mathbf{0}$. But this would mean that $\mathbf{u}_r$ is a linear combination of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{r-1}$, and hence also a linear combination of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{r-1}$, in contradiction with the statement that the set of u's is linearly independent.

1.13. Quadratic forms. A homogeneous expression of second degree, of the form

$$F \equiv a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \cdots + 2a_{n-1,n}x_{n-1}x_n, \tag{92}$$

is called a *quadratic form* in $x_1, x_2, \cdots, x_n$. In two-dimensional space the equation $F =$ constant represents a general second-degree curve (conic) with center at the origin, while in three-dimensional space the equation $F =$ constant represents a general *quadric surface* with center at the origin. Many problems associated with such forms are intimately related to problems associated with sets of linear equations.

We may notice first that if we write

$$y_i = \frac{1}{2}\frac{\partial F}{\partial x_i} \qquad (i = 1, 2, \cdots, n),$$

we obtain the equations

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1, \\ a_{12}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{1n}x_1 + a_{2n}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned}\right\}. \tag{93}$$

This set of equations can be written in the form

$$\mathbf{a}\,\mathbf{x} = \mathbf{y}, \tag{94}$$

where $\mathbf{a} = [a_{ij}]$ is a *symmetric* matrix. That is, the elements satisfy the symmetry condition

$$a_{ji} = a_{ij}. \tag{95}$$

On the other hand, it is easily seen that (92) is equivalent to the relation $F \equiv (\mathbf{x}, \mathbf{y})$; that is, (92) can be written in the form

$$F \equiv \mathbf{x}^T\,\mathbf{a}\,\mathbf{x}. \tag{96}$$

In many cases it is desirable to express $x_1, x_2, \cdots, x_n$ as linear combinations of new coordinates $x_1', x_2', \cdots, x_n'$ in such a way that F is expressed as a linear combination of only *squares* of the new coordinates, the cross-product terms being eliminated. A form of this type is said to be a *canonical form*. Let the vector $\mathbf{x}$ be expressed in terms of $\mathbf{x}'$ by the equation

$$\mathbf{x} = \mathbf{Q}\,\mathbf{x}', \tag{97}$$

where $\mathbf{Q}$ is a square matrix of order n. The introduction of (97) into (96) then gives

$$F = (\mathbf{Q}\,\mathbf{x}')^T\,\mathbf{a}\,\mathbf{Q}\,\mathbf{x}' = \mathbf{x}'^T\,\mathbf{Q}^T\,\mathbf{a}\,\mathbf{Q}\,\mathbf{x}' \tag{98}$$

or

$$F = \mathbf{x}'^T\,\mathbf{a}'\,\mathbf{x}', \tag{99}$$

where the new matrix $\mathbf{a}'$ is defined by the equation

$$\mathbf{a}' = \mathbf{Q}^T\,\mathbf{a}\,\mathbf{Q}. \tag{100}$$

Thus we see that, if F is to involve only squares of the variables x_i', the matrix $\mathbf{Q}$ in (97) must be so chosen that $\mathbf{Q}^T\,\mathbf{a}\,\mathbf{Q}$ is a *diagonal matrix;* that is, so that all elements for which $i \neq j$ vanish.

We show next that if the characteristic numbers and corresponding characteristic vectors of the *symmetric matrix* $\mathbf{a}$ are known, a matrix $\mathbf{Q}$ having this property can be very easily constructed. Suppose that the characteristic numbers of $\mathbf{a}$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$, repeated roots of the characteristic equation being numbered separately, and denote the corresponding members of the orthogonalized set of n characteristic unit vectors by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. We then have the relations

$$\mathbf{a}\,\mathbf{e}_1 = \lambda_1\mathbf{e}_1, \quad \cdots, \quad \mathbf{a}\,\mathbf{e}_n = \lambda_n\mathbf{e}_n. \tag{101}$$

Let a matrix $\mathbf{Q}$ be constructed in such a way that the elements of the unit vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are the elements of the successive *columns* of $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{n1} \\ e_{12} & e_{22} & \cdots & e_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ e_{1n} & e_{2n} & \cdots & e_{nn} \end{bmatrix}. \tag{102}$$

Then, if use is made of (101), it is easily seen that

$$\mathbf{a}\,\mathbf{Q} = \begin{bmatrix} \lambda_1 e_{11} & \lambda_2 e_{21} & \cdots & \lambda_n e_{n1} \\ \lambda_1 e_{12} & \lambda_2 e_{22} & \cdots & \lambda_n e_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_1 e_{1n} & \lambda_2 e_{2n} & \cdots & \lambda_n e_{nn} \end{bmatrix} \tag{103a}$$

or

$$\mathbf{a}\,\mathbf{Q} = \mathbf{Q} \cdot \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}. \tag{103b}$$

This relation follows directly from the fact that the product of $\mathbf{a}$ into the kth column of $\mathbf{Q}$ is the kth column of the right-hand member of (103a) [see Problem 24(a)]. Since the vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are linearly independent, it follows that $|\mathbf{Q}| \neq 0$. Thus the inverse $\mathbf{Q}^{-1}$ exists, and by premultiplying the equal members of (103) by $\mathbf{Q}^{-1}$ we obtain the result

$$\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q} = [\lambda_i\,\delta_{ij}]. \tag{104}$$

Hence, the matrix $\mathbf{a}$ is diagonalized by the indicated operations, the diagonal elements being merely the characteristic numbers of $\mathbf{a}$.

However, the *desired* diagonalization (100) was to be of the form $\mathbf{Q}^T\,\mathbf{a}\,\mathbf{Q}$. Thus, the matrix $\mathbf{Q}$ defined by (102) is not acceptable for present purposes unless it can be shown that

$$\mathbf{Q}^T = \mathbf{Q}^{-1} \qquad \text{or} \qquad \mathbf{Q}^T\,\mathbf{Q} = \mathbf{I}. \tag{105}$$

But the typical term q_{ij} of the product $\mathbf{Q}^T\,\mathbf{Q}$ is of the form

$$q_{ij} = \sum_{k=1}^{n} e_{ik}e_{jk},$$

and since the $\mathbf{e}$'s are *orthogonal* the indicated sum *vanishes* unless $i = j$, in which case the sum is *unity* since the $\mathbf{e}$'s are *unit vectors*.

Hence there follows $q_{ij} = \delta_{ij}$; that is, $\mathbf{Q}^T\,\mathbf{Q} = [\delta_{ij}] = \mathbf{I}$, as is required by (105). Further, since $|\mathbf{Q}| = |\mathbf{Q}^T|$, there follows from (105) the useful result

$$|\mathbf{Q}|^2 = 1: \qquad |\mathbf{Q}| = \pm 1. \tag{106}$$

It follows that the matrix $\mathbf{Q}$ defined by (102) does indeed have the property that *the quadratic form*

$$F = \mathbf{x}^T\,\mathbf{a}\,\mathbf{x} \tag{107}$$

is reduced by the change in variables

$$\mathbf{x} = \mathbf{Q}\,\mathbf{x}' \tag{108}$$

to the form

$$F = \mathbf{x}'^T\,\mathbf{a}'\,\mathbf{x}'$$

where $\mathbf{a}' = [\lambda_i\,\delta_{ij}]$, *that is, to the form*

$$F = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \cdots + \lambda_n x_n'^2, \tag{109}$$

where the numbers λ_i *are the characteristic numbers of* $\mathbf{a}$.

A matrix whose columns comprise the elements of n linearly independent characteristic vectors of a given matrix $\mathbf{a}$, of order n, is called a *modal matrix* of $\mathbf{a}$. In particular, when those n vectors are mutually orthogonal and of unit length, it is convenient to say that the modal matrix is *normalized*. Thus the matrix $\mathbf{Q}$ is a *normalized modal matrix* of $\mathbf{a}$.

We notice that if $\lambda = 0$ is a p-fold root of the characteristic equation the form (109) has only $n - p$ nonvanishing terms. It is shown in Section 1.21 that this situation arises if and only if the symmetric matrix $\mathbf{a}$ is of *rank* $r = n - p$.

The new variables x_i' are related to the original ones, in accordance with (97) and (105), by the equation

$$\mathbf{x}' = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^T\mathbf{x}, \tag{110}$$

and hence are of the form

$$\left.\begin{aligned} x_1' &= e_{11}x_1 + e_{12}x_2 + \cdots + e_{1n}x_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n' &= e_{n1}x_1 + e_{n2}x_2 + \cdots + e_{nn}x_n \end{aligned}\right\} \tag{111}$$

or

$$x_i' = (\mathbf{e}_i, \mathbf{x}) \qquad (i = 1, 2, \cdots, n). \tag{111a}$$

When all the characteristic numbers of $\mathbf{a}$ are *distinct*, the normalized modal matrix $\mathbf{Q}$ is uniquely determined except for the ordering of the columns. However, if a root is of multiplicity p, the corresponding p orthogonalized unit vectors can be chosen in infinitely many ways, as was shown in Section 1.12.

We remark that the modal matrix $\mathbf{Q}$ specified by (102) is not the *only* matrix which can be used in (108) to reduce a quadratic form to a sum of squares. However, it is the only such matrix which possesses the useful property that $\mathbf{Q}^T = \mathbf{Q}^{-1}$. A matrix having this property is called an *orthogonal matrix*.

1.14. A numerical example. To illustrate the preceding reduction in a specific numerical case, we consider the quadratic form

$$F = 25x_1^2 + 34x_2^2 + 41x_3^2 - 24x_2x_3.$$

The corresponding matrix $\mathbf{a}$ is then of the form

$$\mathbf{a} = \begin{bmatrix} 25 & 0 & 0 \\ 0 & 34 & -12 \\ 0 & -12 & 41 \end{bmatrix}$$

and the equations $\mathbf{a}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{0}$ become

$$\begin{aligned} (25 - \lambda)x_1 \qquad\qquad\quad &= 0, \\ (34 - \lambda)x_2 \quad - 12x_3 &= 0, \\ -12x_2 + (41 - \lambda)x_3 &= 0. \end{aligned}$$

The characteristic equation $|\,\mathbf{a} - \lambda\,\mathbf{I}\,| = 0$ then takes the form

$$(25 - \lambda)(\lambda^2 - 75\lambda + 1250) = 0,$$

from which the characteristic numbers are

$$\lambda_1 = \lambda_2 = 25, \qquad \lambda_3 = 50.$$

When $\lambda = \lambda_1 = \lambda_2 = 25$, the equations $\mathbf{a}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{0}$ become

$$\begin{aligned} 0 &= 0, \\ 9x_2 - 12x_3 &= 0, \\ -12x_2 + 16x_3 &= 0, \end{aligned}$$

with the general solution $x_1 = C_1$, $x_2 = C_2$, $x_3 = \frac{3}{4}C_2$. In vector form we may write $\mathbf{x} = C_1\mathbf{u}_1 + C_2\mathbf{u}_2$, where $\mathbf{u}_1 = \{1, 0, 0\}$ and $\mathbf{u}_2 = \{0, 1, \frac{3}{4}\}$. Since it happens that $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal, we need only divide them by their lengths $l_1 = 1$ and $l_2 = \frac{5}{4}$ to obtain the two orthogonal unit characteristic vectors

$$\mathbf{e}_1 = \{1, 0, 0\}, \qquad \mathbf{e}_2 = \{0, \tfrac{4}{5}, \tfrac{3}{5}\}.$$

In a similar way, a unit characteristic vector corresponding to $\lambda = \lambda_3 = 50$ is found to be

$$\mathbf{e}_3 = \{0, \tfrac{3}{5}, -\tfrac{4}{5}\}.$$

Hence the normalized modal matrix $\mathbf{Q}$ of equation (102) can be taken in the form

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{4}{5} & \frac{3}{5} \\ 0 & \frac{3}{5} & -\frac{4}{5} \end{bmatrix}$$

and the new coordinates defined by (111) are then given by

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= \tfrac{4}{5}x_2 + \tfrac{3}{5}x_3, \\ x_3' &= \tfrac{3}{5}x_2 - \tfrac{4}{5}x_3. \end{aligned}$$

With this choice of the new coordinates, (109) states that the quadratic form under consideration takes the form

$$F \equiv 25x_1'^2 + 25x_2'^2 + 50x_3'^2.$$

In particular, it follows that the quadric surface with the equation $25x_1^2 + 34x_2^2 + 41x_3^2 - 24x_2x_3 = 25$ takes the standard form $x_1'^2 + x_2'^2 + 2x_3'^2 = 1$ with the introduction of the new coordinates. It is shown in Section 1.19 that *the new x'y'z'-coordinate system defined by* (108) *is also a rectangular system when* $\mathbf{Q}$ *is an orthogonal matrix* and that *length and angle are preserved by the transformation.* Hence the quadric surface just considered is an *oblate spheroid* with semiaxes of length 1, 1, $\sqrt{2}/2$.

It may be noticed that, by the usual method of "completing squares," we may, for example, also reduce the form F as follows:

$$\begin{aligned} F &= 25x_1^2 + 34[x_2^2 - \tfrac{24}{34}x_2x_3 + (\tfrac{12}{34})^2x_3^2] + (41 - \tfrac{144}{34})x_3^2 \\ &= 25x_1^2 + 34(x_2 - \tfrac{6}{17}x_3)^2 + \tfrac{625}{17}x_3^2. \end{aligned}$$

Hence, if we introduce new variables by the relations

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= x_2 - \tfrac{6}{17}x_3, \\ x_3' &= x_3, \end{aligned}$$

we can reduce F to the form

$$F = 25x_1'^2 + 34x_2'^2 + \tfrac{625}{17}x_3'^2.$$

However, here the matrix $\mathbf{Q}$ for which $\mathbf{x} = \mathbf{Q}\,\mathbf{x}'$, and which takes the *triangular* form

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{6}{17} \\ 0 & 0 & 1 \end{bmatrix},$$

is *not* an orthogonal matrix. Consequently, as is shown in Section

1.19, the new $x'y'z'$-coordinate system is *not* a rectangular system in this case; that is, the new coordinate axes are *not mutually perpendicular*. Nevertheless, the matrix $\mathbf{Q}$ does have the property that $\mathbf{Q}^T \mathbf{a} \mathbf{Q}$ is a diagonal matrix.

1.15. Equivalent matrices and transformations. Two matrices $\mathbf{a}$ and $\mathbf{b}$ which can be obtained from each other by a finite number of successive applications of the *elementary operations* (Section 1.7) to rows and/or columns are said to be *equivalent* (but not necessarily *equal*) *matrices*.

It can be shown that any such sequence of operations on the *rows* of $\mathbf{a}$ can be effected by *pre*multiplying $\mathbf{a}$ by some nonsingular matrix $\mathbf{P}$, while corresponding operations on *columns* can always be effected by *post*multiplying $\mathbf{a}$ by a nonsingular matrix $\mathbf{Q}$. This result is a consequence of the easily established fact that an elementary operation on rows (columns) of $\mathbf{a}$ may be accomplished by first performing that operation on the *unit* matrix $\mathbf{I}$ of the same order, and then premultiplying (postmultiplying) $\mathbf{a}$ by the resultant matrix (see Problems 18 and 19).

The converse of the preceding statement is also true; that is, *the matrices* $\mathbf{a}$ *and* $\mathbf{b}$ *are equivalent if and only if nonsingular matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *exist such that* $\mathbf{b} = \mathbf{P} \mathbf{a} \mathbf{Q}$.

Since the elementary operations do not change the rank of a matrix, it follows that *two equivalent matrices have the same rank*.

Transformations of the form $\mathbf{P} \mathbf{a} \mathbf{Q}$ are classified according to restrictions imposed on $\mathbf{P}$ and $\mathbf{Q}$. Thus if $\mathbf{P} = \mathbf{Q}^T = \mathbf{Q}^{-1}$, as in the reduction of Section 1.13, the transformation is called an *orthogonal* transformation. If only $\mathbf{P} = \mathbf{Q}^T$, as is required by equation (100), the resulting transformation $\mathbf{Q}^T \mathbf{a} \mathbf{Q}$ is called a *congruence* transformation, whereas a transformation of the form $\mathbf{Q}^{-1} \mathbf{a} \mathbf{Q}$, for which $\mathbf{P} = \mathbf{Q}^{-1}$, is called a *similarity* transformation. This terminology is motivated by certain geometrical considerations. We notice that *an orthogonal transformation is both a congruence and a similarity transformation*.

Conjunctive and *unitary* transformations, which are of importance in dealing with matrices of *complex* elements, are defined in the following section.

1.16. Hermitian matrices. We now consider a matrix with *complex* elements which satisfy the relation

$$a_{ji} = \bar{a}_{ij}. \tag{112}$$

Such a matrix is hence of the special form

$$\mathbf{h} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \bar{a}_{12} & a_{22} & a_{23} & \cdots & a_{2n} \\ \bar{a}_{13} & \bar{a}_{23} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \bar{a}_{1n} & \bar{a}_{2n} & \bar{a}_{3n} & \cdots & a_{nn} \end{bmatrix}, \tag{113}$$

and is known as a *Hermitian* matrix. Thus a Hermitian matrix has the property that two elements situated symmetrically with respect to the principal diagonal are *complex conjugates*. In particular, (112) requires that the elements in the principal diagonal ($i = j$) be *real*.

We see that the *complex conjugate* of the matrix $\mathbf{h}$, obtained by replacing each element by its complex conjugate, and denoted by $\bar{\mathbf{h}}$, is equal to the transpose of $\mathbf{h}$:

$$\mathbf{h}^T = \bar{\mathbf{h}}. \tag{114}$$

The product

$$H \equiv \bar{\mathbf{x}}^T \mathbf{h} \mathbf{x} \tag{115}$$

is known as a *Hermitian form*. In two dimensions, the general Hermitian form is thus given by

$$\begin{aligned} H &\equiv [\bar{x}_1 \quad \bar{x}_2] \begin{bmatrix} a_{11} & a_{12} \\ \bar{a}_{12} & a_{22} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} \\ &\equiv a_{11}\bar{x}_1 x_1 + (a_{12}\bar{x}_1 x_2 + \bar{a}_{12}\bar{x}_2 x_1) + a_{22}\bar{x}_2 x_2. \end{aligned} \tag{116}$$

Although the elements a_{ij} and variables x_i may be complex, *the values assumed by a Hermitian form are always real*. To establish this fact, we recall first that *the conjugate of a product of complex quantities is equal to the product of the conjugates*. Thus, if H were complex, and given by (115), then its conjugate $\bar{H}$ would be given by

$$\bar{H} = \mathbf{x}^T \bar{\mathbf{h}} \bar{\mathbf{x}} = \mathbf{x}^T \mathbf{h}^T \bar{\mathbf{x}} = (\mathbf{h}\mathbf{x})^T \bar{\mathbf{x}} = \bar{\mathbf{x}}^T (\mathbf{h}\mathbf{x}) = H. \tag{117}$$

But $\bar{H} = H$ only if H is real, as was to be shown.

Also, we can show that the characteristic numbers of a Hermitian matrix are real. For if $\mathbf{u}_1$ is a characteristic vector corresponding to λ_1, we must have

$$\mathbf{h}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1, \tag{118}$$

and hence also, after premultiplying both sides by $\bar{\mathbf{u}}_1^T$,

$$\bar{\mathbf{u}}_1^T \mathbf{h}\, \mathbf{u}_1 = \lambda_1 \bar{\mathbf{u}}_1^T \mathbf{u}_1. \tag{119}$$

But since $\bar{\mathbf{u}}_1^T \mathbf{h}\, \mathbf{u}_1$ and $\bar{\mathbf{u}}_1^T \mathbf{u}_1$ are both real, and $\bar{\mathbf{u}}_1^T \mathbf{u}_1 \neq 0$, λ_1 must also be real.

Further, let $\mathbf{u}_2$ be a characteristic vector corresponding to a second characteristic number $\lambda_2 \neq \lambda_1$, so that

$$\mathbf{h}\, \mathbf{u}_2 = \lambda_2 \mathbf{u}_2. \tag{120}$$

If the transposed conjugate of (118) is postmultiplied by $\mathbf{u}_2$, there follows

$$(\bar{\mathbf{h}}\, \bar{\mathbf{u}}_1)^T \mathbf{u}_2 = \lambda_1 \bar{\mathbf{u}}_1^T \mathbf{u}_2,$$

while premultiplication of (120) by $\bar{\mathbf{u}}_1^T$ leads to the relation

$$\bar{\mathbf{u}}_1^T \mathbf{h}\, \mathbf{u}_2 = \lambda_2 \bar{\mathbf{u}}_1^T \mathbf{u}_2.$$

By subtracting these equations from each other, and using equations (34) and (114), there follows

$$\begin{aligned}(\lambda_2 - \lambda_1)\bar{\mathbf{u}}_1^T \mathbf{u}_2 &= \bar{\mathbf{u}}_1^T \mathbf{h}\, \mathbf{u}_2 - (\bar{\mathbf{h}}\, \bar{\mathbf{u}}_1)^T \mathbf{u}_2 \\ &= \bar{\mathbf{u}}_1^T \mathbf{h}\, \mathbf{u}_2 - \bar{\mathbf{u}}_1^T \bar{\mathbf{h}}^T \mathbf{u}_2 \\ &= 0.\end{aligned}$$

Hence we conclude that *two characteristic vectors of a Hermitian matrix, corresponding to different characteristic numbers, are orthogonal in the Hermitian sense:*

$$(\bar{\mathbf{u}}_1, \mathbf{u}_2) \equiv \bar{\mathbf{u}}_1^T \mathbf{u}_2 = 0. \tag{121}$$

The vectors $\mathbf{u}_i$ can then be divided by their absolute lengths $l_i = \sqrt{(\bar{\mathbf{u}}_i, \mathbf{u}_i)}$, to give a set of orthogonal *unit* vectors $\mathbf{e}_i$ corresponding to successive nonrepeated roots of the characteristic equation. Corresponding to a root of multiplicity s there exists a set of s linearly independent characteristic vectors (see Section 1.21), which can be orthogonalized and reduced to absolute length unity, by a procedure completely analogous to that given in Section 1.12. Thus we may again obtain a set of n mutually orthogonal unit characteristic vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, orthogonality and length being defined in the Hermitian sense.

To solve the equation

$$\mathbf{h}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{c}, \tag{122}$$

we may then assume the expansion

$$\mathbf{x} = \sum_{k=1}^{n} \alpha_k \mathbf{e}_k, \tag{123}$$

as in the real case, so that (122) takes the form

$$\sum_{k=1}^{n} (\lambda_k - \lambda)\alpha_k \mathbf{e}_k = \mathbf{c}$$

and there follows

$$(\lambda_k - \lambda)\alpha_k = (\bar{\mathbf{e}}_k, \mathbf{c}).$$

Thus, if $\lambda \neq \lambda_k$, the solution becomes

$$\mathbf{x} = \sum_{k=1}^{n} \frac{(\bar{\mathbf{e}}_k, \mathbf{c})}{\lambda_k - \lambda} \mathbf{e}_k \tag{124}$$

in analogy with (89). If $\lambda = \lambda_p$, no solution exists unless $\mathbf{c}$ is such that $(\bar{\mathbf{e}}_p, \mathbf{c}) = 0$, in which case α_p is arbitrary, and infinitely many solutions exist.

The reduction of a Hermitian form to a sum of the *canonical form*

$$H = \lambda_1 \bar{x}'_1 x'_1 + \lambda_2 \bar{x}'_2 x'_2 + \cdots + \lambda_n \bar{x}'_n x'_n \tag{125}$$

may be accomplished by a method analogous to that of Section 1.13. Thus, if we write

$$\mathbf{x} = \mathbf{U}\,\mathbf{x}', \tag{126}$$

the form H of (115) becomes

$$H = (\bar{\mathbf{U}}\,\bar{\mathbf{x}}')^T \mathbf{h}\,\mathbf{U}\,\mathbf{x}' = \bar{\mathbf{x}}'^T (\bar{\mathbf{U}}^T \mathbf{h}\,\mathbf{U})\mathbf{x}'. \tag{127}$$

This form will be of type (125) if and only if the product matrix $\bar{\mathbf{U}}^T \mathbf{h}\,\mathbf{U}$ is a *diagonal* matrix. As in Section 1.12, a permissible choice of $\mathbf{U}$ consists in the normalized modal matrix formed by arranging the n orthogonalized unit characteristic vectors of $\mathbf{h}$ as its columns. For this matrix it can be shown that

$$\bar{\mathbf{U}}^T = \mathbf{U}^{-1} \qquad \text{or} \qquad \bar{\mathbf{U}}^T \mathbf{U} = \mathbf{I}. \tag{128}$$

A matrix $\mathbf{U}$ having the property (128) is called a *unitary (or Hermitian orthogonal) matrix*, and the product $\bar{\mathbf{U}}^r \mathbf{h}\, \mathbf{U}$ is then called a *unitary transformation of* $\mathbf{h}$. More generally, a transformation of the form $\bar{\mathbf{U}}^r \mathbf{h}\, \mathbf{U}$, where $\mathbf{U}$ does not necessarily satisfy (128), is called a *conjunctive* transformation.

1.17. Definite forms. If the quadratic form $\mathbf{x}^r \mathbf{a}\, \mathbf{x}$, associated with a real symmetric matrix $\mathbf{a}$, is *nonnegative* for all real values of the variables x_i, and is *zero* only if *each* of those n variables is zero, then that quadratic form is said to be *positive definite*. It is then conventional to say also that the matrix $\mathbf{a}$ is positive definite.

Similarly, a *Hermitian* matrix $\mathbf{a}$ is said to be positive definite if the associated *Hermitian* form $\bar{\mathbf{x}}^r \mathbf{a}\, \mathbf{x}$ is nonnegative for any real or complex vector $\mathbf{x}$, and vanishes only when $\mathbf{x} = 0$.

If a real quadratic form $A = \mathbf{x}^r \mathbf{a}\, \mathbf{x}$ is reducible by a transformation of the form $\mathbf{x} = \mathbf{Q}\, \mathbf{x}'$, where $\mathbf{Q}$ is a nonsingular real square matrix, to the sum of *squares* of the n new variables, each with a *positive* coefficient, then it is clear that A is a positive definite form relative to the real variables $x_1', \ldots, x_n'$. But from the relation $\mathbf{x}' = \mathbf{Q}^{-1} \mathbf{x}$, which is a consequence of the assumed nonvanishing of $|\, \mathbf{Q}\, |$, we see that a real vector $\mathbf{x}$ then corresponds always to a real vector $\mathbf{x}'$, and that the vectors $\mathbf{x} = 0$ and $\mathbf{x}' = 0$ then correspond uniquely. Hence it follows in this case that A is also positive definite relative to the *original* real variables $x_1, \ldots, x_n$.

Similarly, if a Hermitian form is reducible by a nonsingular complex transformation to the canonical form (125), wherein all coefficients are positive, the form is then nonnegative for any *complex* values of the variables, and is zero if and only if all the n variables vanish.

We notice that if the coefficients of the squares of any of the n variables are zero, then the vanishing of the form does not imply the vanishing of *those* variables, and hence the form is then *not* positive definite relative to the entire set of n variables.

It then follows from the results of the preceding sections [see equations (109) and (125)] that *a quadratic or Hermitian form is positive definite if and only if the characteristic numbers of the corresponding matrix are all positive.*

Positive definite forms are of particular importance in applications, and are found to possess certain useful properties. In par-

ticular, we show next that if at least one of the *two* real quadratic forms

$$A = \mathbf{x}^T \mathbf{a}\, \mathbf{x}, \qquad B = \mathbf{x}^T \mathbf{b}\, \mathbf{x} \tag{129a,b}$$

is *positive definite*, then it is always possible to reduce the two forms *simultaneously* to linear combinations of only squares of new variables, that is, to canonical forms, by a nonsingular real transformation. For this purpose, suppose that the form B is positive definite. Then, by proceeding exactly as in Section 1.13, we first set

$$\mathbf{x} = \mathbf{Q}\, \mathbf{y}, \tag{130}$$

where $\mathbf{Q}$ is the *normalized modal matrix* of $\mathbf{b}$, defined in that section, and so reduce B to the form

$$B = \mu_1 y_1^2 + \mu_2 y_2^2 + \cdots + \mu_n y_n^2, \tag{131}$$

where here μ_i is written for the ith characteristic number of the symmetric matrix $\mathbf{b}$. Since B is positive definite, the μ's are all positive. Hence we may make the substitution

$$\eta_i = \sqrt{\mu_i}\, y_i \qquad (i = 1, 2, \cdots, n), \tag{132}$$

and thus reduce (131) to the form

$$B = \eta_1^2 + \eta_2^2 + \cdots + \eta_n^2 = \mathbf{n}^T \mathbf{n}. \tag{133}$$

At the same time, the substitution (130) reduces A to the form

$$A = (\mathbf{Q}\, \mathbf{y})^T \mathbf{a}\, \mathbf{Q}\, \mathbf{y} = \mathbf{y}^T (\mathbf{Q}^T \mathbf{a}\, \mathbf{Q}) \mathbf{y} \tag{134}$$

and the subsequent substitution (132) reduces this form to the expression

$$A = \mathbf{n}^T (\mathbf{Q}'^T \mathbf{a}\, \mathbf{Q}') \mathbf{n}, \tag{135}$$

where $\mathbf{Q}'$ is a matrix obtained from $\mathbf{Q}$ by dividing each element of the ith column of $\mathbf{Q}$ by $\sqrt{\mu_i}$. Hence, if we write

$$\mathbf{g} = \mathbf{Q}'^T \mathbf{a}\, \mathbf{Q}', \tag{136}$$

equation (135) takes the form

$$A = \mathbf{n}^T \mathbf{g}\, \mathbf{n}. \tag{137}$$

Now $\mathbf{g}$ is a symmetric matrix, since

$$\mathbf{g}^T = (\mathbf{Q}'^T \mathbf{a}\, \mathbf{Q}')^T = \mathbf{Q}'^T \mathbf{a}^T \mathbf{Q}' = \mathbf{Q}'^T \mathbf{a}\, \mathbf{Q}' = \mathbf{g}. \tag{138}$$

Hence we may reduce (137) to canonical form by setting

$$\mathbf{n} = \mathbf{R}\,\boldsymbol{\alpha}, \tag{139}$$

where R is made up of the characteristic vectors of g just as Q is formed from those of b, and (137) is reduced to the form

$$A = \lambda_1\alpha_1^2 + \lambda_2\alpha_2^2 + \cdots + \lambda_n\alpha_n^2 \tag{140}$$

where λ_i is the ith characteristic number of the matrix g.

At the same time, the final substitution (139) reduces (133) to

$$B = \mathbf{n}^T\mathbf{n} = (\mathbf{R}\,\boldsymbol{\alpha})^T(\mathbf{R}\,\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T\mathbf{R}^T\mathbf{R}\,\boldsymbol{\alpha}. \tag{141}$$

But since the matrix R is an *orthogonal* matrix, there follows $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, and hence we have the result

$$B = \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2. \tag{142}$$

Thus, finally, with the substitution

$$\mathbf{x} = \mathbf{Q}\,\mathbf{y} = \mathbf{Q}'\,\mathbf{n} = \mathbf{Q}'\,\mathbf{R}\,\boldsymbol{\alpha}, \tag{143}$$

the two forms (129a, b) are simultaneously reduced to the canonical forms (140) and (142).

If we define the diagonal matrix

$$\mathbf{m} = \begin{bmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & m_n \end{bmatrix}, \qquad m_i = \frac{1}{\sqrt{\mu_i}} \tag{144}$$

it follows that

$$\mathbf{Q}' = \mathbf{Q}\,\mathbf{m} \tag{145}$$

and (143) becomes

$$\mathbf{x} = \mathbf{Q}\,\mathbf{m}\,\mathbf{R}\,\boldsymbol{\alpha}. \tag{146}$$

Since Q and R are orthogonal matrices, with determinants equal to unity in absolute value [see equation (106)], and since clearly $|\,\mathbf{m}\,| \neq 0$, it follows that the transformation (143) is indeed non-singular.

In certain applications to dynamical problems (see Section 2.12) the positive definite form B (*kinetic energy*) involves the time derivative $d\mathbf{x}/dt$ in place of x, whereas the form A (*potential energy*) involves only x itself. The above reduction is still applicable,

however, since $\mathbf{x}$ and $d\mathbf{x}/dt$ are transformed in the same way at each step of the process.

Another method of accomplishing the same reduction, which is usually more conveniently applied in practice, is presented in Section 1.25 (see page 77).

1.18. Discriminants and invariants. It is frequently of importance to determine whether a quadratic or Hermitian form which involves cross-product terms is or is not a *positive definite* form, without reducing it to a canonical form or determining the characteristic numbers of the associated matrix. This problem is to be considered in the present section.

If we write the characteristic equation $|\mathbf{a} - \lambda \mathbf{I}| = 0$ of a square matrix $\mathbf{a}$ in the form

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$

$$\equiv (-1)^n[\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - \cdots + (-1)^n\beta_n] = 0, \quad (147)$$

and denote the n roots of this equation as $\lambda_1, \lambda_2, \ldots, \lambda_n$, numbering multiple roots separately, it follows that

$$\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - \cdots + (-1)^n\beta_n \equiv (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n). \quad (148)$$

By comparing coefficients of λ in the two sides of (148), it can be shown that

$$\left.\begin{aligned} \beta_1 &= \lambda_1 + \lambda_2 + \cdots + \lambda_n, \\ \beta_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_{n-1}\lambda_n, \\ \beta_3 &= \lambda_1\lambda_2\lambda_3 + \cdots + \lambda_{n-2}\lambda_{n-1}\lambda_n, \\ &\cdots\cdots\cdots\cdots\cdots \\ \beta_n &= \lambda_1\lambda_2\lambda_3 \cdots \lambda_n \end{aligned}\right\}. \quad (149)$$

Now, for either a *real symmetric* or a *Hermitian* matrix, we have shown that the roots of (147) are all *real*. Hence, by Descartes' rule of signs, we see in such cases that *the roots of the characteristic*

equation (147) *are all positive if and only if the quantities* β_1, β_2, . . . , β_n *are all positive.*

From (147) it follows that β_n is the value of $|\mathbf{a} - \lambda \mathbf{I}|$ when $\lambda = 0$; that is, β_n is the value of the determinant of $\mathbf{a}$:

$$\beta_n = |a_{ij}|. \tag{150}$$

Further, it is easily seen that the coefficient of λ^{n-1} in the expansion of the determinant in (147) is merely

$$(-1)^{n+1}(a_{11} + a_{22} + \cdots + a_{nn});$$

that is, β_1 is the sum of the *diagonal elements* of $\mathbf{a}$:

$$\beta_1 = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{k=1}^{n} a_{kk}. \tag{151}$$

This sum is called the *trace* of $\mathbf{a}$.

More generally, it can be shown that β_i *is the sum of all determinants formed from square arrays of order* i *whose principal diagonals lie along the principal diagonal of* $\mathbf{a}$. Such determinants are called the *principal minors* of $\mathbf{a}$.

Thus it follows that *a quadratic or Hermitian form is positive definite if and only if these sums, relevant to the associated matrix, are all positive.*

In illustration, the quadratic form

$$F = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{23}x_2x_3 + 2a_{13}x_1x_3 \tag{152}$$

in three dimensions, which is associated with the real matrix

$$\mathbf{a} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}, \tag{153}$$

is positive definite if and only if the three conditions

$$a_{11} + a_{22} + a_{33} > 0, \tag{154a}$$

$$(a_{11}a_{22} - a_{12}^2) + (a_{22}a_{33} - a_{23}^2) + (a_{11}a_{33} - a_{13}^2) > 0, \tag{154b}$$

$$|a_{ij}| > 0, \tag{154c}$$

are satisfied.

It is readily verified by direct expansion that the determinant of the *symmetric* matrix (153) can be written in the form

$$|a_{ij}| = \frac{(a_{11}a_{22} - a_{12}^2)(a_{11}a_{33} - a_{13}^2) - (a_{11}a_{23} - a_{12}a_{13})^2}{a_{11}},$$

and also in two further equivalent forms obtained by cyclic permutation of the subscripts. Suppose that we require only that

$$a_{11} > 0, \qquad a_{11}a_{22} - a_{12}^2 > 0, \qquad |a_{ij}| > 0. \tag{155a,b,c}$$

It then follows from (155a,b) that we must have $a_{22} > 0$, and also, by referring to the above form for $|a_{ij}|$, we see that (155a,b,c) imply that $a_{11}a_{33} - a_{13}^2 > 0$. By considering the permutation of that form in which $1 \to 2$, $2 \to 3$, $3 \to 1$, we then deduce similarly that (155a,b,c) also imply the inequalities $a_{22}a_{33} - a_{23}^2 > 0$ and $a_{33} > 0$. Thus it follows that *the three conditions* (155) *imply the three conditions* (154).

By considering the conditions that (152) still be positive definite when, first, only *one* variable differs from zero and when, second, only *two* variables differ from zero, it is easily shown that *each* diagonal term a_{ii} must be positive and also that *each* principal minor of second order must be positive. Hence these conditions imply and must be implied by either the conditions (154) or the more convenient conditions (155).

More generally, if for *any* real symmetric (or Hermitian) matrix $\mathbf{a}$ we define the mth *discriminant* Δ_m to be the determinant of the matrix $\mathbf{D}_m$ obtained by deleting all elements which do not simultaneously lie in the first m rows and columns of $\mathbf{a}$, it can be shown that *the real symmetric* (*or Hermitian*) *matrix* $\mathbf{a}$, *and the corresponding quadratic* (*or Hermitian*) *form, is positive definite if and only if each of the* n *discriminants* Δ_m *is positive.* If and only if this is so, *all* the principal minors of $\mathbf{a}$ are positive.

To establish the sufficiency of this criterion, we need only prove that, if $\mathbf{D}_m$ is positive definite and $\Delta_{m+1} \equiv |\mathbf{D}_{m+1}|$ is positive, then $\mathbf{D}_{m+1}$ is also positive definite. Suppose, on the contrary, that $\mathbf{D}_{m+1}$ is *not* positive definite. Then, since $|\mathbf{D}_{m+1}|$ is the *product* of the characteristic numbers of $\mathbf{D}_{m+1}$, it follows that an *even* number of these characteristic numbers must be negative. Let γ_1 and γ_2 be two such numbers, and denote by $\mathbf{u}_1$ and $\mathbf{u}_2$ corresponding orthogonal unit characteristic vectors of $\mathbf{D}_{m+1}$, length

and orthogonality being defined in the Hermitian sense. If we define the $(m+1)$-dimensional vector

$$\mathbf{x}^* = c_1\mathbf{u}_1 + c_2\mathbf{u}_2,$$

where at least one of the c's does not vanish, and notice that then $\mathbf{D}_{m+1}\,\mathbf{u}_1 = \gamma_1\mathbf{u}_1$ and $\mathbf{D}_{m+1}\,\mathbf{u}_2 = \gamma_2\mathbf{u}_2$, there follows easily

$$\bar{\mathbf{x}}^{*T}\,\mathbf{D}_{m+1}\,\mathbf{x}^* = \bar{c}_1c_1\gamma_1 + \bar{c}_2c_2\gamma_2 < 0,$$

for any c_1 and c_2. Thus the vector $\mathbf{x}^*$ renders the Hermitian form associated with $\mathbf{D}_{m+1}$ negative. Now let c_1 and c_2 be related in such a way that the component x^*_{m+1} vanishes. If we notice that the Hermitian form $\bar{\mathbf{x}}^T\,\mathbf{D}_{m+1}\,\mathbf{x}$ reduces to the form $\bar{\mathbf{x}}^T\,\mathbf{D}_m\,\mathbf{x}$ when $x_{m+1} = 0$, we conclude that the m-dimensional vector made up of the first m components of the $\mathbf{x}^*$ so determined renders the Hermitian form associated with $\mathbf{D}_m$ negative. Since $\mathbf{D}_m$ is positive definite, this situation is impossible, and the desired contradiction is obtained.

The specialization of the preceding argument to the case of a real symmetric matrix, and its associated real quadratic form, is obtained by deleting the bars indicating complex conjugates. In this case, $\mathbf{u}_1$ and $\mathbf{u}_2$ are real, and the constants c_1 and c_2 are also to be real.

Whereas the requirements that a form or matrix be positive definite thus need not be stated in terms of the *sums* β_i, these sums nevertheless are of considerable importance in themselves. We see from (149) that each β_i is a symmetric function, of degree i, of the characteristic numbers of $\mathbf{a}$. Also, it follows from (147) that *for any two square matrices* $\mathbf{a}$ *and* $\mathbf{b}$ *such that* $|\,\mathbf{a} - \lambda\mathbf{I}\,| = |\,\mathbf{b} - \lambda\mathbf{I}\,|$ *for all values of* λ, *the* n *quantities* β_i *are the same.*

In order to determine conditions under which this situation exists, let $\mathbf{a}$ and $\mathbf{b}$ be two equivalent matrices. This means that nonsingular matrices $\mathbf{P}$ and $\mathbf{Q}$ exist such that $\mathbf{b} = \mathbf{P}\,\mathbf{a}\,\mathbf{Q}$. Hence we have, for any value of λ,

$$\begin{aligned}\mathbf{b} - \lambda\mathbf{I} &= \mathbf{P}\,\mathbf{a}\,\mathbf{Q} - \lambda\mathbf{I} \\ &= \mathbf{P}(\mathbf{a} - \lambda\,\mathbf{P}^{-1}\,\mathbf{Q}^{-1})\mathbf{Q}\end{aligned}$$

and also

$$|\,\mathbf{b} - \lambda\mathbf{I}\,| = |\,\mathbf{P}\,|\,|\,\mathbf{Q}\,|\,|\,\mathbf{a} - \lambda\,\mathbf{P}^{-1}\,\mathbf{Q}^{-1}\,|. \tag{156}$$

Thus, if $\mathbf{P}$ and $\mathbf{Q}$ are such that $\mathbf{P}^{-1}\mathbf{Q}^{-1} = \mathbf{I}$ or $\mathbf{P} = \mathbf{Q}^{-1}$, so that

$$\mathbf{b} = \mathbf{Q}^{-1}\mathbf{a}\,\mathbf{Q}, \tag{157}$$

there follows also $\mathbf{P}\mathbf{Q} = \mathbf{I}$, and hence $|\mathbf{P}|\,|\mathbf{Q}| = 1$, and (156) takes the form

$$|\mathbf{b} - \lambda\mathbf{I}| = |\mathbf{a} - \lambda\mathbf{I}|, \tag{158}$$

for all values of λ.

A transformation of the form (157) has been defined as a *similarity transformation,* and the matrices $\mathbf{a}$ and $\mathbf{b}$ are said to be *similar*. Since (158) states that $\mathbf{a}$ and $\mathbf{b}$ have the same characteristic equation, it follows that *the quantities β_i are invariant under* (unchanged by) *any similarity transformation.* This result has important consequences in many physical considerations.

Since *orthogonal* and *unitary* transformations are special types of similarity transformations, in which also $\mathbf{Q}^{-1} = \mathbf{Q}^{T}$ and $\mathbf{Q}^{-1} = \bar{\mathbf{Q}}^{T}$, respectively, the preceding statement applies to them.

1.19. Coordinate transformations. That similarity transformations are of frequent interest in practice, is indicated by the following consideration.

The elements of any vector $\mathbf{x}$ in a space of n dimensions may be interpreted as the components of that vector in the directions of the n mutually orthogonal unit vectors $\mathbf{i}_1 = \{1, 0, \ldots, 0\}, \ldots, \mathbf{i}_n = \{0, 0, \ldots, 1\}$ which lie along the axes of the coordinates $x_1, x_2, \ldots, x_n$. That is, we may write

$$\mathbf{x} = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + \cdots + x_n\mathbf{i}_n = \sum_{r=1}^{n} x_r\mathbf{i}_r. \tag{159}$$

Now let a new coordinate system be so chosen that the unit vectors $\mathbf{i}'_1, \mathbf{i}'_2, \ldots, \mathbf{i}'_n$ in the directions of the axes of the new coordinates $x'_1, x'_2, \ldots, x'_n$ are related to the original unit vectors by the equations

$$\left.\begin{aligned} \mathbf{i}'_1 &= Q_{11}\mathbf{i}_1 + Q_{21}\mathbf{i}_2 + \cdots + Q_{n1}\mathbf{i}_n, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathbf{i}'_n &= Q_{1n}\mathbf{i}_1 + Q_{2n}\mathbf{i}_2 + \cdots + Q_{nn}\mathbf{i}_n \end{aligned}\right\} \tag{160}$$

or

$$\mathbf{i}'_i = \sum_{r=1}^{n} Q_{ri}\mathbf{i}_r. \tag{161}$$

The quantity represented by the vector $\mathbf{x}$ can then be expressed in terms of its components $x_1', x_2', \ldots, x_n'$ along the new axes. If we denote the vector specified by this array of components by $\mathbf{x}'$, there follows

$$\mathbf{x}' = x_1'\mathbf{i}_1' + x_2'\mathbf{i}_2' + \cdots + x_n'\mathbf{i}_n' = \sum_{k=1}^{n} x_k'\mathbf{i}_k'. \tag{162}$$

To determine the new components in terms of the original ones, we first introduce (161) into (162):

$$\mathbf{x}' = \sum_{k=1}^{n}\sum_{r=1}^{n} x_k' Q_{rk}\mathbf{i}_r = \sum_{r=1}^{n}\left(\sum_{k=1}^{n} Q_{rk}x_k'\right)\mathbf{i}_r. \tag{163}$$

Then, since (163) and (159) represent the same quantity, and since the vectors $\mathbf{i}_r$ are mutually orthogonal, their respective coefficients in (159) and (163) must be equal, so that

$$x_r = \sum_{k=1}^{n} Q_{rk}x_k'. \tag{164}$$

Thus, if we write $\mathbf{x}' = \{x_1', x_2', \ldots, x_n'\}$ for the vector comprising the components of $\mathbf{x}$ in the directions of the new coordinate axes specified by (160), there follows

$$\mathbf{x} = \mathbf{Q}\,\mathbf{x}', \tag{165}$$

where $\mathbf{Q}$ is the *transformation matrix*

$$\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1n} \\ Q_{21} & Q_{22} & \cdots & Q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ Q_{n1} & Q_{n2} & \cdots & Q_{nn} \end{bmatrix}, \tag{166}$$

of which the coefficient matrix in (160) is the *transpose*. We notice that each *column* of (166) contains *the components of a new unit vector along the original coordinate axes.*

Here we interpret the matrix $\mathbf{Q}$ of (165) as relating the components of a vector along the original coordinate axes to the components of the *same vector* along the new coordinate axes. In other considerations we may suppose that no *change of axes* is involved, and that an equation of the form (165) merely transforms one vector into another one, both vectors then being referred to the same axes.

Which interpretation is to be attached to such an equation in practice clearly depends upon the nature of the problem involved.

In order that equations (160) be solvable for the vectors $\mathbf{i}_r$ in terms of the vectors $\mathbf{i}'_r$, the determinant $|\mathbf{Q}|$ must not vanish; that is, *the matrix* $\mathbf{Q}$ *must be nonsingular.* Hence $\mathbf{Q}^{-1}$ then exists, and we have also, from (165),

$$\mathbf{x}' = \mathbf{Q}^{-1}\mathbf{x}. \tag{167}$$

Suppose now that two vectors are related by an equation of the form

$$\mathbf{y} = \mathbf{a}\,\mathbf{x}, \tag{168}$$

when the components refer to the original coordinate frame, and that the corresponding relationship between the components referred to a new coordinate frame (160) is required. (We may, for example, imagine that $\mathbf{y}$ represents *force* and $\mathbf{x}$ *acceleration.* In *Newtonian* mechanics, the matrix $\mathbf{a}$ would then be a *scalar* matrix.) By replacing $\mathbf{x}$ by $\mathbf{Q}\,\mathbf{x}'$ and $\mathbf{y}$ by $\mathbf{Q}\,\mathbf{y}'$, under the assumption that $\mathbf{x}$ and $\mathbf{y}$ transform in the same way, we obtain the relation

$$\mathbf{Q}\,\mathbf{y}' = \mathbf{a}\,\mathbf{Q}\,\mathbf{x}'$$

and hence, after premultiplying both sides by $\mathbf{Q}^{-1}$, we obtain the desired result

$$\mathbf{y}' = (\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q})\mathbf{x}'. \tag{169}$$

Thus we see that the matrix relating $\mathbf{x}'$ and $\mathbf{y}'$ is obtained from that relating $\mathbf{x}$ and $\mathbf{y}$ by a *similarity transformation.* In particular, it follows that the invariance properties discussed at the close of the preceding section apply in the present case. That is, the quantities β_i of that section, pertaining to the matrix $\mathbf{a}$ of (168), are invariant under a nonsingular linear coordinate transformation. This result is of great importance.

If the *new* unit vectors are *mutually orthogonal,* we readily obtain from (166) the result

$$\mathbf{Q}^r\,\mathbf{Q} = \mathbf{I} \qquad \text{or} \qquad \mathbf{Q}^r = \mathbf{Q}^{-1}, \tag{170}$$

so that $\mathbf{Q}$ is then an *orthogonal* matrix. Thus *a transformation from one set of orthogonal axes to another is accomplished by an orthogonal transformation.* We may verify that in such a transformation the

length of a vector in the new system is the same as its length in the original system (that is, that there is no change in *scale*). For if $\mathbf{x} = \mathbf{Q}\,\mathbf{x}'$ there follows

$$l^2 = \mathbf{x}^T\mathbf{x} = (\mathbf{Q}\,\mathbf{x}')^T\mathbf{Q}\,\mathbf{x}' = \mathbf{x}'^T\mathbf{Q}^T\mathbf{Q}\,\mathbf{x}' = \mathbf{x}'^T\mathbf{x}' = l'^2. \quad (171)$$

Also, the magnitude of the scalar product of two vectors is the same in both systems (that is, the magnitude of an "angle" is also preserved), since

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} = \mathbf{x}'^T\mathbf{Q}^T\mathbf{Q}\,\mathbf{y}' = \mathbf{x}'^T\mathbf{y}' = (\mathbf{x}', \mathbf{y}'). \quad (172)$$

As these results suggest, it can be shown that any orthogonal transformation in space of n dimensions can be interpreted as a combination of *rotations* and *reflections*.

1.20. Diagonalization of symmetric matrices. In the two sections which follow, we reconsider, from the point of view of the preceding section, the problem of reducing a real symmetric matrix (or the associated quadratic form) to a canonical form, and establish a result stated without proof in Section 1.11 (see page 32) and used in succeeding sections. For completeness, certain notions and results already discussed are first summarized.

Associated with a real square matrix $\mathbf{a}$, we consider first a set of equations of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= y_n \end{aligned}\right\} \quad (173)$$

which, when written in the matrix form

$$\mathbf{a}\,\mathbf{x} = \mathbf{y} \quad (174)$$

establishes a relationship between the vectors $\mathbf{x}$ and $\mathbf{y}$. If we require also that

$$\mathbf{y} = \lambda\,\mathbf{x}, \quad (175)$$

so that the "transform" of $\mathbf{x}$ is to be parallel to $\mathbf{x}$, we are led to a consideration of the equation

$$(\mathbf{a} - \lambda\,\mathbf{I})\mathbf{x} = \mathbf{0}, \quad (176)$$

which can be satisfied nontrivially if and only if λ is a root of the characteristic equation

$$|\mathbf{a} - \lambda \mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{177}$$

If $\mathbf{a}$ is a *symmetric* matrix, it is associated also with the quadratic form

$$F = \mathbf{x}^T \mathbf{a}\, \mathbf{x}, \tag{178}$$

the value of which is merely $\mathbf{x}^T \mathbf{y} = (\mathbf{x}, \mathbf{y})$; that is, F is the scalar product of $\mathbf{x}$ and its transform. The form F involves only squared terms if and only if $\mathbf{a}$ is a diagonal matrix.

Suppose now that a *coordinate transformation* of the form (165) is made, new coordinates being introduced by the matrix equation

$$\mathbf{x} = \mathbf{Q}\, \mathbf{x}', \tag{179}$$

and that the new axes are required to be orthogonal. This means that the new unit vectors, as given by (160), are to be mutually orthogonal, and that the transformation matrix $\mathbf{Q}$ must be an orthogonal matrix:

$$\mathbf{Q}^T = \mathbf{Q}^{-1} \quad \text{or} \quad \mathbf{Q}\, \mathbf{Q}^T = \mathbf{I}. \tag{180}$$

Length and magnitude of angle are then preserved.

If we write accordingly

$$\mathbf{y} = \mathbf{Q}\, \mathbf{y}', \tag{181}$$

the vectors $\mathbf{x}'$ and $\mathbf{y}'$ are related by the vector equation

$$(\mathbf{Q}^{-1}\, \mathbf{a}\, \mathbf{Q})\mathbf{x}' = \mathbf{y}', \tag{182}$$

or by corresponding scalar equations which may be written in the form

$$\left.\begin{aligned} \alpha_{11}x_1' + \alpha_{12}x_2' + \cdots + \alpha_{1n}x_n' &= y_1', \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ \alpha_{n1}x_1' + \alpha_{n2}x_2' + \cdots + \alpha_{nn}x_n' &= y_n' \end{aligned}\right\}. \tag{183}$$

The result of introducing (179) into (178), which is of the form $F = \mathbf{x}^T \mathbf{y} = \mathbf{x}'^T \mathbf{Q}^T \mathbf{a}\, \mathbf{Q}\, \mathbf{x}'$, is identical with the product $\mathbf{x}'^T \mathbf{y}'$,

where $\mathbf{y}'$ is given by (182), only under the restriction that (180) be satisfied, so that the transformation is orthogonal.*

The present problem consists in attempting to determine a new coordinate system, specified by the orthogonal matrix $\mathbf{Q}$, such that the coefficient matrix of (183) is a *diagonal* matrix; that is, such that $\alpha_{ij} = 0$ when $i \neq j$. The form (178) will then involve only *squares* of the new coordinates x_i'.

Let λ_1 be a root of the characteristic equation (177), and let a *unit* vector which satisfies (176) when $\lambda = \lambda_1$ be denoted by $\mathbf{e}_1 = \{e_{11}, e_{12}, \cdots, e_{1n}\}$. Then we may require that the direction of the *first* axis in the new coordinate system coincide with the direction of $\mathbf{e}_1$ in the original system. Hence, in accordance with (160), we must take

$$Q_{11} = e_{11}, \qquad Q_{21} = e_{12}, \qquad \cdots, \qquad Q_{n1} = e_{1n}. \tag{184}$$

Now it must follow that the transform of the vector $\{1, 0, \ldots, 0\}$, in the new system, is λ_1 times itself, and hence is the vector $\{\lambda_1, 0, \ldots, 0\}$. When this condition is imposed on (183), there follows

$$\alpha_{11} = \lambda_1, \qquad \alpha_{21} = \alpha_{31} = \cdots = \alpha_{n1} = 0. \tag{185}$$

Hence, if the first column of $\mathbf{Q}$ is made up of the elements of $\mathbf{e}_1$, there follows

$$\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q} = \begin{bmatrix} \lambda_1 & \alpha_{12} & \cdots & \alpha_{1n} \\ 0 & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix}. \tag{186}$$

But if $\mathbf{a}$ is *symmetric*, this matrix must also be symmetric; for since also $\mathbf{Q}$ is orthogonal there follows

$$(\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q})^T = (\mathbf{Q}^T\,\mathbf{a}\,\mathbf{Q})^T = \mathbf{Q}^T\,\mathbf{a}^T\,\mathbf{Q} = \mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q}.$$

Hence, *in this case*, there follows also

$$\alpha_{12} = \alpha_{13} = \cdots = \alpha_{1n} = 0, \tag{187}$$

* When two vectors $\mathbf{x}$ and $\mathbf{y}$ undergo the *same transformation*, as in (179) and (181), the two sets of variables which comprise their components are said to be *cogredient*. When the vectors are transformed separately in such a way that the condition $(\mathbf{x}, \mathbf{y}) = (\mathbf{x}', \mathbf{y}')$ is satisfied, the two sets of variables are said to be *contragredient*. As is seen here, when two vectors undergo the same *orthogonal* transformation their components are *both cogredient and contragredient*.

and (186) becomes

$$\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \alpha_{2n} & \cdots & \alpha_{nn} \end{bmatrix}. \tag{188}$$

It is easily shown (see also Section 1.13) that, if the *second* column of $\mathbf{Q}$ comprises the elements of a second unit characteristic vector $\mathbf{e}_2$, all elements in the second row and second column of (188) except α_{22} reduce to zero and α_{22} is identified with the corresponding characteristic number λ_2.

However, since an orthogonal coordinate transformation preserves the magnitude of angles, the essential difficulty is involved in showing that the second axis direction ($x_1' = 0$, $x_2' = 1$, $x_3' = \ldots = x_n' = 0$) *can* in fact be identified with the direction of $\mathbf{e}_2$; that is, *that* $\mathbf{e}_2$ *is orthogonal to* $\mathbf{e}_1$. If these vectors correspond to *different* characteristic numbers, $\lambda_1 \neq \lambda_2$, we have shown (Section 1.11) that they *are* orthogonal. Thus, if *all* the characteristic numbers of $\mathbf{a}$ are distinct, a generalization of the above argument leads easily to the result of Section 1.13, which states that if the elements of *each* column of $\mathbf{Q}$ are then formed from distinct characteristic unit vectors, the matrix (186) takes the diagonal form $[\lambda_i\,\delta_{ij}]$.

In the case when λ_1 is a root of (177) of multiplicity s, it remains to prove that s corresponding *linearly independent* characteristic vectors exist, so that s mutually orthogonal unit vectors can be determined as linear combinations of them. The truth of this assertion was assumed in Section 1.13.

1.21. Multiple characteristic numbers. Suppose that λ_1 is a repeated characteristic number of $\mathbf{a}$, so that $(\lambda - \lambda_1)^2$ is a factor of $|\,\mathbf{a} - \lambda\mathbf{I}\,|$. From (188) it follows that, if $\mathbf{Q}$ is *any* orthogonal matrix such that the elements of its first column are the components of *one* characteristic unit vector corresponding to λ_1, we have

$$\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q} - \lambda\,\mathbf{I} = \begin{bmatrix} \lambda_1 - \lambda & 0 & \cdots & 0 \\ 0 & \alpha_{22} - \lambda & \cdots & \alpha_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \alpha_{2n} & \cdots & \alpha_{nn} - \lambda \end{bmatrix}. \tag{189}$$

But the determinant of this matrix is identical with $|\,\mathbf{a} - \lambda\,\mathbf{I}\,|$ [see

equation (158)] and hence must possess a factor $(\lambda - \lambda_1)^2$. Thus it follows that the cofactor of $\lambda_1 - \lambda$ in (189) must vanish when $\lambda = \lambda_1$, so that the rank of the matrix (189) cannot be greater than $n - 2$ when $\lambda = \lambda_1$. Since the matrix (189) is *equivalent* to the matrix $\mathbf{a} - \lambda\mathbf{I}$,

$$\mathbf{Q}^{-1}\mathbf{a}\mathbf{Q} - \lambda\mathbf{I} = \mathbf{Q}^{-1}(\mathbf{a} - \lambda\mathbf{I})\mathbf{Q},$$

the same statement applies to that matrix.

Thus it follows that if λ_1 is a multiple characteristic number of $\mathbf{a}$, the equation

$$(\mathbf{a} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

possesses at least two linearly independent solutions when $\lambda = \lambda_1$. Hence a second characteristic unit vector $\mathbf{e}_2$, corresponding to λ_1, can indeed be determined in such a way that it is orthogonal to $\mathbf{e}_1$, and the reduction can be advanced by one step.

Consequently, if the first *two* columns of $\mathbf{Q}$ comprise the elements of $\mathbf{e}_1$ and $\mathbf{e}_2$, equation (189) then reduces to

$$\mathbf{Q}^{-1}\mathbf{a}\mathbf{Q} - \lambda\mathbf{I} = \begin{bmatrix} \lambda_1 - \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda_1 - \lambda & 0 & \cdots & 0 \\ 0 & 0 & \alpha_{33} - \lambda & \cdots & \alpha_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \alpha_{3n} & \cdots & \alpha_{nn} - \lambda \end{bmatrix}. \tag{190}$$

If the multiplicity of λ_1 is greater than two, the preceding argument leads to the conclusion that the matrix $\mathbf{a} - \lambda\mathbf{I}$ is of rank not greater than $n - 3$ when $\lambda = \lambda_1$, so that at least three linearly independent corresponding characteristic vectors can be obtained.

By inductive reasoning, we thus deduce that if λ_1 is a characteristic number of a *symmetric* matrix $\mathbf{a}$, of multiplicity s, and $\mathbf{a}$ is of order n, then the rank of the matrix $\mathbf{a} - \lambda\mathbf{I}$ is not greater than $n - s$ when $\lambda = \lambda_1$, so that *at least* s linearly independent characteristic vectors, corresponding to λ_1, can be obtained. However, the rank also cannot be *less* than $n - s$, for if this were so, more than s linearly independent characteristic vectors would correspond to λ_1, in which case the total number of linearly independent characteristic vectors corresponding to *all* characteristic numbers would be greater

than the dimension n of the space involved. Hence we obtain the following important result:

If λ_1 is a characteristic number, of multiplicity s, of a symmetric matrix $\mathbf{a}$ of order n, then the rank of the matrix $\mathbf{a} - \lambda \mathbf{I}$ is exactly $n - s$ when $\lambda = \lambda_1$; that is, there exist exactly s linearly independent corresponding characteristic vectors.

This statement does not apply, in general, to a *nonsymmetric* matrix, as was shown by an example in Section 1.11. However, an argument analogous to that given above shows that *the statement does apply also to Hermitian matrices.*

In the general nonsymmetric case, it is shown in Section 1.26 that a matrix $\mathbf{a}$ with n *distinct* characteristic numbers possesses n *linearly independent* characteristic vectors. If a *modal matrix* $\mathbf{Q}$ is formed, in such a way that the components of successive vectors comprise successive columns of $\mathbf{Q}$, the matrix $\mathbf{a}$ can be diagonalized by the *similarity* transformation $\mathbf{Q}^{-1}\,\mathbf{a}\,\mathbf{Q}$, the resultant diagonal elements being the characteristic numbers of $\mathbf{a}$ (see Problem 51). However, in consequence of the fact that the n characteristic vectors are generally not orthogonal, it follows that $\mathbf{Q}^{-1} \neq \mathbf{Q}^T$, in general, so that the matrix $\mathbf{Q}$ is generally *not* an *orthogonal matrix.*

If certain characteristic numbers of a nonsymmetric and non-Hermitian matrix are repeated, there may be less than n linearly independent characteristic vectors, so that complete diagonalization in this way is impossible. In any case, it can be shown that *any* square matrix can be transformed by a *similarity* transformation (which is not necessarily orthogonal) to a *canonical matrix* with the following properties:

1. All elements *below* the principal diagonal are zero.
2. The diagonal elements are the characteristic numbers of the matrix.
3. All elements *above* the principal diagonal are zero *except* possibly those elements which are adjacent to *two* equal diagonal elements.
4. The latter elements are each either zero or unity.

A matrix having these four properties is known as a *Jordan canonical matrix.*

In illustration, for a matrix of order five for which $\lambda_1 = \lambda_2 = \lambda_3$ and $\lambda_4 = \lambda_5$, but $\lambda_1 \neq \lambda_4$, this canonical form would be

$$\begin{bmatrix} \lambda_1 & \alpha_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & \alpha_2 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & \alpha_3 \\ 0 & 0 & 0 & 0 & \lambda_4 \end{bmatrix},$$

where each of the elements α_1, α_2, and α_3 is either unity or zero, according as λ_1 corresponds to one, two, or three independent characteristic vectors, and λ_4 to one or two independent characteristic vectors. Reductions to certain other standard forms have also been studied.*

1.22. Functions of symmetric matrices. In this section, we restrict attention to real *symmetric* matrices, which are of principal interest in applications. We notice first that, as is easily shown, *the sum of two symmetric matrices of the same order is also symmetric*, while *the product of two symmetric matrices of the same order is symmetric if those matrices are commutative.*

Positive integral powers of a square matrix $\mathbf{a}$ are defined by iteration:

$$\mathbf{a}^2 = \mathbf{a}\,\mathbf{a}, \quad \mathbf{a}^3 = \mathbf{a}\,\mathbf{a}^2, \quad \cdots, \quad \mathbf{a}^{n+1} = \mathbf{a}\,\mathbf{a}^n, \quad \cdots. \tag{191}$$

In consequence of this definition, there follows also

$$\mathbf{a}^r\,\mathbf{a}^s = \mathbf{a}^s\,\mathbf{a}^r = \mathbf{a}^{r+s}, \tag{192}$$

when r and s are positive integers. *Negative integral* powers are defined only for *nonsingular* matrices, for which a unique inverse $\mathbf{a}^{-1}$ exists, and are then defined by the relation

$$\mathbf{a}^{-n} = (\mathbf{a}^{-1})^n. \tag{193}$$

If we define also

$$\mathbf{a}^0 = \mathbf{I}, \tag{194}$$

then (192) applies to any nonsingular matrix, for *any* integers r and s. It is clear that *any integral power of a symmetric matrix is also symmetric.*

Polynomial functions of $\mathbf{a}$ are then defined as linear combinations of nonnegative integral powers of $\mathbf{a}$. Any polynomial in $\mathbf{a}$ can hence be expressed as a symmetric matrix of the same order as $\mathbf{a}$.

* See Reference 6.

Suppose now that $\mathbf{a}$ is of order n, and let its characteristic numbers be denoted by $\lambda_1, \lambda_2, \ldots, \lambda_n$ (not necessarily distinct), with corresponding orthogonalized characteristic unit vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. That is, let λ_i and $\mathbf{e}_i$ be such that

$$\mathbf{a}\,\mathbf{e}_i = \lambda_i \mathbf{e}_i \tag{195}$$

for $i = 1, 2, \ldots, n$. If we multiply both sides of (195) by $\mathbf{a}$, and use (195) to simplify the resulting right-hand member, there follows

$$\mathbf{a}^2\,\mathbf{e}_i = \lambda_i \mathbf{a}\,\mathbf{e}_i = \lambda_i^2 \mathbf{e}_i, \tag{196}$$

and, by repeating this process, we deduce from (195) the relation

$$\mathbf{a}^r\,\mathbf{e}_i = \lambda_i^r \mathbf{e}_i \tag{197}$$

for any positive integer r. Similarly, if $\mathbf{a}$ is *nonsingular*, the result of multiplying both sides of (195) by $\mathbf{a}^{-1}$ becomes

$$\mathbf{a}^{-1}\,\mathbf{e}_i = \lambda_i^{-1} \mathbf{e}_i \tag{198}$$

and, by iteration, we find that (197) is then true for any integer r.

Thus we deduce that *if λ_i is a characteristic number of $\mathbf{a}$, with a corresponding characteristic vector $\mathbf{e}_i$, then λ_i^r is a characteristic number of $\mathbf{a}^r$, with the same characteristic vector $\mathbf{e}_i$.* For a symmetric matrix of order n, there are exactly n linearly independent characteristic vectors. Hence it follows in this case that $\mathbf{a}^r$ cannot possess *additional* characteristic numbers or vectors.

Next, consider any polynomial in $\mathbf{a}$, of degree m, of the form

$$P(\mathbf{a}) = \alpha_0 \mathbf{a}^m + \alpha_1 \mathbf{a}^{m-1} + \cdots + \alpha_{m-1}\mathbf{a} + \alpha_m \mathbf{I}. \tag{199}$$

If we consider the product of the matrix $P(\mathbf{a})$ with any characteristic vector of $\mathbf{a}$, and use (197), we obtain the relation

$$P(\mathbf{a})\mathbf{e}_i = \alpha_0 \lambda_i^m \mathbf{e}_i + \alpha_1 \lambda_i^{m-1} \mathbf{e}_i + \cdots + \alpha_{m-1}\lambda_i \mathbf{e}_i + \alpha_m \mathbf{e}_i$$

or

$$P(\mathbf{a})\mathbf{e}_i = P(\lambda_i)\mathbf{e}_i. \tag{200}$$

Hence it follows that the equation

$$[P(\mathbf{a}) - \mu\,\mathbf{I}]\,\mathbf{x} = \mathbf{0} \tag{201}$$

possesses a nontrivial solution when $\mu = P(\lambda_i)$, and that a solution of (201) in this case is an arbitrary multiple of $\mathbf{e}_i$. But, as in the preceding argument, no additional solutions of (201) can exist.

Thus it follows that *if* $\mathbf{a}$ *is symmetric, then* $P(\mathbf{a})$ *has the same characteristic vectors as* $\mathbf{a}$, and also, *if the characteristic numbers of* $\mathbf{a}$ *are* $\lambda_1, \ldots, \lambda_n$, *then those of* $P(\mathbf{a})$ *are* $P(\lambda_1), \ldots, P(\lambda_n)$.

Let the determinant $|\mathbf{a} - \lambda \mathbf{I}|$, the vanishing of which determines the characteristic numbers of $\mathbf{a}$, be denoted by $F(\lambda)$:

$$F(\lambda) = |\mathbf{a} - \lambda \mathbf{I}|. \tag{202}$$

Then $F(\lambda)$ is a polynomial, of degree n in λ, which vanishes when $\lambda = \lambda_i$,

$$F(\lambda_i) = 0 \qquad (i = 1, 2, \ldots, n). \tag{203}$$

If now we identify the polynomial P with F, equation (200) becomes

$$F(\mathbf{a})\mathbf{e}_i = \mathbf{0} \qquad (i = 1, 2, \ldots, n). \tag{204}$$

Thus if we write temporarily $\mathbf{b} \equiv F(\mathbf{a})$, it follows that the equation $\mathbf{b}\,\mathbf{x} = \mathbf{0}$ possesses the n linearly independent solutions $\mathbf{x} = \mathbf{e}_1, \ldots, \mathbf{e}_n$. But since $\mathbf{b}$ is a symmetric matrix of order n, the results of Section 1.8 show that $\mathbf{b}$ must be of rank $n - n = 0$. Hence $\mathbf{b} = F(\mathbf{a})$ must be the *zero matrix*, and it follows that

$$F(\mathbf{a}) = \mathbf{0}. \tag{205}$$

That is, *if the characteristic equation of a symmetric matrix* $\mathbf{a}$ *is* $F(\lambda) = 0$, *then the matrix* $\mathbf{a}$ *satisfies the equation* $F(\mathbf{a}) = \mathbf{0}$.

This curious and useful result is known as the *Cayley-Hamilton theorem*, and is often stated briefly as follows: "A matrix satisfies its characteristic equation."

It is important to notice that in deducing (205) from (204) we have made use of the fact that a *symmetric* matrix always has n linearly independent characteristic vectors. Since this statement does not apply to nonsymmetric matrices with repeated characteristic numbers, the preceding proof does not apply in such cases. However, it can be proved by somewhat less direct methods that *the Cayley-Hamilton theorem is true for any square matrix*.

If $F(\lambda)$ possesses a factor $(\lambda - \lambda_r)^s$, where $s > 1$, so that λ_r is of multiplicity s, the same argument shows that the *symmetric* matrix $\mathbf{a}$ also satisfies the *reduced characteristic equation* $G(\mathbf{a}) = \mathbf{0}$, where $G(\lambda) = F(\lambda)/(\lambda - \lambda_r)^{s-1}$. (The matrix $\mathbf{a}$ considered in Section 1.14 may be used as an example.) This statement is *not* necessarily true

if $\mathbf{a}$ is nonsymmetric, as may be illustrated by the matrix considered on page 32.

As a verification of the theorem, we notice that, for the matrix

$$\mathbf{a} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \tag{206}$$

we have

$$F(\lambda) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3, \tag{207}$$

and the equation $\mathbf{a}^2 - 4\mathbf{a} + 3\mathbf{I} = \mathbf{0}$ becomes

$$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} - \begin{bmatrix} 8 & 4 \\ 4 & 8 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

We notice that this theorem permits any power of a matrix $\mathbf{a}$, and hence *any polynomial* in $\mathbf{a}$, to be expressed as a linear combination of the matrices $\mathbf{I}, \mathbf{a}, \mathbf{a}^2, \ldots, \mathbf{a}^{n-1}$, where n is the order of $\mathbf{a}$.

Thus, for the matrix (206) considered above, we have the successive results

$$\begin{aligned} \mathbf{a}^2 &= 4\mathbf{a} - 3\mathbf{I}, \\ \mathbf{a}^3 &= 4\mathbf{a}^2 - 3\mathbf{a} = 4(4\mathbf{a} - 3\mathbf{I}) - 3\mathbf{a} = 13\mathbf{a} - 12\mathbf{I}, \end{aligned} \tag{208}$$

and so forth. In addition, we obtain the relation $\mathbf{a} - 4\mathbf{I} + 3\mathbf{a}^{-1} = \mathbf{0}$. Hence we deduce that

$$\mathbf{a}^{-1} = -\tfrac{1}{3}\mathbf{a} + \tfrac{4}{3}\mathbf{I},$$

and obtain successive *negative* integral powers of $\mathbf{a}$ by successive multiplications and simplifications.

A convenient determination of the constants of combination, in the case of a general polynomial, is afforded by a result next to be obtained. In place of determining the constants involved in the representation $P(\mathbf{a}) = c_1\mathbf{a}^{n-1} + c_2\mathbf{a}^{n-2} + \ldots + c_n\mathbf{I}$, it is desirable for present purposes to assume the equivalent form

$$\begin{aligned} P(\mathbf{a}) = {} & C_1[(\mathbf{a} - \lambda_2\mathbf{I})(\mathbf{a} - \lambda_3\mathbf{I}) \cdots (\mathbf{a} - \lambda_n\mathbf{I})] \\ & + C_2[(\mathbf{a} - \lambda_1\mathbf{I})(\mathbf{a} - \lambda_3\mathbf{I}) \cdots (\mathbf{a} - \lambda_n\mathbf{I})] + \cdots \\ & + C_n[(\mathbf{a} - \lambda_1\mathbf{I})(\mathbf{a} - \lambda_2\mathbf{I}) \cdots (\mathbf{a} - \lambda_{n-1}\mathbf{I})], \end{aligned} \tag{209}$$

where each bracketed quantity, and hence also the complete right-hand side, is clearly a polynomial of degree $n - 1$ in $\mathbf{a}$. To determine the n C's, we postmultiply the equal members of (209) successively by each of the n characteristic vectors $\mathbf{e}_1, \cdots, \mathbf{e}_n$ of the matrix $\mathbf{a}$.

If both members are postmultiplied by $\mathbf{e}_k$, and use is made of the relation $\mathbf{a}\,\mathbf{e}_k = \lambda_k \mathbf{e}_k$, it is found that the coefficients of all C's except C_k then contain the factor $(\lambda_k - \lambda_k)$, and hence vanish. Thus there follows, after a simple calculation,

$$P(\mathbf{a})\mathbf{e}_k = C_k[(\lambda_k - \lambda_1) \cdots (\lambda_k - \lambda_{k-1})(\lambda_k - \lambda_{k+1}) \cdots (\lambda_k - \lambda_n)]\mathbf{e}_k, \tag{210}$$

for $k = 1, 2, \cdots, n$. But reference to equation (200) then shows that the coefficient of $\mathbf{e}_k$ on the right must be equal to $P(\lambda_k)$. Thus *if the characteristic numbers of the matrix* $\mathbf{a}$ *are all distinct*, we obtain the result

$$C_k = \frac{P(\lambda_k)}{\prod\limits_{r \neq k} (\lambda_k - \lambda_r)} \qquad (k = 1, 2, \cdots, n), \tag{211}$$

where the notation $\prod\limits_{r \neq k}$ denotes the product of those factors for which r takes on the values 1 through n, excluding k. If this result is introduced into (209), the desired representation is obtained in the form

$$P(\mathbf{a}) = \sum_{k=1}^{n} P(\lambda_k) Z_k(\mathbf{a}), \tag{212}$$

with the convenient abbreviation

$$Z_k(\mathbf{a}) = \frac{\prod\limits_{r \neq k} (\mathbf{a} - \lambda_r \mathbf{I})}{\prod\limits_{r \neq k} (\lambda_k - \lambda_r)} \qquad (k = 1, 2, \cdots, n). \tag{213}$$

Cases in which certain characteristic numbers are repeated require special treatment.*

* See Reference 1. It can be shown that this representation (with appropriate modifications for repeated characteristic numbers) is valid for *any* square matrix $\mathbf{a}$.

To verify this result in the case of the matrix (206), we notice that $\lambda_1 = 3$ and $\lambda_2 = 1$. To evaluate $P(\mathbf{a}) = \mathbf{a}^3$, we first calculate

$$Z_1 = \frac{\mathbf{a} - \lambda_2\mathbf{I}}{3 - 1} = \frac{1}{2}(\mathbf{a} - \mathbf{I}), \qquad Z_2 = \frac{\mathbf{a} - \lambda_1\mathbf{I}}{1 - 3} = -\frac{1}{2}(\mathbf{a} - 3\mathbf{I}).$$

Hence, with $P(3) = 27$ and $P(1) = 1$, there follows

$$\mathbf{a}^3 = \tfrac{27}{2}(\mathbf{a} - \mathbf{I}) - \tfrac{1}{2}(\mathbf{a} - 3\mathbf{I}) = 13\mathbf{a} - 12\mathbf{I},$$

in accordance with (208). The usefulness of (212) would clearly be better illustrated in the calculation of $\mathbf{a}^{100}$.

It should be noticed that the quantities Z_k depend only on $\mathbf{a}$, and are *not* dependent upon the form of the polynomial P chosen. The result (212) is known as *Sylvester's formula*.

Having defined polynomial functions, we may next define other functions of $\mathbf{a}$ by *infinite series* such as

$$\sum_{m=0}^{\infty} \alpha_m \mathbf{a}^m = \lim_{M \to \infty} \sum_{m=0}^{M} \alpha_m \mathbf{a}^m, \tag{214}$$

for those matrices for which the indicated limit exists. We omit discussion of the convergence of such series. However, if $\mathbf{a}$ is of order n, it is clear that the sum of M terms of the series can be expressed as a polynomial of maximum degree $n - 1$ in $\mathbf{a}$, regardless of the value of M, in consequence of the preceding results. Hence we see that *if the series converges, the function represented by the series must also be so expressible, and hence must be determinable from* (212) *if the characteristic numbers of* $\mathbf{a}$ *are distinct.*

In particular, it can be shown that the series

$$e^{\mathbf{a}} = \sum_{m=0}^{\infty} \frac{\mathbf{a}^m}{m!} \tag{215}$$

converges for *any* square matrix $\mathbf{a}$. Suppose that $\mathbf{a}$ is a matrix of order *two*, with distinct characteristic numbers λ_1 and λ_2. Then (213) gives

$$Z_1 = \frac{\mathbf{a} - \lambda_2\mathbf{I}}{\lambda_1 - \lambda_2}, \qquad Z_2 = \frac{\mathbf{a} - \lambda_1\mathbf{I}}{\lambda_2 - \lambda_1},$$

and from (212), we obtain the evaluation

$$e^{\mathbf{a}} = \frac{1}{\lambda_1 - \lambda_2}[(e^{\lambda_1} - e^{\lambda_2})\mathbf{a} - (\lambda_2 e^{\lambda_1} - \lambda_1 e^{\lambda_2})\mathbf{I}]. \tag{216}$$

The corresponding evaluation when $\lambda_3 = \lambda_2$ can be obtained from this result as the limiting form when $\lambda_2 \to \lambda_1$ (see Problem 55).

1.23. Numerical solution of characteristic-value problems. In the process of dealing with a characteristic-value problem of the form

$$\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}, \tag{217}$$

it is necessary first to determine roots of the characteristic equation

$$|\,\mathbf{a} - \lambda\,\mathbf{I}\,| = 0, \tag{218}$$

and then, for each such value of λ, to obtain a nontrivial solution vector of (217). If $\mathbf{a}$ is of order n, equation (218) is an algebraic equation of the same degree in λ, and the numerical determination of the characteristic numbers generally involves considerable labor when $n > 2$. Further, the actual *expansion* of (218) may be tedious in such cases.

In this section we outline a numerical iterative method which avoids these steps, and which is frequently useful in practice. This method is analogous to the method, associated with the names of *Vianello* and *Stodola*, which is applied to corresponding problems involving differential equations.*

Suppose first that the *dominant* characteristic number, that is, the characteristic number with largest magnitude, is required. To initiate the procedure, we choose an initial approximation to the corresponding characteristic vector, say $\mathbf{x}^{(1)}$. In the absence of advance knowledge as to the nature of this vector, we may, for example, start with the vector $\{0, 0, \ldots, 1\}$ or $\{1, 1, \ldots, 1\}$. This initial approximation is then introduced into the *left*-hand member of (217). If we then set

$$\mathbf{y}^{(1)} = \mathbf{a}\,\mathbf{x}^{(1)}, \tag{219}$$

the requirement that (217) be approximately satisfied becomes

$$\mathbf{y}^{(1)} \approx \lambda\,\mathbf{x}^{(1)}. \tag{220}$$

If the respective components of $\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ are nearly in a constant ratio, we may expect that the approximation $\mathbf{x}^{(1)}$ is good, and that this ratio is an approximation to the true value of λ.

It is conventional to choose $\mathbf{x}^{(1)}$ in such a way that one com-

* See Reference 8.

ponent is *unity*, and to choose, as a first approximation to the dominant characteristic value of λ, the corresponding component of $\mathbf{y}^{(1)}$. A more efficient determination is outlined in the following section [equations (232a,b)].

A convenient multiple of $\mathbf{y}^{(1)}$ is then taken as the next approximation $\mathbf{x}^{(2)}$, and the process is repeated until satisfactory agreement between successive approximations is obtained. As will be shown, in the case when $\mathbf{a}$ is real and symmetric, this method will lead inevitably to the *dominant* characteristic value of λ and to the corresponding characteristic vector, unless the vector $\mathbf{x}^{(1)}$ happens to be exactly orthogonal to that vector, except in the unusual case when the *negative* of the dominant characteristic number is *also* a characteristic number.

If the *smallest* value of λ is required, we first transform (217) to the equation

$$\mathbf{x} = \lambda\, \mathbf{a}^{-1}\, \mathbf{x}.$$

With the notations

$$\mathbf{b} = \mathbf{a}^{-1}, \qquad \kappa = \frac{1}{\lambda} \tag{221a,b}$$

this equation takes the form

$$\mathbf{b}\, \mathbf{x} = \kappa\, \mathbf{x}. \tag{222}$$

The *largest* characteristic value of κ for this equation can then be determined by the iterative method, and is then the reciprocal of the *smallest* characteristic value of λ for (217).

This inversion clearly fails if $\mathbf{a}$ is *singular*, that is, if $\lambda = 0$ is a characteristic number of (217). A method which is useful in this case is presented in the following section (page 73).

To illustrate the basic procedure, we seek the largest characteristic value of λ for the system

$$\left.\begin{aligned} x_1 + \ \ x_2 + \ \ x_3 &= \lambda\, x_1, \\ x_1 + 2x_2 + 2x_3 &= \lambda\, x_2, \\ x_1 + 2x_2 + 3x_3 &= \lambda\, x_3 \end{aligned}\right\}. \tag{223}$$

With the initial approximation $\mathbf{x}^{(1)} = \{1, 1, 1\}$, there follows

$$\mathbf{y}^{(1)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 3 \\ 5 \\ 6 \end{Bmatrix} = 6 \begin{Bmatrix} \frac{1}{2} \\ \frac{5}{6} \\ 1 \end{Bmatrix}. \tag{224}$$

If we determine λ such that the x_3-components of $\lambda\,\mathbf{x}^{(1)}$ and $\mathbf{y}^{(1)}$ are equal, we have $\lambda^{(1)} = 6$. Next, with $\mathbf{x}^{(2)} = \{\frac{1}{2},\ \frac{5}{6},\ 1\}$, there follows

$$\mathbf{y}^{(2)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{Bmatrix} \frac{1}{2} \\ \frac{5}{6} \\ 1 \end{Bmatrix} = \begin{Bmatrix} \frac{7}{3} \\ \frac{25}{6} \\ \frac{31}{6} \end{Bmatrix} = \frac{31}{6} \begin{Bmatrix} \frac{14}{31} \\ \frac{25}{31} \\ 1 \end{Bmatrix}. \tag{225}$$

The second approximation to the dominant characteristic number is then $\lambda^{(2)} = \frac{31}{6} \doteq 5.17$. The third step then gives

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{Bmatrix} \frac{14}{31} \\ \frac{25}{31} \\ 1 \end{Bmatrix} = \begin{Bmatrix} \frac{70}{31} \\ \frac{128}{31} \\ \frac{157}{31} \end{Bmatrix} = \frac{157}{31} \begin{Bmatrix} \frac{70}{157} \\ \frac{128}{157} \\ 1 \end{Bmatrix} \tag{226}$$

and also $\lambda^{(3)} = \frac{157}{31} \doteq 5.06$. The ratios $x_1 : x_2 : x_3$ according to the four approximations are (1:1:1), (0.500:0.833:1), and (0.446: 0.803:1). The next cycle leads to the value $\lambda^{(4)} \doteq 5.05$, and to the ratios 0.445:0.802:1, which hence may be expected to be accurate to three significant figures.

1.24. Additional techniques. In order to improve and extend the procedure just outlined, in the case when $\mathbf{a}$ is *real and symmetric*, it is desirable to consider the analytical basis of the procedure in that case.* For this purpose, we may suppose that $\mathbf{e}_1$, $\mathbf{e}_2$, . . . , $\mathbf{e}_n$ are the true orthogonalized characteristic unit vectors of the problem (217), corresponding to the characteristic numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$, arranged in increasing order of magnitude. If the initial assumption $\mathbf{x}^{(1)}$ is imagined to be expressed in the form

$$\mathbf{x}^{(1)} = \sum_{k=1}^{n} c_k \mathbf{e}_k, \tag{227a}$$

then the vector $\mathbf{y}^{(1)} = \mathbf{a}\,\mathbf{x}^{(1)}$ must accordingly be given by

$$\mathbf{y}^{(1)} = \sum_{k=1}^{n} c_k \mathbf{a}\,\mathbf{e}_k = \sum_{k=1}^{n} \lambda_k c_k \mathbf{e}_k. \tag{227b}$$

Next, if a multiple of $\mathbf{y}^{(1)}$, say $\alpha\,\mathbf{y}^{(1)}$, is taken to be $\mathbf{x}^{(2)}$, there then follows similarly

$$\mathbf{x}^{(2)} = \alpha \sum_{k=1}^{n} \lambda_k c_k \mathbf{e}_k, \qquad \mathbf{y}^{(2)} = \alpha \sum_{k=1}^{n} \lambda_k^2 c_k \mathbf{e}_k. \tag{228a,b}$$

* Certain other cases are considered in Sections 1.25 and 1.26.

More generally, after r steps we have

$$\mathbf{x}^{(r)} = \beta \sum_{k=1}^{n} \lambda_k^{r-1} c_k \mathbf{e}_k = \beta \lambda_n^{r-1} \sum_{k=1}^{n} \left(\frac{\lambda_k}{\lambda_n}\right)^{r-1} c_k \mathbf{e}_k$$

or

$$\mathbf{x}^{(r)} = \beta \lambda_n^{r-1} \left[c_n \mathbf{e}_n + \left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{r-1} c_{n-1} \mathbf{e}_{n-1} + \cdots + \left(\frac{\lambda_1}{\lambda_n}\right)^{r-1} c_1 \mathbf{e}_1 \right] \tag{229a}$$

and, correspondingly,

$$\mathbf{y}^{(r)} = \beta \lambda_n^{r} \left[c_n \mathbf{e}_n + \left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{r} c_{n-1} \mathbf{e}_{n-1} + \cdots + \left(\frac{\lambda_1}{\lambda_n}\right)^{r} c_1 \mathbf{e}_1 \right]. \tag{229b}$$

Since λ_n is the dominant characteristic number, the powers $(\lambda_k/\lambda_n)^r$ tend to zero when $k \neq n$, and the expressions tend to multiples of $\mathbf{e}_n$ as r increases except in the very special case when the initial assumption $\mathbf{x}^{(1)}$ happens to be exactly orthogonal to $\mathbf{e}_n$, so that $c_n = 0$.

If λ_n is a multiple root of the characteristic equation, it is easily seen that the process will still lead to *one* corresponding characteristic vector. The case when λ_n and $-\lambda_n$ are both characteristic numbers requires special treatment.* However, in most practical cases the characteristic numbers are all nonnegative.

The *rate* of convergence of the method clearly depends upon the magnitude of the ratio of the two largest characteristic numbers. In case this ratio is near unity, and the convergence rate is slow, the matrix $\mathbf{a}$ may be first raised to an integral power p. The characteristic numbers of the new matrix $\mathbf{a}^p$ are then $\lambda_1^p, \ldots, \lambda_n^p$, and the ratio of the dominant and subdominant numbers is clearly increased.

We may notice from (229a,b) that, if at any stage of the iteration the true vector $\mathbf{e}_n$ were known, the condition

$$(\mathbf{e}_n, \mathbf{y}^{(r)}) = \lambda(\mathbf{e}_n, \mathbf{x}^{(r)}) \tag{230}$$

* It is apparent from (229a) that if $\lambda_{n-1} = -\lambda_n$ the sequence $\mathbf{x}^{(1)}, \mathbf{x}^{(3)}, \ldots$ converges to a multiple of $c_n\mathbf{e}_n + c_{n-1}\mathbf{e}_{n-1}$, whereas the sequence $\mathbf{x}^{(2)}, \mathbf{x}^{(4)}, \ldots$ converges to a multiple of $c_n\mathbf{e}_n - c_{n-1}\mathbf{e}_{n-1}$ (see Problem 62).

would lead to the relation

$$\beta \lambda_n{}^r c_n = \lambda \beta \lambda_n{}^{r-1} c_n \qquad \text{or} \qquad \lambda = \lambda_n, \tag{231}$$

and hence would determine λ_n *exactly*. Clearly, any multiple of $\mathbf{e}_n$ would serve the same purpose. Thus it may be expected that a reasonably good approximation to λ_n would be obtained by replacing $\mathbf{e}_n$ by a convenient multiple of either the approximation $\mathbf{x}^{(r)}$ or the better approximation $\mathbf{y}^{(r)}$ in (230). This procedure gives the alternative formulas

$$(\mathbf{x}^{(r)}, \mathbf{y}^{(r)}) \approx \lambda_n (\mathbf{x}^{(r)}, \mathbf{x}^{(r)}) \tag{232a}$$

or

$$(\mathbf{y}^{(r)}, \mathbf{y}^{(r)}) \approx \lambda_n (\mathbf{x}^{(r)}, \mathbf{y}^{(r)}), \tag{232b}$$

of which the second is in general the more nearly accurate. It can be shown that the approximation given by (232a) is always conservative in absolute value (when $\mathbf{a}$ is symmetric). The same is true of that given by (232b) if the matrix $\mathbf{a}$ is also positive definite (see Problem 79).

We list in the following table the results of applying (A) the preceding method, (B) the formula of (232a), and (C) the formula of (232b), to the illustrative example:

r	(A)	(B)	(C)
1	6.000	4.667	5.000
2	5.167	5.043	5.048
3	5.065	5.049	5.049
4	5.051	5.049	5.049

It may be seen that if (232a) or (232b) is used, the successive approximations to λ_n converge more rapidly than do the approximations to $\mathbf{e}_n$. This statement is generally true. Thus these formulas are useful in those cases when an accurate value of the dominant characteristic number is required, but comparable accuracy in the determination of the corresponding characteristic vector is not needed.

To obtain the *smallest* characteristic number of (223), we write $\kappa = 1/\lambda$, resolve the equations in the form

$$\left.\begin{aligned} 2x_1 - x_2 &= \kappa x_1, \\ -x_1 + 2x_2 - x_3 &= \kappa x_2, \\ -x_2 + x_3 &= \kappa x_3 \end{aligned}\right\}, \tag{233}$$

and determine the largest characteristic value of κ by the preceding methods.

Suppose now that *one* characteristic vector, say one which corresponds to a dominant characteristic number, is known *exactly*. Then for a *symmetric* matrix, all other characteristic vectors may be considered to be orthogonal to $\mathbf{e}_n$.* Hence, if we impose the constraint

$$(\mathbf{e}_n, \mathbf{x}) = 0 \tag{234}$$

on the problem (217), the resultant problem will possess those characteristic numbers and corresponding characteristic vectors which are in addition to λ_n and $\mathbf{e}_n$. But (234) permits one of the components, say x_r, to be expressed as a linear combination of the others. Hence we may eliminate x_r from the scalar equations corresponding to (217), disregard the rth resulting equation, and obtain a set of $n - 1$ equations involving only $n - 1$ components. The dominant characteristic number, and a corresponding characteristic vector, are then obtained as before, the component x_r being determined finally from (234).

Whereas the coefficient matrix associated with the new set of $n - 1$ equations is generally nonsymmetric, the convergence of the iterative method is assured in this case by results to be obtained in Section 1.26.

In particular, in the case when $|\mathbf{a}| = 0$ so that $\lambda = 0$ is a characteristic number of $\mathbf{a}$, we may replace $\mathbf{e}_n$ in (234) by the corresponding characteristic vector. Unless $\lambda = 0$ is of multiplicity greater than one, the corresponding reduced set of equations can then be inverted for the purpose of determining the smallest nonzero characteristic number. In the more general case, a number of unknowns equal to the multiplicity of the number $\lambda = 0$ must be eliminated in this way.

The procedure may be repeated until the solution is concluded or until only two components remain, at which stage the characteristic equation is quadratic in λ and the analysis can be conveniently completed without matrix iteration. Thus, if $\mathbf{a}$ is of order *three*, only one iterative process is needed. If $\mathbf{a}$ is of order *four*, we may conveniently determine the largest and smallest characteristic

* If the characteristic numbers are distinct, this *must* be so; otherwise, we may impose this condition without loss of generality.

numbers and their corresponding vectors. The conditions $(\mathbf{e}_1, \mathbf{x}) = 0$ and $(\mathbf{e}_4, \mathbf{x}) = 0$ then permit the elimination of two components, and the reduction of the problem to one involving only two components.

In practice, the determination of the primary characteristic vector is only approximately effected. It is found that the numerical determination of a subdominant characteristic vector will often involve repeated subtraction of nearly equal quantities, particularly if the two relevant characteristic numbers are nearly equal. In such cases, it may be necessary to calculate the components of the dominant characteristic vector to a degree of accuracy much higher than that required for the subdominant characteristic vector.

To illustrate the reduction in the preceding example, we notice that the dominant characteristic vector is given by $\{0.445, 0.802, 1\}$ to three significant figures. Hence (234) here becomes

$$0.445x_1 + 0.802x_2 + x_3 = 0. \tag{235}$$

If we eliminate x_3 between (235) and (223), and notice that the third equation is then a consequence of the first two (to the three significant figures retained), we obtain the reduced problem

$$\left.\begin{aligned} 0.555x_1 + 0.198x_2 &= \lambda x_1, \\ 0.110x_1 + 0.396x_2 &= \lambda x_2 \end{aligned}\right\}. \tag{236}$$

The dominant characteristic number of (236), and the two components of the corresponding characteristic vector, can then be obtained by matrix iteration, if this is desired, the component x_3 being determined in terms of them by (235). Otherwise, since the characteristic equation of (236) is quadratic, that equation can be solved by the quadratic formula, and the ratio of the x_1 and x_2 components of the corresponding characteristic vectors can be obtained directly.

1.25. Generalized characteristic-value problems. In certain fields we encounter characteristic-value problems of the more general form

$$\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}, \tag{237}$$

where $\mathbf{a}$ and $\mathbf{b}$ are real square matrices of order n. Such a problem reduces to the type considered previously when $\mathbf{b} = \mathbf{I}$. The charac-

teristic equation corresponding to (237) is of the form

$$| \mathbf{a} - \lambda \mathbf{b} | = 0. \tag{238}$$

In the important practical cases in which both $\mathbf{a}$ and $\mathbf{b}$ are *symmetric*, so that $\mathbf{a}^T = \mathbf{a}$ and $\mathbf{b}^T = \mathbf{b}$, we next establish a useful generalization of the results of Section 1.11. If λ_1 and λ_2 are distinct characteristic numbers corresponding, respectively, to the characteristic vectors $\mathbf{e}_1$ and $\mathbf{e}_2$, there follows

$$\mathbf{a}\,\mathbf{e}_1 = \lambda_1 \mathbf{b}\,\mathbf{e}_1, \qquad \mathbf{a}\,\mathbf{e}_2 = \lambda_2 \mathbf{b}\,\mathbf{e}_2$$

and hence also

$$(\mathbf{a}\,\mathbf{e}_1)^T\,\mathbf{e}_2 = \lambda_1 (\mathbf{b}\,\mathbf{e}_1)^T\,\mathbf{e}_2, \qquad \mathbf{e}_1{}^T\,\mathbf{a}\,\mathbf{e}_2 = \lambda_2 \mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2,$$

or, making use of the symmetry in $\mathbf{a}$ and $\mathbf{b}$,

$$\mathbf{e}_1{}^T\,\mathbf{a}\,\mathbf{e}_2 = \lambda_1 \mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2, \qquad \mathbf{e}_1{}^T\,\mathbf{a}\,\mathbf{e}_2 = \lambda_2 \mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2. \tag{239}$$

By subtracting the first equation from the second in (239), we then obtain the relation

$$(\lambda_2 - \lambda_1)\mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2 = 0. \tag{240}$$

Thus since $\lambda_1 \neq \lambda_2$ by assumption, we conclude that $\mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2 = 0$. That is, *if* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *are characteristic vectors, corresponding to two distinct characteristic numbers of the problem* $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$, *where* $\mathbf{a}$ *and* $\mathbf{b}$ *are symmetric, there follows*

$$\mathbf{e}_1{}^T\,\mathbf{b}\,\mathbf{e}_2 = 0. \tag{241}$$

It is convenient to speak of the left-hand member of (241) as the *scalar product of* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *relative to* $\mathbf{b}$, and to say that when (241) is satisfied *the vectors* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *are orthogonal relative to* $\mathbf{b}$. The ordinary type of orthogonality is thus relative to the *unit matrix* I.

In consequence of (241) and (239), we deduce that *the vectors* $\mathbf{e}_1$ *and* $\mathbf{e}_2$ *are also orthogonal relative to the matrix* $\mathbf{a}$.

The left-hand member of (241) is conveniently denoted by $(\mathbf{e}_1, \mathbf{e}_2)_\mathbf{b}$. More generally, we write

$$(\mathbf{u}, \mathbf{v})_\mathbf{b} \equiv \mathbf{u}^T\,\mathbf{b}\,\mathbf{v} = \mathbf{v}^T\,\mathbf{b}\,\mathbf{u} \tag{242}$$

for the scalar product of $\mathbf{u}$ and $\mathbf{v}$ relative to a *symmetric* matrix $\mathbf{b}$. In particular, when $\mathbf{v} = \mathbf{u}$ we define the product

$$|_\mathbf{b}{}^2 \equiv (\mathbf{u}, \mathbf{u})_\mathbf{b} \equiv \mathbf{u}^T\,\mathbf{b}\,\mathbf{u} \tag{243}$$

to be the square of the generalized *length* of $\mathbf{u}$, relative to $\mathbf{b}$. In order that this quantity be necessarily *positive* except only when $\mathbf{u}$ is the *zero vector*, the matrix $\mathbf{b}$ must be *positive definite*. This is the case which most frequently arises in practice.*

In the remainder of this section, we assume that $\mathbf{a}$ *is real and symmetric and* $\mathbf{b}$ *real, symmetric, and positive definite.* In particular, this implies that $\mathbf{b}$ is *nonsingular*. The generalized length of a vector, relative to $\mathbf{b}$, is then real and positive unless the vector is a zero vector, in which case its generalized length is zero.

By a method analogous to that used in Section 1.11, it is then easily shown that *the characteristic numbers of* (237) *are real*. Further, by an argument similar to that used in Section 1.21, it can be shown that *to a characteristic number of multiplicity* s *there correspond* s *linearly independent characteristic vectors*. Then, by methods completely analogous to those of Section 1.12, this set can be orthogonalized relative to $\mathbf{b}$, and normalized in such a way that each vector possesses generalized length unity. It is seen that the condition $|\mathbf{b}| \neq 0$ guarantees that the characteristic equation (238) be of degree n. Hence, in the case under consideration, we may always obtain a set of n mutually orthogonal unit characteristic vectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$, such that

$$(\mathbf{e}_i, \mathbf{e}_j)_{\mathbf{b}} = \delta_{ij}. \tag{244}$$

The *normalized modal matrix* $\mathbf{M}$, associated with (237), may now be defined as the matrix having the components of the kth vector of the set as the elements of its kth *column*. Then in consequence of the relation

$$\mathbf{a}\,\mathbf{e}_i = \lambda_i \mathbf{b}\,\mathbf{e}_i \qquad (i = 1, 2, \cdots, n),$$

there follows

$$\mathbf{a}\,\mathbf{M} = \mathbf{b}\,\mathbf{M}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \equiv \mathbf{b}\,\mathbf{M}\,\mathbf{D}, \tag{245}$$

* Usually at least *one* of the matrices $\mathbf{a}$ and $\mathbf{b}$ is positive definite. If $\mathbf{a}$ is positive definite, we may replace λ by $1/\lambda'$ and interchange the roles of $\mathbf{a}$ and $\mathbf{b}$ throughout this section.

and also, in virtue of (244),

$$\mathbf{M}^T \mathbf{b}\, \mathbf{M} = \mathbf{I}. \tag{246}$$

(See Problem 24.)

We may now verify the fact that, with the change of variables

$$\mathbf{x} = \mathbf{M}\, \boldsymbol{\alpha}, \tag{247}$$

the *two* quadratic forms

$$A = \mathbf{x}^T \mathbf{a}\, \mathbf{x}, \qquad B = \mathbf{x}^T \mathbf{b}\, \mathbf{x} \tag{248a,b}$$

are reduced *simultaneously* to the canonical forms

$$A = \boldsymbol{\alpha}^T \mathbf{D}\, \boldsymbol{\alpha} = \lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 + \cdots + \lambda_n \alpha_n^2, \tag{249a}$$

$$B = \boldsymbol{\alpha}^T \boldsymbol{\alpha} = \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2. \tag{249b}$$

For the substitution of (247) into (248b), and the use of (246), gives immediately

$$B = \boldsymbol{\alpha}^T \mathbf{M}^T \mathbf{b}\, \mathbf{M}\, \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \boldsymbol{\alpha},$$

in accordance with (249b), whereas the substitution of (247) into (248a) gives

$$A = \boldsymbol{\alpha}^T \mathbf{M}^T \mathbf{a}\, \mathbf{M}\, \boldsymbol{\alpha}$$

and the use of (245) leads to the result

$$A = \boldsymbol{\alpha}^T \mathbf{M}^T \mathbf{b}\, \mathbf{M}\, \mathbf{D}\, \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{D}\, \boldsymbol{\alpha},$$

in accordance with (249a).

From (246) it follows that

$$|\, \mathbf{M}\, | = \pm \frac{1}{\sqrt{|\, \mathbf{b}\, |}},$$

so that the transformation (247) is *nonsingular*. Further, since (246) leads to the relation $\mathbf{M}^{-1} = \mathbf{M}^T \mathbf{b}$, the inversion of (247) may be conveniently effected by use of the equation

$$\boldsymbol{\alpha} = \mathbf{M}^T \mathbf{b}\, \mathbf{x}. \tag{247'}$$

Equations (249a,b) are identical with equations (140) and (142) of Section 1.17. It is important to notice that the coefficients λ_i in (249) are the roots of the equation $|\, \mathbf{a} - \lambda\, \mathbf{b}\, | = 0$, and hence are *real*, under the present restrictions on $\mathbf{a}$ and $\mathbf{b}$.

If the matrices **a** *and* **b** *are both positive definite, the characteristic numbers* λ_i *are also necessarily positive.* This result is established by noticing that the relation

$$\mathbf{a}\,\mathbf{e}_i = \lambda_i \mathbf{b}\,\mathbf{e}_i$$

implies the relation

$$\mathbf{e}_i^T\,\mathbf{a}\,\mathbf{e}_i = \lambda_i \mathbf{e}_i^T\,\mathbf{b}\,\mathbf{e}_i.$$

Since both $\mathbf{e}_i^T\,\mathbf{a}\,\mathbf{e}_i$ and $\mathbf{e}_i^T\,\mathbf{b}\,\mathbf{e}_i$ are positive when **a** and **b** are positive definite (and $\mathbf{e}_i \neq \mathbf{0}$), the same is true of λ_i.

The preceding results will be of importance in Section 2.12 of the following chapter.

Any vector **v** in n-dimensional space can be expressed as a linear combination of the vectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$, of the form

$$\mathbf{v} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n = \sum_{k=1}^{n} c_k\mathbf{e}_k. \qquad (250a)$$

In order to evaluate any coefficient c_r, we merely form the generalized scalar product of $\mathbf{e}_r$ into both sides of (250a), and use (244) to obtain the result

$$c_r = (\mathbf{e}_r, \mathbf{v})_{\mathbf{b}} \equiv \mathbf{e}_r^T\,\mathbf{b}\,\mathbf{v} \qquad (r = 1, 2, \cdots, n). \qquad (250b)$$

The case of most common occurrence in practice is that in which **b** is a *diagonal matrix* **g**, say

$$\mathbf{g} = \begin{bmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & g_n \end{bmatrix} = [g_i\,\delta_{ij}] = [g_j\,\delta_{ij}], \qquad (251a)$$

so that the equations $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{g}\,\mathbf{x}$ take the special form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= \lambda\, g_1 x_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= \lambda\, g_n x_n \end{aligned}\right\}, \qquad (251b)$$

where $a_{ji} = a_{ij}$.

The generalized scalar product $(\mathbf{x}, \mathbf{y})_{\mathbf{g}}$ then takes the form

$$(\mathbf{x}, \mathbf{y})_{\mathbf{g}} = g_1x_1y_1 + g_2x_2y_2 + \cdots + g_nx_ny_n, \qquad (251c)$$

while the generalized length of $\mathbf{x}$ is given by

$$l_g^2 = (\mathbf{x}, \mathbf{x})_g = g_1 x_1^2 + g_2 x_2^2 + \cdots + g_n x_n^2. \tag{251d}$$

The condition that $\mathbf{g}$ be positive definite requires that the diagonal elements be positive:

$$g_i > 0 \qquad (i = 1, 2, \cdots, n). \tag{251e}$$

It may be noticed that in certain cases a set of equations of the matrix form $\mathbf{a}'\mathbf{x} = \lambda\mathbf{x}$, where $\mathbf{a}'$ is a *nonsymmetric* square matrix, can be reduced to a set of the matrix form $\mathbf{a}\mathbf{x} = \lambda\mathbf{g}\mathbf{x}$, where $\mathbf{a}$ is symmetric and $\mathbf{g}$ is a diagonal matrix with positive diagonal elements, by multiplying the ith equation of the original set by a suitably chosen *positive* constant g_i. When $\mathbf{a}'$ is of order *two*, this reduction is clearly always possible if $a'_{12}a'_{21} > 0$; it is possible in other cases only when the coefficients satisfy certain compatibility conditions. If and only if such a reduction is possible, $\mathbf{a}'$ can be expressed as a product $\mathbf{d}\,\mathbf{a}$, where $\mathbf{d} \equiv \mathbf{g}^{-1}$ is a *diagonal* matrix with positive diagonal elements, and $\mathbf{a}$ is symmetric.

To conclude this section, we indicate the extension of the numerical methods of the preceding sections to the treatment of a characteristic-value problem of the form

$$\mathbf{a}\mathbf{x} = \lambda\mathbf{b}\mathbf{x}, \tag{252}$$

where again $\mathbf{a}$ is a symmetric matrix of order n, and $\mathbf{b}$ is a positive definite, symmetric, matrix of the same order.

Since by assumption, $\mathbf{b}$ is nonsingular, equation (252) can be reduced to the form

$$\mathbf{b}^{-1}\mathbf{a}\mathbf{x} = \lambda\mathbf{x}, \tag{253}$$

which is of the type considered previously. However, the matrix $\mathbf{b}^{-1}\mathbf{a}$ will now *not* be symmetric, in general. In the case of (251b) the reduction to the form (253) involves only division of both sides of the ith equation by g_i.

In order to investigate the convergence of the iterative procedure in this case, let the normalized characteristic unit vectors be denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, corresponding, respectively, to $\lambda_1, \lambda_2, \ldots, \lambda_n$, so that

$$\mathbf{a}\,\mathbf{e}_r = \lambda_r \mathbf{b}\,\mathbf{e}_r. \tag{254}$$

The initial approximation $\mathbf{x}^{(1)}$ can then be imagined to be expressed as a linear combination of these vectors, in the form

$$\mathbf{x}^{(1)} = \sum_{k=1}^{n} c_k \mathbf{e}_k. \tag{255a}$$

Then if we denote $\mathbf{b}^{-1}\,\mathbf{a}\,\mathbf{x}^{(1)}$ by $\mathbf{y}^{(1)}$, there follows

$$\mathbf{y}^{(1)} \equiv \mathbf{b}^{-1}\,\mathbf{a}\,\mathbf{x}^{(1)} = \sum_{k=1}^{n} c_k \mathbf{b}^{-1}\,\mathbf{a}\,\mathbf{e}_k = \sum_{k=1}^{n} c_k \mathbf{b}^{-1} \lambda_k \mathbf{b}\,\mathbf{e}_k$$

or

$$\mathbf{y}^{(1)} \equiv \mathbf{b}^{-1}\,\mathbf{a}\,\mathbf{x}^{(1)} = \sum_{k=1}^{n} \lambda_k c_k \mathbf{e}_k. \tag{255b}$$

By comparing (255a,b) with (227a,b) of the preceding section, we see that the arguments presented in that section again apply here, to show that successive approximations will indeed converge to a multiple of the dominant vector $\mathbf{e}_n$.

In this case, however, it is seen that the requirement

$$(\mathbf{e}_n, \mathbf{y}^{(r)})_{\mathbf{b}} = \lambda (\mathbf{e}_n, \mathbf{x}^{(r)})_{\mathbf{b}},$$

in place of (230), would give $\lambda = \lambda_n$ exactly. Hence (232a,b) should here be modified to the alternative conditions

$$(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})_{\mathbf{b}} \approx \lambda_n (\mathbf{x}^{(r)}, \mathbf{x}^{(r)})_{\mathbf{b}} \tag{256a}$$

or, better,

$$(\mathbf{y}^{(r)}, \mathbf{y}^{(r)})_{\mathbf{b}} \approx \lambda_n (\mathbf{x}^{(r)}, \mathbf{y}^{(r)})_{\mathbf{b}}. \tag{256b}$$

Similarly, equation (234) must be replaced by the relation

$$(\mathbf{e}_n, \mathbf{x})_{\mathbf{b}} = 0, \tag{257}$$

which permits reduction of the order of the system when one characteristic vector has been obtained.

The same statements apply to the inversion of (253),

$$\frac{1}{\lambda}\mathbf{x} = \mathbf{a}^{-1}\,\mathbf{b}\,\mathbf{x},$$

which is used in determining the smallest characteristic value of λ when $\mathbf{a}$ is nonsingular.

1.26. Characteristic numbers of nonsymmetric matrices. Whereas the characteristic equation $|\mathbf{a} - \lambda \mathbf{I}| = 0$ of a non-

symmetric square matrix $\mathbf{a}$ of order n is of degree n, we have seen that when the roots of this equation are not *distinct* the total number of linearly independent characteristic vectors may be less than n. In the present section we exclude the exceptional cases, which rarely occur in practice, and suppose that the n characteristic numbers of $\mathbf{a}$ are *real and distinct*. The corresponding characteristic vectors are then linearly independent.

In order to establish this fact, we assume the contrary and deduce a contradiction. Suppose that the characteristic numbers $\lambda_1, \ldots, \lambda_n$ are all distinct, and denote the corresponding characteristic vectors by $\mathbf{e}_1, \ldots, \mathbf{e}_n$. We then have the relations $\mathbf{a}\,\mathbf{e}_i = \lambda_i \mathbf{e}_i$ for $i = 1, 2, \ldots, n$. *Assume* that the first r characteristic vectors are linearly independent, but that

$$\mathbf{e}_{r+1} = \sum_{k=1}^{r} c_k \mathbf{e}_k,$$

where at least one c_k is not zero. By premultiplying the equal members of this relation by $\mathbf{a}$, there then follows

$$\lambda_{r+1}\mathbf{e}_{r+1} = \sum_{k=1}^{r} c_k \lambda_k \mathbf{e}_k,$$

and hence also, by comparing these relations,

$$\sum_{k=1}^{r} c_k(\lambda_{r+1} - \lambda_k)\mathbf{e}_k = \mathbf{0}.$$

But, since $\mathbf{e}_1, \ldots, \mathbf{e}_r$ are linearly independent, the coefficient of *each* $\mathbf{e}_k$ must vanish. Since at least one c_k is not zero, at least one λ_k must equal λ_{r+1}, in contradiction with the assumption that the λ's are distinct.

In correspondence with the characteristic-value problem

$$\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}, \tag{258}$$

we may consider the problem

$$\mathbf{a}^T\,\mathbf{x}' = \lambda\,\mathbf{x}', \tag{258'}$$

associated with the *transpose* of $\mathbf{a}$. In virtue of the fact that the two matrices $[\mathbf{a} - \lambda\,\mathbf{I}]$ and $[\mathbf{a}^T - \lambda\,\mathbf{I}]$ differ only in that rows and columns are interchanged, their determinants possess the same

expansion, so that (258) *and* (258′) *possess the same characteristic numbers.* Let λ_1 and λ_2 denote any two distinct characteristic numbers, and let corresponding solutions of (258) and (258′) be denoted by $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}'_1$, $\mathbf{e}'_2$, respectively. We then have the relations

$$\mathbf{a}\,\mathbf{e}_1 = \lambda_1\mathbf{e}_1, \qquad \mathbf{a}^T\,\mathbf{e}'_2 = \lambda_2\mathbf{e}'_2,$$

from which there follows

$$(\lambda_2 - \lambda_1)\mathbf{e}_1{}^T\,\mathbf{e}'_2 = 0. \tag{259}$$

Hence we conclude that *any characteristic vector of* (258) *is orthogonal to any characteristic vector of* (258′) *which corresponds to a different characteristic number:*

$$(\mathbf{e}_i,\ \mathbf{e}'_j) = 0 \qquad (\lambda_i \neq \lambda_j). \tag{260}$$

This property permits the generalization of the methods of Sections 1.23 and 1.24 to the more general case considered here. While the problems considered in Section 1.25 are included in this generalization, the methods given in that section are usually preferable when they are applicable.

In particular, it is seen that the coefficients in the representation

$$\mathbf{v} = \sum_{k=1}^{n} c_k\mathbf{e}_k \tag{261a}$$

are determined by forming the scalar product of $\mathbf{e}'_k$ with the two members of this equation, in the form

$$c_k(\mathbf{e}_k,\ \mathbf{e}'_k) = (\mathbf{v},\ \mathbf{e}'_k). \tag{261b}$$

A development analogous to that of Section 24 then shows that the matrix iteration procedure again converges in this case to the characteristic vector corresponding to the dominant characteristic number of $\mathbf{a}$. In fact, such a development shows that the convergence of this procedure is insured if the matrix $\mathbf{a}$ possesses n linearly independent characteristic vectors, and only *real* characteristic numbers (which need not be *distinct*).* While formulas

* By a somewhat more involved analysis, which may be based on the generalized Sylvester formula mentioned in Section 1.22 (cf. Problem 54), it can be shown that the iterative procedure converges to the dominant characteristic number of any real matrix *if that number is real, and if no unequal characteristic number has equal absolute value.* If the dominant number is *repeated*, however,

analogous to (232a,b) can be devised for more accurate estimates of λ_n, their use involves a considerable increase in calculation.

The essential modification in procedure is involved in the calculation of subdominant characteristic quantities. In the more general case considered here, the constraint condition (234) must be replaced by the equation

$$(\mathbf{e}'_n, \mathbf{x}) = 0. \tag{262}$$

Thus after the (approximate) determination of the dominant characteristic number λ_n, and the corresponding vector solution $\mathbf{e}_n$, a vector $\mathbf{e}'_n$ satisfying the related equation $\mathbf{a}^T \mathbf{x}' = \lambda_n \mathbf{x}'$ must be determined (see Problem 71). The constraint (262), which corresponds to the fact that all other characteristic vectors of $\mathbf{a}$ are orthogonal to $\mathbf{e}'_n$, then permits the elimination of one of the unknowns in the system of equations (and the neglect of one of the resulting equations) so that the order of the system is reduced by unity.

1.27. A physical application. Several applications of the preceding methods will be found in the chapters which follow. In this section, we present one such application to a mechanical problem.

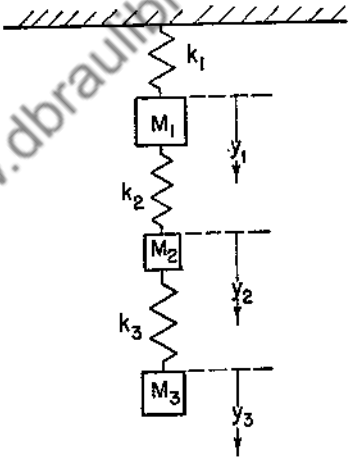


FIGURE 1.1

We consider the problem of determining the natural modes of free vibration of the mechanical system indicated in Figure 1.1, in which the masses M_1, M_2, and M_3 are connected in series to a fixed support, by linear springs with spring constants k_1, k_2, and k_3. The effects of viscous damping are neglected. If we denote the displacements of the respective masses from their equilibrium positions by $y_1(t)$, $y_2(t)$, and $y_3(t)$, respectively, the differential

the rate of convergence is not always *exponential* and the procedure is often not practical unless the multiplicity of that number is somehow known in advance. Modifications (similar to those of Problem 62) which are useful in this special situation, as well as in the case when the dominant numbers are conjugate complex, are given in Reference 1.

equations of motion are of the form

$$\left.\begin{aligned} M_1\frac{d^2y_1}{dt^2} &= k_2(y_2 - y_1) - k_1y_1 = -(k_1 + k_2)y_1 + k_2y_2, \\ M_2\frac{d^2y_2}{dt^2} &= k_3(y_3 - y_2) - k_2(y_2 - y_1) = k_2y_1 - (k_2 + k_3)y_2 + k_3y_3, \\ M_3\frac{d^2y_3}{dt^2} &= -k_3(y_3 - y_2) = k_3y_2 - k_3y_3 \end{aligned}\right\}. \tag{263}$$

The *natural modes* of vibration are those in which the masses oscillate in phase with a common frequency, and hence are specified by equations of the form

$$\begin{aligned} y_1(t) &= x_1 \sin(\omega t + \alpha), \\ y_2(t) &= x_2 \sin(\omega t + \alpha), \\ y_3(t) &= x_3 \sin(\omega t + \alpha), \end{aligned} \tag{264}$$

where the amplitudes x_1, x_2, and x_3 and the common circular frequency ω are to be determined. By introducing (264) into (263), and canceling the common resultant time factors, we obtain the equations

$$\left.\begin{aligned} (k_1 + k_2)x_1 - k_2x_2 &= M_1\omega^2 x_1, \\ -k_2x_1 + (k_2 + k_3)x_2 - k_3x_3 &= M_2\omega^2 x_2, \\ -k_3x_2 + k_3x_3 &= M_3\omega^2 x_3 \end{aligned}\right\}. \tag{265}$$

It should be noticed that the matrix of the coefficients in the left-hand members is *symmetric*. Also it is found that the "discriminants" Δ_m defined in Section 1.18 are of the form

$$\begin{aligned} \Delta_1 &= k_1 + k_2, \\ \Delta_2 &= k_1k_2 + k_2k_3 + k_3k_1, \\ \Delta_3 &= k_1k_2k_3, \end{aligned}$$

so that the matrix of coefficients is also *positive definite* when the spring constants are positive.

In the special case when $k_1 = k_2 = k_3 \equiv k$, and $M_1 = M_2 = M_3 \equiv M$, these equations reduce to equations (233) of Section 1.23

if we set

$$\kappa = \frac{1}{\lambda} = \frac{M\omega^2}{k}. \tag{266}$$

Hence the characteristic values of λ discussed in the example of that section are inversely proportional to the squares of the natural frequencies of the physical system under consideration, and the components of the characteristic vectors are in the same ratio as the three amplitudes x_1, x_2, and x_3 in a corresponding mode of vibration.

In the *fundamental mode*, corresponding to the *smallest* natural frequency, and hence to the *dominant* characteristic value of λ as defined by (266), the circular frequency is hence given by

$$\frac{M\omega_1^2}{k} = \frac{1}{5.05}: \qquad \omega_1 = 0.445\sqrt{\frac{k}{M}}.$$

Here the three masses all move in the same direction, the respective displacements from equilibrium at any instant being in the ratio $0.445:0.802:1$.

By completing the analysis indicated in Section 1.24, we find that in the *second* mode there follows

$$\omega_2 = 1.247\sqrt{\frac{k}{M}}.$$

In this mode the first two masses move in the same direction, whereas the third mass moves in the opposite direction, the displacements being in the ratio $-1.247:-0.555:1$. In the *third* mode there follows

$$\omega_3 = 1.802\sqrt{\frac{k}{M}}.$$

Here the first and third masses move in the same direction, and the second mass in the opposite direction, the displacements being in the ratio $1.802:-2.247:1$.

The most general motion of the system, possible in the absence of externally applied forces, is then a superposition of the three modes just described, in which the phase angle α of equation (264) may take on different values in the individual modes.

If the three masses are unequal, equations (265) are of the form of equations (251b), with g_i proportional to M_i. To illustrate the treatment of this case, we suppose that

$$k_1 = k_2 = k_3 \equiv k, \qquad M_1 = M_2 \equiv M, \qquad M_3 = 2M,$$

so that equations (265) become

$$\left.\begin{aligned} 2x_1 - x_2 &= \frac{M\omega^2}{k} x_1, \\ -x_1 + 2x_2 - x_3 &= \frac{M\omega^2}{k} x_2, \\ -x_2 + x_3 &= 2\frac{M\omega^2}{k} x_3 \end{aligned}\right\}.$$

In this case we may write

$$g_1 = 1, \qquad g_2 = 1, \qquad g_3 = 2.$$

In order to determine the *fundamental* mode *directly*, we must first *invert* these equations in the form

$$\left.\begin{aligned} x_1 + x_2 + 2x_3 &= \lambda x_1, \\ x_1 + 2x_2 + 4x_3 &= \lambda x_2, \\ x_1 + 2x_2 + 6x_3 &= \lambda x_3 \end{aligned}\right\},$$

where

$$\lambda = \frac{k}{M\omega^2}.$$

Except for refined successive estimates of $\lambda_3 = k/(M\omega_1^2)$, the calculation proceeds exactly as before. The results of successive steps are tabulated below to three significant figures:

	$\mathbf{x}^{(1)}$	$\mathbf{y}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{y}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{y}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{y}^{(4)}$	$\mathbf{x}^{(5)}$
x_1	1	4	0.444	3.22	0.402	3.15	0.399	3.15	0.399
x_2	1	7	0.777	6.00	0.750	5.90	0.747	5.89	0.747
x_3	1	9	1	8.00	1	7.90	1	7.89	1

Thus, after four cycles, the modal column {0.399, 0.747, 1} is repeated. The dominant characteristic value of λ is seen to be 7.89, so that the fundamental circular frequency of the physical system is

$$\omega_1 = 0.356\sqrt{\frac{k}{M}}.$$

If only this value were of interest, and accurate values of the corresponding mode components were not required, the use of either (256a), in the form

$$\lambda(x_1^2 + x_2^2 + 2x_3^2) \approx (x_1y_1 + x_2y_2 + 2x_3y_3),$$

or (256b), in the form

$$\lambda(x_1y_1 + x_2y_2 + 2x_3y_3) \approx (y_1^2 + y_2^2 + 2y_3^2),$$

would yield the above result for the dominant value of λ after only the second cycle.

If the remaining modes are required, the orthogonality relation (257) becomes

$$0.399x_1 + 0.747x_2 + 2.000x_3 = 0,$$

and permits the reduction of the order of the system to two.

1.28. Function space. In this section, we develop certain analogies between vector space and the so-called "function space" and point out certain essential difficulties involved in the treatment of the latter.

If, in ordinary three-dimensional space, we consider any *two* vectors $\mathbf{u}$ and $\mathbf{v}$ which are not scalar multiples of each other, we see that the totality of all vectors of the form $c_1\mathbf{u} + c_2\mathbf{v}$ comprises a double infinity of vectors, namely, all vectors in that space which are parallel to the plane of $\mathbf{u}$ and $\mathbf{v}$. If $\mathbf{w}$ is *any* third vector which is not parallel to the plane of $\mathbf{u}$ and $\mathbf{v}$, then *all* vectors in that space are comprised in the representation $c_1\mathbf{u} + c_2\mathbf{v} + c_3\mathbf{w}$. In the language of linear vector spaces, we say that "any three linearly independent vectors form a basis in three-dimensional space."

Similarly, if we consider two functions $f(x)$ and $g(x)$, defined over an interval (a, b) and not multiples of each other over that interval, those functions which are of the form $c_1f(x) + c_2g(x)$ comprise a doubly infinite set of functions. However, this set very obviously falls far short of comprising *all* functions which are defined over (a, b). The question next arises as to the possibility of choosing a set of functions such that *any* function, satisfying appropriate regularity conditions, can be expressed as a linear combination of these functions over a given interval, that is, as to

the possibility of choosing a "basis" in "function space" associated with that interval. Certainly any such set of functions must have infinitely many members; that is, function space comprises infinitely many dimensions. Also, as in vector space, we would expect the choice to be by no means a *unique* one.

In *vector* space of n dimensions, great advantage is attained by choosing as a basis a set of n mutually *orthogonal* vectors, that is, a set of vectors such that the *scalar product* of any two distinct vectors in the set is zero. This fact suggests that we introduce an analogous definition of the scalar product of two *functions*, relative to the interval under consideration. It is found that a particularly useful definition is of the form

$$(f, g) = \int_a^b f\, g\, dx. \tag{267}$$

This definition is a natural generalization of the vector definition

$$(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{n} u_k v_k$$

as the dimension of the space and the number of components involved become infinitely large.

Thus (assuming here and henceforth that the functions involved are such that the integrals involved *exist*) we are led to say that *two functions $f(x)$ and $g(x)$ are orthogonal over an interval (a, b) if the integral $\int_a^b f\, g\, dx$ vanishes.*

In particular, when $f = g$ we may think of the number $\int_a^b f^2\, dx$ as the "square of the length" of $f(x)$ in the function space associated with the interval (a, b). It is more conventional to speak of this quantity as the *norm* of f, and to write

$$\text{norm } f \equiv \| f \| \equiv (f, f) = \int_a^b f^2\, dx. \tag{268}$$

A function whose norm is *unity* is said to be *normalized*, and is seen to be analogous to a *unit vector* in vector space.*

We notice that if the norm of f is *zero*, then the integral of the nonnegative function f^2 over the interval (a, b) must vanish. This means that $f(x)$ cannot differ from zero over any range of positive

* In some references, the norm of f is defined as the positive *square root* of (f, f).

length in (a, b). In particular, if f is *continuous* everywhere in (a, b), and has a zero norm over that interval, then f must *vanish everywhere* in (a, b). However, it is clear that if $f(x)$ were zero everywhere except at a finite number of points in (a, b), the integral of f^2 over that interval would still vanish. It is convenient to speak of a function $f(x)$ for which $\int_a^b f^2\,dx = 0$ as a *trivial function*, and to say that such a function vanishes "almost everywhere" in (a, b).*

A set of n functions is said to be *linearly independent* in (a, b) if *no* linear combination of those functions (with at least one non-vanishing coefficient) is identically zero over that interval. Given any such set of functions $f_k(x)$, we can then determine a set of n new functions $\phi_k(x)$, each of which is a linear combination of certain of the f's, such that the ϕ's are mutually orthogonal and normalized in (a, b). The procedure is completely analogous to the Schmidt procedure of Section 1.12. We call such a set of functions an *orthonormal* set. Any two functions of the set then have the property that

$$(\phi_i, \phi_j) \equiv \int_a^b \phi_i \phi_j\,dx = \delta_{ij}, \tag{269}$$

where δ_{ij} is the Kronecker delta of equation (39).

Now for any (sufficiently regular) function $f(x)$ defined in (a, b) we may calculate the scalar product of that function with each function ϕ_k:

$$c_k = (f, \phi_k) = \int_a^b f\,\phi_k\,dx. \tag{270}$$

The functions ϕ_k are analogous to a set of n mutually orthogonal unit vectors in space, and we may think of the numbers $c_1, c_2, \ldots, c_n$ as the scalar components of $f(x)$ relative to those functions. We refer to these numbers as the *Fourier constants* of $f(x)$ relative to the functions $\phi_k(x)$ in (a, b).

There then exists an n-fold infinity of functions which can be generated as a linear combination of the n ϕ's. If for any such function $F(x)$ we write

$$F(x) = \sum_{k=1}^{n} a_k \phi_k(x) \qquad (a < x < b), \tag{271}$$

* For a more precise definition of this term it is desirable to consider an extension of the usual concepts of integration. The points at which a trivial function differs from zero may be *infinite* in number, so long as they are not "densely" distributed.

each coefficient a_r can be determined by forming the scalar product of ϕ_r with both members of (271), and using (269) to obtain the result

$$a_r = (F, \phi_r). \tag{272}$$

Thus the coefficient a_k in (271) is the scalar component of $F(x)$ relative to $\phi_k(x)$.

For a more general function $f(x)$, we may assume an *approximation* of the form

$$f(x) \approx \sum_{k=1}^{n} a_k \phi_k(x) \qquad (a < x < b), \tag{273}$$

and determine the coefficients a_k in such a way that the norm of the difference between the two members of (273) over (a, b) is as small as possible:

$$\Delta \equiv \left| f(x) - \sum_{k=1}^{n} a_k \phi_k(x) \right| = \int_a^b \left[f(x) - \sum_{k=1}^{n} a_k \phi_k(x) \right]^2 dx = \text{min.} \tag{274}$$

The approximation to be obtained, over the interval (a, b), is thus the best possible in the "least squares" sense.

If we think of a function $f(x)$ as a "vector" in function space, extending from an origin to a "point" P in that space (see Problems 86–91), we can interpret (274) as choosing, from all points which can be attained by vectors of the form $\sum_{k=1}^{n} a_k \phi_k(x)$, that point whose distance from P is as small as possible.

Equation (274) is equivalent to the requirement that the expression

$$\Delta \equiv \int_a^b f^2\, dx - 2 \sum_{k=1}^{n} a_k \int_a^b f\, \phi_k\, dx + \int_a^b \left[\sum_{k=1}^{n} a_k \phi_k(x) \right]^2 dx$$

take on a minimum value. But since the functions ϕ_k are orthonormal it follows that only squared terms in the last integrand have nonzero integrals. Hence, with the notation of (270), we obtain the result

$$\Delta = \int_a^b f^2\,dx - 2\sum_{k=1}^{n} a_k c_k + \sum_{k=1}^{n} a_k^2,$$

which can be put in the more convenient form

$$\Delta = \int_a^b f^2\,dx - \sum_{k=1}^{n} c_k^2 + \sum_{k=1}^{n} (c_k - a_k)^2. \tag{275}$$

From this result it is clear that, since f and the c's are fixed, Δ takes a minimum value when the coefficients a_k are chosen such that

$$a_k = c_k. \tag{276}$$

Thus it follows that *the best approximation* (273) *in the least-squares sense is obtained when* a_k *is taken as the Fourier constant of* $f(x)$ *relative to* $\phi_k(x)$ *over* (a, b).

The norm of the deviation between $f(x)$ and its *best* n-term approximation of the form (273) is then obtained, by introducing (276) into (275), in the form

$$\left\| f(x) - \sum_{k=1}^{n} c_k \phi_k(x) \right\|_{\text{min}} = \int_a^b f^2\,dx - \sum_{k=1}^{n} c_k^2. \tag{277}$$

From the definition (274) it is clear that (277) cannot be *negative;* that is, we must have

$$\int_a^b f^2\,dx - \sum_{k=1}^{n} c_k^2 \geqq 0. \tag{278}$$

This relation is known as *Bessel's inequality.*

Suppose now that the dimension n of the orthonormal set ϕ_1, ϕ_2, . . . , ϕ_n is increased without limit. The positive series in (278) must increase with n (unless the corresponding c's vanish) so that the error involved *decreases,* but since the series cannot become greater than the fixed number $\int_a^b f^2\,dx$, we conclude that the series $\sum_1^\infty c_k^2$ always converges to some positive number not greater than $\int_a^b f^2\,dx$. However, there is no assurance that the limit to which this series converges will actually coincide with this integral, so that the right-hand member of (277) then tends to zero as n increases. That is, it is *not* sufficient merely to have a set of *infinitely many* mutually orthogonal functions.

In illustration, we may recall that the functions $\cos(k\pi x/a)$ $(k = 1, 2, 3, \ldots)$ constitute an orthogonal set of functions over the interval $(0, a)$; that is, we have the relation

$$\int_0^a \cos\frac{r\pi x}{a}\cos\frac{s\pi x}{a}\,dx = 0 \qquad (r \neq s).$$

The norm of each function over $(0, a)$ is $a/2$, so that the functions

$$\phi_k(x) = \sqrt{\frac{2}{a}}\cos\frac{k\pi x}{a} \qquad (k = 1, 2, \cdots) \tag{279}$$

form an infinite orthonormal set over $(0, a)$. However, for the simple function $f(x) = 1$, the relevant Fourier constants are all *zeros*, since here

$$c_k = \sqrt{\frac{2}{a}}\int_0^a 1 \cdot \cos\frac{k\pi x}{a}\,dx = 0 \qquad (k = 1, 2, \cdots).$$

Hence, in this case the right-hand member of (277) is constantly equal to a, regardless of the value of n.

In space of n dimensions, if we construct a set of n mutually orthogonal vectors, then the possibility of expressing any other vector as a linear combination of these vectors is a consequence of the fact that no other vector can be linearly independent of them; that is, there exists no vector in that space, other than the zero vector, which is simultaneously orthogonal to these n vectors. However, in function space (of infinitely many dimensions) the difficulty consists in the fact that a function may simultaneously be orthogonal to an *infinite number* of mutually orthogonal functions. Thus, in the above case, the function $f(x) = 1$ is orthogonal to *all* the functions in the set (279) over the interval $(0, a)$. However, it can be shown that this function is the *only* nontrivial function which has this property, so that for the extended set

$$1, \sqrt{\frac{2}{a}}\cos\frac{\pi x}{a}, \sqrt{\frac{2}{a}}\cos\frac{2\pi x}{a}, \cdots, \sqrt{\frac{2}{a}}\cos\frac{n\pi x}{a}, \cdots \tag{280}$$

there is no nontrivial function whose Fourier constants *all* vanish. Such a set of orthogonal functions is said to be *complete*.

It is easily verified that the set (280) is also orthogonal (but not normalized) over the larger interval $(-a, a)$. However, it is obvious that *any odd function* of x [for which $f(-x) = -f(x)$] will

possess zero Fourier constants relative to this set, over that interval. To complete the set, it is found to be sufficient to add the functions

$$\sqrt{\frac{2}{a}} \sin \frac{\pi x}{a}, \sqrt{\frac{2}{a}} \sin \frac{2\pi x}{a}, \cdots, \sqrt{\frac{2}{a}} \sin \frac{n\pi x}{a}, \cdots. \tag{281}$$

Either of the sets (280) and (281) is complete over $(0, a)$, while the combination of the two sets is complete over $(-a, a)$ or, as a matter of fact, over *any* interval of length $2a$. These results are consequences of the known theory of *Fourier series*.

It can be shown that if the set of functions $\phi_1, \phi_2, \cdots, \phi_n, \cdots$, is *complete* in (a, b), then the right-hand member of (277) *does* indeed tend to zero in (a, b) for any function $f(x)$ which is of integrable square over that interval. The proof of this theorem is involved. Furthermore, it is difficult in practice actually to *establish* the completeness of a given infinite orthogonal set of functions. For this reason, no attempt is made here to pursue the general theory.

However, it is important to realize that one further difficulty exists. Even though we prove that the right-hand member of (277) tends to zero as n increases, so that

$$\lim_{n\to\infty} \int_a^b \Big[f(x) - \sum_{k=1}^{n} c_k \phi_k(x) \Big]^2 dx = 0, \tag{282}$$

we cannot then conclude that the integrand tends to zero everywhere in (a, b), but only that it tends to a trivial function over that interval. That is, there may be no specific value of x in (a, b) for which we are then certain that the statement

$$f(x) = \lim_{n\to\infty} \sum_{k=1}^{n} c_k \phi_k(x)$$

is true. We know only that the mean square error in (a, b) tends to zero, and we say accordingly that if (282) is true then the series *converges in the mean* to $f(x)$. However, if $f(x)$ is *continuous* throughout the interval (a, b), and if we can prove that the series $\sum_1^\infty c_k \phi_k(x)$ also represents a continuous function over that interval,* then the

* This will be the case, in particular, if the functions ϕ_k are continuous and if the series converges *uniformly* in the interval (a, b).

difference between these two functions is a continuous function with zero norm, and hence is indeed zero *everywhere* in (a, b), so that the series then converges to $f(x)$ in the true sense at each point of (a, b). Unfortunately, the conditions stated are not always fulfilled in practice.

While the knowledge that a series represents a function which differs from $f(x)$ in (a, b) by a trivial function is often all that is required (for such purposes as *integration*), it is nevertheless frequently desirable to determine whether or not the series actually represents $f(x)$ *at a given point.* The treatment of problems of this type is again beyond the scope of the present work.

The problems just discussed have been satisfactorily solved, in the mathematical literature, for a very large class of sets of orthogonal functions which frequently arise in practice. Certain known results are summarized, for convenient reference, in the following section.

It should first be pointed out that, in analogy with the corresponding situation in vector space, it is often desirable to modify somewhat the definition of the *norm* of a function. In particular, if $f(x)$ is a *complex* function of a real variable x, of the form $u(x) + i\,v(x)$, the norm of f is usually defined to be the *real* quantity

$$\| f \| = (\bar{f}, f) = \int_a^b \bar{f} f \, dx, \tag{283}$$

where a bar indicates that the complex conjugate is to be taken. We speak of (283) as the *Hermitian norm* of f. The Hermitian scalar product of two complex functions f and g is then defined to be one of the two different quantities $(\bar{f}, g)$ and $(f, \bar{g})$, these two quantities being complex conjugates. In particular, f and g are said to be *orthogonal in the Hermitian sense* if

$$(\bar{f}, g) = (f, \bar{g}) = 0. \tag{284}$$

In problems analogous to those discussed in Section 1.25, but involving *differential* equations, sets of functions are often generated for which the members $\phi_1, \phi_2, \ldots, \phi_n, \ldots$ are not orthogonal in the sense of (269), but for which a relation holds of the form

$$\int_a^b r(x)\phi_i(x)\phi_j(x)\,dx = 0 \qquad (i \neq j). \tag{285}$$

We may define the left-hand member of (285) to be the generalized or *weighted* scalar product of ϕ_i and ϕ_j. The function $r(x)$ is called the *weighting function*, and in practical applications is *nonnegative* in the relevant interval (a, b). The functions $\phi_n(x)$, for which (285) holds, are said to be *orthogonal in* (a, b) *with respect to the weighting function* $r(x)$. Finally, the norm of any function f relative to the function r is defined to be

$$\| f \|_r \equiv \int_a^b r f^2 \, dx, \tag{286}$$

and a function with *unit* norm, so defined, is said to be *normalized* relative to $r(x)$. The weighted scalar product of f and g is conveniently indicated by the notation

$$(f, g)_r \equiv \int_a^b r f g \, dx. \tag{287}$$

1.29. Sturm-Liouville problems. In this section we summarize briefly certain known results concerning sets of orthogonal functions generated by certain types of boundary-value problems involving *linear differential equations.*

A problem which consists of a homogeneous linear differential equation of the form

$$\frac{d}{dx}\left(p \frac{dy}{dx}\right) + q y + \lambda r y = 0, \tag{288}$$

together with *homogeneous boundary conditions* of a rather general type, prescribed at the end points of an interval (a, b), generally possesses a nontrivial solution only if the parameter λ is assigned one of a certain set of permissible values. For such a value of λ, say $\lambda = \lambda_k$, the conditions of the problem are satisfied by an expression of the form $y = C \phi_k(x)$ where C is an arbitrary constant. The permissible values of λ are known as its *characteristic values* (or "eigenvalues") and the corresponding functions $\phi_k(x)$, which then satisfy the conditions of the problem when $\lambda = \lambda_k$, are known as the *characteristic functions* (or "eigenfunctions").

In most cases occurring in practice, the functions $p(x)$ and $r(x)$ are positive in the interval (a, b), except possibly at one or both of the end points.

If we define a *linear differential operator* of second order by the equation

$$L = \frac{d}{dx}\left(p\,\frac{d}{dx}\right) + q, \tag{289}$$

equation (288) takes the operational form

$$L\,y + \lambda\, r\, y = 0, \tag{290}$$

and is seen to be analogous to equation (237) of Section 1.25.

We show next that, when suitable boundary conditions are prescribed at the ends of an interval (a, b), the characteristic functions of the resulting problem have properties analogous to those discussed in Section 1.25. For this purpose, let $\phi_i(x)$ and $\phi_j(x)$ be two characteristic functions, satisfying the conditions of the problem in correspondence with *distinct* characteristic numbers λ_i and λ_j. We then have the relations

$$\frac{d}{dx}\left(p\,\frac{d\phi_i}{dx}\right) + q\,\phi_i + \lambda_i r\,\phi_i = 0 \tag{291a}$$

and

$$\frac{d}{dx}\left(p\,\frac{d\phi_j}{dx}\right) + q\,\phi_j + \lambda_j r\,\phi_j = 0. \tag{291b}$$

If we multiply (291a) by ϕ_j and (291b) by ϕ_i, and subtract the resultant equations from each other, there follows

$$\begin{aligned}(\lambda_j - \lambda_i) r\,\phi_i\phi_j &= \phi_j \frac{d}{dx}\left(p\,\frac{d\phi_i}{dx}\right) - \phi_i \frac{d}{dx}\left(p\,\frac{d\phi_j}{dx}\right)\\ &= \frac{d}{dx}\left[p\left(\phi_j\,\frac{d\phi_i}{dx} - \phi_i\,\frac{d\phi_j}{dx}\right)\right],\end{aligned} \tag{292}$$

and the result of integrating both members of (292) over the interval (a, b) takes the form

$$(\lambda_j - \lambda_i)\int_a^b r\,\phi_i\phi_j\,dx = \left[p\left(\phi_j\,\frac{d\phi_i}{dx} - \phi_i\,\frac{d\phi_j}{dx}\right)\right]_a^b. \tag{293}$$

Thus, since we have assumed that $\lambda_j \neq \lambda_i$, we conclude that if the specified boundary conditions require the right-hand member of (293) to vanish, then *the characteristic functions ϕ_i and ϕ_j are orthogonal relative to the weighting function $r(x)$:*

$$(\phi_i,\ \phi_j)_r \equiv \int_a^b r\,\phi_i\phi_j\,dx = 0 \qquad (\lambda_i \neq \lambda_j). \tag{294}$$

Appropriate boundary conditions which may be seen to give rise to this situation include the following:

1. At each end of the interval we may require that either y or dy/dx or a linear combination $\alpha y + \beta\, dy/dx$ *vanish*.

2. If it happens that $p(x)$ vanishes at $x = a$ or at $x = b$, we may require instead merely that y and dy/dx remain finite at that point, and impose one of the conditions 1 at the other point.

3. If it happens that $p(b) = p(a)$, we may require merely that $y(b) = y(a)$ and $y'(b) = y'(a)$.

In most practical cases [in particular, if p, q, and r are *regular** and both p and r *positive* throughout (a, b)], when the interval (a, b) is of *finite length* it is found that in each of the listed cases there exists an *infinite* set of distinct characteristic numbers $\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots$. If also the function $q(x)$ is *nonpositive* in (a, b), and if

$$[p\ \phi_i\ \phi_i']_a^b \leqq 0,$$

the λ's are all *nonnegative*. Furthermore, except in the case of the periodicity condition 3, to each characteristic number there corresponds *one and only one* characteristic function, an arbitrary multiple of which satisfies all the specified conditions when λ is assigned the appropriate characteristic value. In case 3, *two* linearly independent characteristic functions generally correspond to each characteristic number. Such pairs of functions can then always be orthogonalized, if this is desirable, by the Schmidt procedure.

A problem of the general type just considered is known as a *Sturm-Liouville problem*.

The importance of such problems stems from the known fact that the sets of orthogonal functions generated by these problems are *complete*, in the sense of the preceding section, and further, that a positive statement can be made in such cases concerning actual *convergence* of the series representation of a sufficiently well-behaved function $f(x)$ to the value of the function *at all points where* $f(x)$ *is continuous*.

In actual practice, it is often inconvenient to *normalize* the characteristic functions (so that their norm relative to r is unity). In such cases, the coefficients in a series representation

* A function $f(x)$ is said to be *regular* at $x = x_0$ if it can be represented by a power series over an interval including x_0.

$$f(x) = \sum_{k=1}^{\infty} C_k \phi_k(x) \qquad (a < x < b) \tag{295}$$

are given by the formula

$$C_i \int_a^b r\, \phi_i^2\, dx = \int_a^b r f\, \phi_i\, dx \tag{296a}$$

or, symbolically,

$$C_i \parallel \phi_i \parallel_r = (f, \phi_i)_r. \tag{296b}$$

This result is obtained by multiplying both sides of (295) by the product $r\,\phi_i$, integrating the results formally term-by-term over (a, b), and taking into account the orthogonality of the characteristic functions relative to the weighting function $r(x)$. We notice that (296b) reduces to the obvious generalization of (270) when $\parallel \phi_i \parallel_r = 1$. The theorem to which reference was made above can then be stated as follows:

Let the functions $p(x)$, $q(x)$, and $r(x)$ in (288) be regular in the finite interval (a, b), and let $p(x)$ and $r(x)$ be positive in that interval, including the end points. Then, if $f(x)$ is piecewise differentiable in (a, b), the series (295) converges to $f(x)$ at all points inside that interval where $f(x)$ is continuous, and to the mean value $\frac{1}{2}[f(x+) + f(x-)]$ at any point where a finite jump occurs.*

While the stated conclusions follow also under even milder restrictions on $f(x)$, the condition given here is satisfied by most functions arising in practice.

To illustrate this result, we may consider the differential equation

$$\frac{d^2y}{dx^2} + \lambda y = 0, \tag{297}$$

which is the special case of (288) in which $p(x) = r(x) = 1$ and $q(x) = 0$. If we consider the interval $(0, a)$, and impose the boundary conditions

$$y(0) = 0, \qquad y(a) = 0, \tag{298}$$

it is easily verified that the characteristic values of λ, for which this problem possesses a solution other than the trivial solution $y(x) \equiv 0$, are of the form $\lambda_k = k^2\pi^2/a^2$, where k is any positive integer, and that the corresponding characteristic functions are given by

* The convergence is absolute and uniform in any *interior* subinterval which does not include a point of discontinuity as an interior or end point.

$$\phi_k(x) = \sin \frac{k\pi x}{a}.$$

Thus we obtain in this way a derivation of the *Fourier sine-series* representation

$$f(x) = \sum_{k=1}^{\infty} C_k \sin \frac{k\pi x}{a} \qquad (0 < x < a), \tag{299}$$

where, with $r(x) = 1$, equation (296) determines the coefficients in the form

$$C_k = \frac{2}{a} \int_0^a f(x) \sin \frac{k\pi x}{a}\, dx. \tag{300}$$

In a similar way, the conditions $y'(0) = y'(a) = 0$ associated with (297) give rise to the Fourier *cosine-series* representation, while the periodicity conditions $y(-a) = y(a)$ and $y'(-a) = y'(a)$, relevant to the interval $(-a, a)$, lead to the *general* Fourier series representation over that interval, involving both sines and cosines of period $2a$.

By considering other appropriate special forms of (288), expansions in terms of *Bessel functions*, *Legendre polynomials*, and so forth, may be established. The latter two cases are exceptional in that the coefficient functions p, q, and r do not satisfy the requirements specified in the preceding theorem. However, it has been found that the conclusions of the theorem are still valid in these and certain other exceptional cases.

Elementary discussions of such developments may be found in Reference 8 (Chapter 5). For more detailed treatments of these topics, References 4 and 5 are suggested.

In those cases when the interval (a, b) is of *infinite* length, or when other conditions of the stated theorem are violated, it frequently happens that the characteristic values of λ are no longer *discretely* distributed, but that all values of λ in some *continuous* range are characteristic values. In such cases, the superposition of characteristic functions is accomplished by *integration*, rather than summation. In particular, for the problem discussed relative to equation (297), it is found that *all positive values of* λ are characteristic values when the fundamental interval is of infinite length, and one is led to the *Fourier integral* representation. In certain

other exceptional cases the characteristic values may again be discretely distributed, or there may be both *continuously* distributed and *discretely* distributed characteristic values of λ.

Finally, we remark that the preceding discussion can be generalized to apply to characteristic functions of boundary-value problems governed by certain linear ordinary differential equations of higher order, as well as to characteristic functions of two or more variables associated with certain *partial* differential equations.

Analogous characteristic-value problems governed by linear *difference equations*, and by linear *integral equations*, are to be treated in Chapters 3 and 4.

REFERENCES

1. Frazer, R. A., W. J. Duncan, and A. R. Collar: *Elementary Matrices*, Cambridge University Press, London, 1946.
2. Bocher, M.: *Introduction to Higher Algebra*, The Macmillan Company, New York, 1936.
3. Birkhoff, G., and S. Maclane: *A Survey of Modern Algebra*, The Macmillan Company, New York, 1941.
4. Courant, R., and D. Hilbert: *Methoden der mathematischen Physik*, Interscience Publishers, Inc., New York, 1943.
5. Frank, Ph., and R. von Mises: *Die Differential- und Integralgleichungen der Mechanik und Physik*, Rosenberg, New York, 1943.
6. Turnbull, H. W., and A. C. Aitken: *An Introduction to the Theory of Canonical Matrices*, Blackie and Son, Ltd., London, 1932.
7. Crout, P. D.: "A Short Method for Evaluating Determinants and Solving Systems of Linear Equations with Real or Complex Coefficients," *Trans. AIEE*, Vol. 60, pp. 1235–1241 (1941).
8. Hildebrand, F. B.: *Advanced Calculus for Engineers*, Prentice-Hall, Inc., New York, 1949.

PROBLEMS

Sections 1.1, 1.2.

1. Illustrate the use of the Gauss-Jordan reduction in obtaining the general solution of each of the following sets of equations:

(a) $x_1 + 2x_2 + 2x_3 = 1,$
$2x_1 + 2x_2 + 3x_3 = 3,$
$x_1 - x_2 + 3x_3 = 5.$

(b) $2x_1 + x_3 = 4,$
$x_1 - 2x_2 + 2x_3 = 7,$
$3x_1 + 2x_2 = 1.$

Section 1.3.

2. Evaluate the following matrix products:

(a) $\begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}$. (b) $\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}\begin{bmatrix} 6 & -2 \\ -3 & 1 \end{bmatrix}$.

(c) $[a_1 \quad a_2 \quad \cdots \quad a_n]\begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$. (d) $\begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}[a_1 \quad a_2 \quad \cdots \quad a_n]$.

(e) $\begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. (f) $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}$.

3. If the product **a b c** is defined, show that it is of the form

$$[a_{ir}][b_{rs}][c_{sj}] = \Big[\sum_r \sum_s a_{ir}b_{rs}c_{sj}\Big].$$

4. It is required to determine values of the function

$$\Phi(x) = \int_a^b K(x, \xi)f(\xi)\,d\xi,$$

at the n points $x_1, x_2 \ldots, x_n$. Show that, if in each case the integral is approximated by the use of Simpson's rule, as a linear combination of the ordinates of N equally spaced points $\xi_1 = a, \xi_2, \ldots, \xi_{N-1}, \xi_N = b$, where N is odd, the calculations can be arranged in the matrix form

$$\begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \cdot \\ \cdot \\ \cdot \\ \Phi_n \end{pmatrix} \approx \frac{b-a}{3N-3}\begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & K_{22} & \cdots & K_{2N} \\ K_{31} & K_{32} & \cdots & K_{3N} \\ \cdots & \cdots & \cdots & \cdots \\ K_{n1} & K_{n2} & \cdots & K_{nN} \end{bmatrix}\begin{pmatrix} f_1 \\ 4f_2 \\ 2f_3 \\ \cdot \\ \cdot \\ f_N \end{pmatrix},$$

where $\Phi_i \equiv \Phi(x_i)$, $K_{ij} \equiv K(x_i, \xi_j)$, and $f_j \equiv f(\xi_j)$.

5. Apply the procedure of Problem 4 to the approximate evaluation of the integral

$$\Phi(x) = \int_0^1 \sqrt{x^2 + \xi^2}\,\sin \pi\xi\,d\xi,$$

for $x = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and 1, with $N = 5$. Retain three significant figures in the calculation.

6. Prove that, if two square matrices of order three are both symmetrically partitioned (as in the text on page 9), then these matrices may be correctly multiplied by treating the submatrices as single elements.

Section 1.4.

7. Prove, by direct expansion or otherwise, that $|\mathbf{a}| \cdot |\mathbf{b}| = |\mathbf{a}\,\mathbf{b}|$ when $\mathbf{a}$ and $\mathbf{b}$ are square matrices of order two.

8. Determine those values of λ for which the following set of equations may possess a nontrivial solution:

$$3x_1 + x_2 - \lambda x_3 = 0,$$
$$4x_1 - 2x_2 - 3x_3 = 0,$$
$$2\lambda x_1 + 4x_2 + \lambda x_3 = 0.$$

For each permissible value of λ, determine the most general solution.

9. Show that the equation of the straight line $ax + by + c = 0$ which passes through the points (x_1, y_1) and (x_2, y_2) can be written in the form

$$\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0.$$

10. Express the requirement, that four points (x_i, y_i) $(i = 1, 2, 3, 4)$ lie simultaneously on a conic of the form $ax^2 + bxy + cy^2 + d = 0$, in terms of the vanishing of a determinant.

Section 1.5.

11. A *symmetric* matrix $\mathbf{a} = [a_{ij}]$ is a square matrix for which $a_{ji} = a_{ij}$.
(a) Show that $\mathbf{a}^T = \mathbf{a}$ if and only if $\mathbf{a}$ is symmetric.
(b) Let $\mathbf{a}$ and $\mathbf{b}$ represent symmetric matrices of order n. Prove that $\mathbf{a}\,\mathbf{b}$ is also symmetric if and only if $\mathbf{a}$ and $\mathbf{b}$ are commutative.

12. Prove that, if $\mathbf{a}$ and $\mathbf{b}$ are square matrices of order n, there follows $\text{Adj}\,(\mathbf{a}\,\mathbf{b}) = (\text{Adj}\,\mathbf{b})(\text{Adj}\,\mathbf{a})$.

13. Let $\mathbf{a}$ and $\mathbf{b}$ represent diagonal matrices of order n.
(a) Prove that $\mathbf{a}\,\mathbf{b}$ is also a diagonal matrix.
(b) Prove that $\mathbf{b}\,\mathbf{a} = \mathbf{a}\,\mathbf{b}$.

Section 1.6.

14. If $\mathbf{d} = [d_i\,\delta_{ij}]$ is a diagonal matrix, prove that its inverse is given by

$$\mathbf{d}^{-1} = \left[\frac{1}{d_i}\,\delta_{ij}\right].$$

15. (a) Prove that $\mathbf{a}\ \mathrm{Adj}\ \mathbf{a} = \mathbf{0}$ if $\mathbf{a}$ is singular, and illustrate by an example.

(b) Prove that $|\mathrm{Adj}\ \mathbf{a}| = |\mathbf{a}|^{n-1}$ ($\mathbf{a}$ of order n) and illustrate by an example.

16. Determine the elements of $\mathbf{a}^T$, Adj $\mathbf{a}$, and $\mathbf{a}^{-1}$ when

$$\mathbf{a} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

17. If $\mathbf{a}\,\mathbf{b} = \mathbf{a}\,\mathbf{c}$, where $\mathbf{a}$ is a square matrix, when does it necessarily follow that $\mathbf{b} = \mathbf{c}$? Give an example in which this conclusion does *not* follow.

Section 1.7.

18. If $\mathbf{a}$ is a square matrix of order n, show that each of the three elementary operations on *rows* of $\mathbf{a}$ can be accomplished by *pre*multiplying $\mathbf{a}$ by a matrix $\mathbf{P}$, where $\mathbf{P}$ is formed by performing that operation on corresponding rows of the *unit* matrix $\mathbf{I}$ of order n. In each case, show also that $\mathbf{P}$ is nonsingular.

19. If $\mathbf{a}$ is a square matrix of order n, show that each of the elementary operations on *columns* of $\mathbf{a}$ can be accomplished by *post*multiplying $\mathbf{a}$ by a matrix $\mathbf{Q}$, where $\mathbf{Q}$ is formed by performing that operation on corresponding columns of the unit matrix $\mathbf{I}$ of order n. In each case, show also that $\mathbf{Q}$ is nonsingular.

20. (a) If $a_{ij} = r_i s_j$, prove that $\mathbf{a}$ is of rank one or zero.

(b) If $\mathbf{a} = [a_{ij}]$ is of rank one, prove that a_{ij} can be written as $r_i s_j$.

[Such a matrix is called a *dyad*.]

Section 1.8.

21. (a) By investigating ranks of relevant matrices, show that the following set of equations possesses a one-parameter family of solutions:

$$\begin{aligned} 2x_1 - x_2 - x_3 &= 2, \\ x_1 + 2x_2 + x_3 &= 2, \\ 4x_1 - 7x_2 - 5x_3 &= 2. \end{aligned}$$

(b) Determine the general solution.

22. (a) Show that the set

$$\begin{aligned} 2x_1 - 2x_2 + x_3 &= \lambda x_1, \\ 2x_1 - 3x_2 + 2x_3 &= \lambda x_2, \\ -x_1 + 2x_2 &= \lambda x_3 \end{aligned}$$

can possess a nontrivial solution only if $\lambda = 1$ or $\lambda = -3$.

(b) Obtain the general solution in each case.

Section 1.9.

23. (a) Prove that if the Gramian of two real vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ vanishes, then $\mathbf{v}_1$ and $\mathbf{v}_2$ are linearly dependent. [Notice that, if $G = 0$, the equations $c_1\mathbf{v}_1^T\mathbf{v}_1 + c_2\mathbf{v}_1^T\mathbf{v}_2 = 0$ and $c_1\mathbf{v}_2^T\mathbf{v}_1 + c_2\mathbf{v}_2^T\mathbf{v}_2 = 0$ possess a nontrivial solution. Multiply the first equation by c_1, the second by c_2, add, and interpret the result.]

(b) Generalize the result of part (a) to the case of n vectors.

24. (a) If $\mathbf{a}$ is an $m \times s$-matrix and $\mathbf{b}$ is an $s \times n$-matrix, and if the elements of the rth column of $\mathbf{b}$ are considered to comprise the elements of a vector $\mathbf{v}_r$, show that the rth column of the product $\mathbf{a}\,\mathbf{b}$ is the vector $\mathbf{a}\,\mathbf{v}_r$:

$$\begin{bmatrix}\mathbf{a}\end{bmatrix}\begin{bmatrix}\vdots & & \vdots\\ \mathbf{v}_1 & \cdots & \mathbf{v}_n\\ \vdots & & \vdots\end{bmatrix} = \begin{bmatrix}\vdots & & \vdots\\ \mathbf{a}\,\mathbf{v}_1 & \cdots & \mathbf{a}\,\mathbf{v}_n\\ \vdots & & \vdots\end{bmatrix}.$$

(b) If the matrices $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are conformable in that order, and if the rth *row* of $\mathbf{a}$ comprises a vector $\mathbf{u}_r$, whereas the sth *column* of $\mathbf{c}$ comprises a vector $\mathbf{v}_s$, show that the typical element P_{rs} of the product $\mathbf{a}\,\mathbf{b}\,\mathbf{c}$ is the scalar $\mathbf{u}_r\,\mathbf{b}\,\mathbf{v}_s$:

$$\begin{bmatrix}\cdots\cdots\mathbf{u}_1\cdots\cdots\rightarrow\\ \\ \\ \cdots\cdots\mathbf{u}_m\cdots\cdots\rightarrow\end{bmatrix}\begin{bmatrix}\mathbf{b}\end{bmatrix}\begin{bmatrix}\vdots & & \vdots\\ \mathbf{v}_1 & \cdots & \mathbf{v}_n\\ \vdots & & \vdots\end{bmatrix} = \begin{bmatrix}\mathbf{u}_1\,\mathbf{b}\,\mathbf{v}_1 & \cdots & \mathbf{u}_1\,\mathbf{b}\,\mathbf{v}_n\\ \cdots & \cdots & \cdots\\ \mathbf{u}_m\,\mathbf{b}\,\mathbf{v}_1 & \cdots & \mathbf{u}_m\,\mathbf{b}\,\mathbf{v}_n\end{bmatrix}.$$

25. Determine the dimension of the vector space generated by each of the following sets of vectors:

(a) $\{1, 1, 0\}$, $\{1, 0, 1\}$, $\{0, 1, 1\}$.
(b) $\{1, 0, 0\}$, $\{0, 1, 0\}$, $\{0, 0, 1\}$, $\{1, 1, 1\}$.
(c) $\{1, 1, 1\}$, $\{1, 0, 1\}$, $\{1, 2, 1\}$.

Section 1.10.

26. Show that the set of equations

$$\begin{aligned} x_1 + x_2 + x_3 &= 3,\\ x_1 + x_2 - x_3 &= 1,\\ 3x_1 + 3x_2 - 5x_3 &= 1 \end{aligned}$$

possesses a one-parameter family of solutions, and verify directly that the vector $\mathbf{c}$ whose elements comprise the right-hand members is orthogonal to all vector solutions of the transposed homogeneous set of equations.

27. (a) Prove that if the set $\mathbf{a}\,\mathbf{x} = \mathbf{0}$ possesses an r-parameter set of nontrivial solutions, then the same is true of the transposed set $\mathbf{a}^T\mathbf{x}' = \mathbf{0}$, and conversely.

(b) Interpret the statement at the end of Section 1.10 in the case when the transposed set $\mathbf{a}^T\mathbf{x}' = \mathbf{0}$ possesses no nontrivial solution.

Section 1.11.

28. Show that the problem

$$x_1 - 2x_2 = \lambda x_1,$$
$$x_1 - x_2 = \lambda x_2$$

does not possess real nontrivial solutions for any values of λ.

29. (a) Determine the characteristic numbers (λ_1, λ_2) and corresponding unit characteristic vectors $(\mathbf{e}_1, \mathbf{e}_2)$ of the matrix

$$\mathbf{a} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$

(b) Verify that $\mathbf{e}_1$ and $\mathbf{e}_2$ are orthogonal.

(c) Use the results of part (a), together with equation (89), to obtain the solution of the following set of equations:

$$5x_1 + 2x_2 = \lambda x_1 + 2,$$
$$2x_1 + 2x_2 = \lambda x_2 + 1.$$

Consider the exceptional cases separately.

30. (a) Suppose that the n characteristic vectors of the symmetric matrix $\mathbf{a}$ are *not* normalized (reduced to unit length). If they are denoted by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, show that (89) must be replaced by the equation

$$\mathbf{x} = \sum_{k=1}^{n} \frac{(\mathbf{v}_k, \mathbf{c})}{\lambda_k - \lambda} \frac{\mathbf{v}_k}{(\mathbf{v}_k, \mathbf{v}_k)}.$$

(b) Verify this result in the case of Problem 29(c).

Section 1.12.

31. Construct a set of three mutually orthogonal unit vectors which are linear combinations of the vectors $\{1, 0, 2, 2\}$, $\{1, 1, 0, 1\}$, and $\{1, 1, 0, 0\}$.

Sections 1.13, 1.14.

32. If F is a (homogeneous) quadratic form in $x_1, x_2, \ldots, x_n$, prove that

$$F = \frac{1}{2} \sum_{k=1}^{n} x_k \frac{\partial F}{\partial x_k}.$$

33. Construct a normalized modal matrix $\mathbf{Q}$ corresponding to the matrix

$$\mathbf{a} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 3 \end{bmatrix},$$

and verify that $\mathbf{Q}^T \mathbf{a} \mathbf{Q} = [\lambda_i \delta_{ij}]$. (Notice the footnote on page 31.)

34. Reduce the quadratic form $F = x_1^2 + 3x_2^2 + 3x_3^2 - 2x_2x_3$ to a canonical form by making an appropriate change in variables, $\mathbf{x} = \mathbf{Q}\mathbf{x}'$, where $\mathbf{Q}$ is an orthogonal matrix.

35. Let $\mathbf{M}$ represent a modal matrix of a symmetric matrix $\mathbf{a}$, the modal columns of which are orthogonal, but not necessarily reduced to unit length. If the characteristic vectors whose elements comprise successive columns are denoted by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, show that

$$\mathbf{M}^T \mathbf{M} = \begin{bmatrix} \mathbf{v}_1^2 & 0 & \cdots & 0 \\ 0 & \mathbf{v}_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{v}_n^2 \end{bmatrix}$$

and

$$\mathbf{M}^T \mathbf{a} \mathbf{M} = \begin{bmatrix} \lambda_1 \mathbf{v}_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2 \mathbf{v}_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \mathbf{v}_n^2 \end{bmatrix}.$$

Hence deduce also that the form $F = \mathbf{x}^T \mathbf{a} \mathbf{x}$ is reduced to the form

$$F = \lambda_1 \mathbf{v}_1^2 x_1'^2 + \lambda_2 \mathbf{v}_2^2 x_2'^2 + \cdots + \lambda_n \mathbf{v}_n^2 x_n'^2$$

by the change in variables $\mathbf{x} = \mathbf{M}\mathbf{x}'$.

[Notice that this form reduces to the canonical form (109) if the vectors $\mathbf{v}_i$ are normalized, so that $\mathbf{M}$ is an orthogonal matrix.]

Section 1.15.

36. Let $\mathbf{a} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Determine nonsingular matrices $\mathbf{P}$ and $\mathbf{Q}$ such that $\mathbf{P} \mathbf{a} \mathbf{Q} = \mathbf{b}$, where $\mathbf{b}$ is obtained by interchanging the two rows of $\mathbf{a}$ and then adding twice the first column to the second column. (See also Problems 18 and 19).

Section 1.16.

37. Determine the characteristic numbers (λ_1, λ_2) and corresponding Hermitian unit characteristic vectors $(\mathbf{e}_1, \mathbf{e}_2)$ of the problem

$$9x_1 + (2 + 2i)x_2 = \lambda x_1,$$
$$(2 - 2i)x_1 + 2x_2 = \lambda x_2,$$

where $i^2 = -1$, and verify that $\mathbf{e}_1$ and $\mathbf{e}_2$ are orthogonal in the Hermitian sense.

38. Describe the modification of the Schmidt orthogonalization procedure of Section 1.12 which applies when orthogonality and unit length are defined in the Hermitian sense.

39. Prove that the normalized modal matrix $\mathbf{U}$ of a Hermitian matrix $\mathbf{h}$ is a unitary matrix [i.e., that equation (128) is satisfied].

40. (a) Show that, if a matrix is both unitary and Hermitian, it must satisfy the equation $\mathbf{U}^2 = \mathbf{I}$.

(b) Prove that any matrix of order two, of this type, is either the positive or negative unit matrix, or else is of the form

$$\mathbf{U} = \begin{bmatrix} a & r\,e^{i\alpha} \\ r\,e^{-i\alpha} & -a \end{bmatrix},$$

where a, r, and α are real and $a^2 + r^2 = 1$.

Section 1.17.

41. Determine whether the form

$$F = x_1^2 + 2x_2^2 + x_3^2 - 2x_1x_2 + 2x_2x_3$$

is positive definite, by examining the characteristic numbers of the associated matrix.

42. Determine a change in variables which reduces the forms

$$A = 3x_1^2 - 2x_1x_2 + 3x_2^2, \qquad B = 2x_1^2 + 2x_2^2$$

simultaneously to the canonical forms

$$A = \lambda_1\alpha_1^2 + \lambda_2\alpha_2^2, \qquad B = \alpha_1^2 + \alpha_2^2,$$

by using the methods of Section 1.17.

Section 1.18.

43. Find the *sum* and *product* of all characteristic numbers of the matrix

$$\mathbf{a} = \begin{bmatrix} 2 & 1 & -1 & 0 \\ 1 & 3 & 4 & 2 \\ -1 & 4 & 1 & 2 \\ 0 & 2 & 2 & 1 \end{bmatrix}.$$

44. Determine whether the matrix $\mathbf{a}$ of Problem 43 is positive definite.

45. A real symmetric matrix $\mathbf{a}$ is said to be *negative definite* if its associated quadratic form $\mathbf{x}^T\,\mathbf{a}\,\mathbf{x}$ is nonpositive for all real $\mathbf{x}$, and is zero only

when $\mathbf{x} = \mathbf{0}$. State conditions under which this situation exists, (a) in terms of the characteristic numbers of $\mathbf{a}$, and (b) in terms of the discriminants of $\mathbf{a}$. (Notice that $\mathbf{a}$ is negative definite if and only if $-\mathbf{a}$ is positive definite.)

Section 1.19.

46. A vector $\mathbf{x}$ has components $\{1, 1, 1\}$ along unit vectors $\mathbf{i}_1$, $\mathbf{i}_2$, and $\mathbf{i}_3$ coinciding with the axes of a rectangular $x_1x_2x_3$-coordinate system. If new axes are chosen in such a way that the new unit vectors are related to the original ones by the equations

$$\mathbf{i}_1' = \frac{(\mathbf{i}_1 + \mathbf{i}_2)}{\sqrt{2}}, \quad \mathbf{i}_2' = \frac{(\mathbf{i}_1 - \mathbf{i}_2)}{\sqrt{2}}, \quad \mathbf{i}_3' = \mathbf{i}_3,$$

determine the components of $\mathbf{x}$ along the new axes. Show also that the new coordinate system is also rectangular.

47. A vector $\mathbf{y}$ is related to the vector $\mathbf{x}$ of Problem 46 by the equation $\mathbf{y} = \mathbf{a}\,\mathbf{x}$, where

$$\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

when the components refer to the original axes. Assuming that $\mathbf{x}$ and $\mathbf{y}$ transform in the same way under the change of axes, determine the components of $\mathbf{y}$ in the new system, *first*, by determining the original components of $\mathbf{y}$ and transforming them directly, and *second*, by using equation (169) in connection with the result of Problem 46.

48. Prove that, if the new unit vectors of (160) are mutually orthogonal, then the matrix (166) is an orthogonal matrix.

49. (a) Show that an orthogonal matrix of order two is necessarily of one of the following two types:

$$\mathbf{Q}^{(+)} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}, \qquad \mathbf{Q}^{(-)} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{bmatrix}.$$

[Notice that $|\mathbf{Q}^{(+)}| = +1$, and $|\mathbf{Q}^{(-)}| = -1$.]

(b) If $\mathbf{x}$ and $\mathbf{x}'$ are considered as two distinct vectors referred to the same axes, and are related by the equation $\mathbf{x} = \mathbf{Q}\,\mathbf{x}'$, verify that $\mathbf{x}$ is rotated into $\mathbf{x}'$ through the angle α by a positive (counterclockwise) rotation if $\mathbf{Q} = \mathbf{Q}^{(+)}$.

(c) If $\mathbf{x}$ and $\mathbf{x}'$ are considered as comprising the components of the same vector, referred to original and rotated axes, respectively, verify that the coordinate transformation $\mathbf{x} = \mathbf{Q}^{(+)}\,\mathbf{x}'$ corresponds to a negative rotation of the original axes, through the angle α.

(d) If $\mathbf{Q} = \mathbf{Q}^{(-)}$ in parts (b) and (c), verify that the transformations then each involve a *reversed* rotation combined with a suitable *reflection*.

Section 1.20.

50. Show that if the first two columns of an orthogonal matrix $\mathbf{Q}$ comprise the elements of two unit characteristic vectors of a symmetric matrix $\mathbf{a}$, then $\mathbf{Q}^{-1}\mathbf{a}\mathbf{Q}$ is of the form

$$\begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \alpha_{33} & \cdots & \alpha_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \alpha_{3n} & \cdots & \alpha_{nn} \end{bmatrix},$$

where λ_1 and λ_2 are the characteristic numbers corresponding respectively to the two characteristic vectors.

Section 1.21.

51. Let $\mathbf{M}$ represent a modal matrix of *any* square matrix $\mathbf{a}$ of order n with n distinct characteristic numbers $\lambda_1, \ldots, \lambda_n$, the successive columns of $\mathbf{M}$ comprising the elements of successive corresponding characteristic vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ which need be neither orthogonal nor of unit length. Prove that $\mathbf{M}^{-1}\mathbf{a}\mathbf{M}$ is then of the form $[\lambda_i \delta_{ij}]$, so that $\mathbf{a}$ is thus diagonalized by a similarity transformation.

[Make appropriate modifications in the argument of equations (101) to (104) of Section 1.13. Notice that $\mathbf{a}$ need not be symmetric.]

Section 1.22.

52. (a) If $\mathbf{a} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, determine the characteristic numbers and corresponding characteristic vectors of the matrix $\mathbf{b} = \mathbf{a}^5 - 3\mathbf{a}^4 + 2\mathbf{a} - \mathbf{I}$.

(b) Determine whether the matrix $\mathbf{b}$ is positive definite.

53. Evaluate $\mathbf{a}^{100}$, where $\mathbf{a}$ is defined in Problem 52.

54. (a) Show that if $\mathbf{a}$ is a symmetric matrix of order n, with distinct characteristic numbers, then

$$\mathbf{a}^N = \sum_{k=1}^{n} \lambda_k^N Z_k(\mathbf{a}),$$

where Z_k is defined by (213) and N is a positive integer.

(b) Let λ_n be the dominant characteristic number (i.e., the characteristic number with largest absolute value). Noticing that for sufficiently large N the nth term in the preceding sum will then predominate, show that if $\mathbf{x}$ is an arbitrary vector there follows

$$\mathbf{a}^N \mathbf{x} \approx \lambda_n^N \mathbf{v}, \qquad \mathbf{a}^{N+1} \mathbf{x} \approx \lambda_n^{N+1} \mathbf{v},$$

where $\mathbf{v} = [Z_n(\mathbf{a})]\mathbf{x}$, when N is large, unless it happens that $\mathbf{v} = \mathbf{0}$.

(c) Hence deduce that, in general, if an arbitrary vector $\mathbf{x}$ is premultiplied repeatedly by a symmetric matrix $\mathbf{a}$, the vector obtained after $N+1$ such multiplications is approximately λ_n times that obtained after N multiplications, where λ_n is the dominant characteristic number of $\mathbf{a}$, and hence also that the vectors obtained after successive multiplications tend, in general, to become multiples of the characteristic vector associated with λ_n.

(d) Show that the exceptional case, in which the vector $\mathbf{x}$ is such that $[Z_n(\mathbf{a})]\mathbf{x} = \mathbf{0}$, will occur if $\mathbf{x}$ happens to be a characteristic vector of $\mathbf{a}$, corresponding to a characteristic number $\lambda_k \neq \lambda_n$, or if $\mathbf{x}$ is a linear combination of those vectors.

[A more complete treatment of this procedure, from a somewhat different point of view, is given in Section 1.23.]

55. Suppose that $\mathbf{a}$ is real and symmetric of order two, with a repeated characteristic number $\lambda_1 = \lambda_2$.

(a) Obtain from (216) the evaluation

$$e^{\mathbf{a}} = e^{\lambda_1}\mathbf{a} - (\lambda_1 - 1)e^{\lambda_1}\mathbf{I}.$$

(b) Prove that $\mathbf{a}$ must in this case be a scalar matrix, $\mathbf{a} = k\mathbf{I}$, and show that the evaluation of part (a) reduces to

$$e^{k\mathbf{I}} = e^k\mathbf{I}.$$

56. Suppose that the elements of a matrix $\mathbf{a}(t) = [a_{ij}(t)]$ are differentiable functions of a variable t.

(a) From the definition

$$\frac{d\mathbf{a}(t)}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{a}(t+\Delta t) - \mathbf{a}(t)}{\Delta t} \equiv \lim_{\Delta t \to 0} \frac{\Delta \mathbf{a}}{\Delta t},$$

prove that $d\mathbf{a}(t)/dt = [da_{ij}/dt]$.

(b) Prove that

$$\frac{d}{dt}(\mathbf{a}\,\mathbf{b}) = \frac{d\mathbf{a}}{dt}\mathbf{b} + \mathbf{a}\frac{d\mathbf{b}}{dt}.$$

(c) Specialize the result of part (b) in the case when $\mathbf{b} = \mathbf{a}$, and give an example to show that $d\mathbf{a}^2/dt \neq 2\mathbf{a}\,d\mathbf{a}/dt$ in general.

Section 1.23.

57. Determine the dominant characteristic number and the corresponding characteristic vector for the system

$$\begin{aligned} x_1 + x_2 + x_3 &= \lambda x_1, \\ x_1 + 3x_2 + 3x_3 &= \lambda x_2, \\ x_1 + 3x_2 + 6x_3 &= \lambda x_3. \end{aligned}$$

(Retain slide-rule accuracy.)

58. Show that the iterative method does not converge to a characteristic vector if $\mathbf{a} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$, regardless of the initial approximation. Explain.

59. Investigate the application of the iterative method to the matrix $\mathbf{a} = \begin{bmatrix} 1 & 1 \\ -2 & -1 \end{bmatrix}$. Explain.

Section 1.24.

60. Determine the two largest characteristic numbers, and corresponding characteristic vectors, of the system

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= \lambda x_1, \\ x_1 + 2x_2 + 2x_3 + 2x_4 &= \lambda x_2, \\ x_1 + 2x_2 + 3x_3 + 3x_4 &= \lambda x_3, \\ x_1 + 2x_2 + 3x_3 + 4x_4 &= \lambda x_4. \end{aligned}$$

(Retain slide-rule accuracy.)

61. Determine all characteristic numbers, and the corresponding characteristic vectors, of the system

$$\begin{aligned} x_1 - x_2 &= \lambda x_1, \\ -x_1 + 2x_2 - x_3 &= \lambda x_2, \\ -x_2 + 2x_3 - x_4 &= \lambda x_3, \\ -x_3 + x_4 &= \lambda x_4. \end{aligned}$$

(Retain slide-rule accuracy.)

62. Suppose that the iterative method of Section 1.27 fails to converge for a real symmetric matrix $\mathbf{a}$, so that λ_n and $-\lambda_n$ are both dominant characteristic numbers. Take $\lambda_n > 0$, and write $\lambda_{n-1} = -\lambda_n$.

Show that, if r is sufficiently large, the input in the rth cycle is given approximately by

$$\mathbf{x}^{(r)} \approx \mathbf{v}_n + \mathbf{v}_{n-1},$$

where $\mathbf{v}_n$ and $\mathbf{v}_{n-1}$ are constant multiples of the unit characteristic vectors relevant to λ_n and $\lambda_{n-1} \equiv -\lambda_n$, respectively, whereas the output is then given approximately by

$$\mathbf{y}^{(r)} \approx \lambda_n(\mathbf{v}_n - \mathbf{v}_{n-1}).$$

Show further that *if this output is taken as the input for the next cycle,* so that

$$\mathbf{x}^{(r+1)} = \mathbf{y}^{(r)},$$

there follows also

$$\mathbf{y}^{(r+1)} \approx \lambda_n^2(\mathbf{v}_n + \mathbf{v}_{n-1}),$$

so that λ_n can then be determined approximately by the relation

$$\mathbf{y}^{(r+1)} \approx \lambda_n{}^2 \mathbf{x}^{(r)},$$

after which approximations to $\mathbf{v}_n$ and $\mathbf{v}_{n-1}$ are given by

$$\mathbf{v}_n \approx \frac{1}{2}\left[\mathbf{x}^{(r)} + \frac{1}{\lambda_n}\mathbf{y}^{(r)}\right], \quad \mathbf{v}_{n-1} \approx \frac{1}{2}\left[\mathbf{x}^{(r)} - \frac{1}{\lambda_n}\mathbf{y}^{(r)}\right],$$

when r is sufficiently large.

63. Illustrate the technique developed in Problem 62 in the case of the symmetric matrix $\mathbf{a} = \begin{bmatrix} -4 & 3 \\ 3 & 4 \end{bmatrix}$.

Section 1.25.

64. Prove that the characteristic numbers of the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$ are real when $\mathbf{a}$ and $\mathbf{b}$ are symmetric, and either $\mathbf{a}$ or $\mathbf{b}$ is positive definite.

65. Determine the characteristic numbers and vectors of the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$, where

$$\mathbf{a} = \begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

and verify that the characteristic vectors are orthogonal relative to both $\mathbf{a}$ and $\mathbf{b}$.

66. Construct a normalized modal matrix associated with Problem 65, where the normalization is relative to $\mathbf{b}$.

67. Use the results of Problems 65 and 66 to determine a change in variables which reduces the quadratic forms

$$A = 5x_1{}^2 + 4x_1x_2 + 3x_2{}^2, \qquad B = x_1{}^2 + 2x_2{}^2$$

simultaneously to the canonical forms

$$A = \lambda_1\alpha_1{}^2 + \lambda_2\alpha_2{}^2, \qquad B = \alpha_1{}^2 + \alpha_2{}^2.$$

68. Show that the condition (256a) is equivalent to the condition

$$(\mathbf{x}^{(r)}, \mathbf{x}^{(r)})_{\mathbf{a}} \approx \lambda_n(\mathbf{x}^{(r)}, \mathbf{x}^{(r)})_{\mathbf{b}}$$

and that (256b) is equivalent to the condition

$$(\mathbf{y}^{(r)}, \mathbf{y}^{(r)})_{\mathbf{b}} \approx \lambda_n(\mathbf{x}^{(r)}, \mathbf{x}^{(r)})_{\mathbf{a}}.$$

Section 1.26.

69. If $\mathbf{a} = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix}$, determine the characteristic vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}$, and the characteristic vectors $\mathbf{e}_1'$ and $\mathbf{e}_2'$ of the associated problem $\mathbf{a}^T\mathbf{x}' = \lambda\,\mathbf{x}'$, and verify the validity of equation (260).

70. (a) Suppose that $\mathbf{a}$ possesses n distinct characteristic numbers $\lambda_1, \ldots, \lambda_n$, with corresponding characteristic vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, and denote corresponding characteristic vectors of $\mathbf{a}^T$ by $\mathbf{v}_1', \ldots, \mathbf{v}_n'$. Obtain the solution of the problem $\mathbf{a}\,\mathbf{x} - \lambda\,\mathbf{x} = \mathbf{c}$ in the form

$$\mathbf{x} = \sum_{k=1}^{n} \frac{(\mathbf{v}_k', \mathbf{c})}{\lambda_k - \lambda} \frac{\mathbf{v}_k}{(\mathbf{v}_k', \mathbf{v}_k)},$$

when $\lambda \neq \lambda_1, \ldots, \lambda_n$. [Compare Problem 30.]

(b) Discuss the situation when λ assumes a characteristic value λ_p. (Notice also that this case is described by the result of replacing $\mathbf{a}$ by $\mathbf{a} - \lambda_p\mathbf{I}$ in the statement at the end of Section 1.10.)

71. With the terminology of Problem 70, use the result at the end of Section 1.8 to show that the elements of $\mathbf{v}_r$ are proportional to the cofactors of respective elements in any *row* of the matrix

$$[\mathbf{a} - \lambda_r\mathbf{I}] \equiv \begin{bmatrix} (a_{11} - \lambda_r) & a_{12} & \cdots & a_{1n} \\ a_{21} & (a_{22} - \lambda_r) & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & (a_{nn} - \lambda_r) \end{bmatrix},$$

whereas the elements of $\mathbf{v}_r'$ are proportional to the cofactors of respective elements in any *column* of that matrix, if not all the relevant cofactors vanish. Verify this conclusion in the example of Problem 69.

Section 1.27.

Determine the natural frequencies and natural modes of vibration of the mechanical system of Figure 1.1 in the following cases:

72. Assume $k_1 = 2k$, $k_2 = k_3 = k$; $M_1 = M_2 = M_3 = M$.

73. Assume $k_1 = 2k$, $k_2 = k_3 = k$; $M_1 = M_2 = M$, $M_3 = 2M$.

74. Assume $k_1 = 0$, $k_2 = k_3 = k$; $M_1 = M_2 = M_3 = M$.

75. Assume $k_1 = 0$, $k_2 = k_3 = k$; $M_1 = M_2 = M$, $M_3 = 2M$.

[In most physical problems of this type, the fundamental mode (corresponding to the *smallest* natural frequency) is usually such that the initial approximation $\{1, 1, 1, \ldots, 1\}$ is a convenient one. In the highest natural mode, the successive masses generally tend to oscillate with opposite phases, so that the initial approximation $\{1, -1, 1, \ldots, \pm 1\}$ usually leads to more rapid convergence. In Problems 74 and 75, the system of masses and springs is unattached to a support, and the characteristic number associated with $\omega = 0$ corresponds to motion of the system as a rigid body.]

Minimal Properties of Characteristic Numbers.

76. Let $\mathbf{a}$ denote a real symmetric matrix of order n, with characteristic numbers $\lambda_1, \ldots, \lambda_n$, arranged in increasing *algebraic* order ($\lambda_1 \leqq \lambda_2 \leqq$

$\cdots \leqq \lambda_n$), and corresponding normalized and orthogonalized characteristic vectors $\mathbf{e}_1, \cdots, \mathbf{e}_n$.

(a) If $\mathbf{x}$ is an arbitrary real vector with n components, and hence expressible in the form

$$\mathbf{x} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n = \sum_{k=1}^{n} c_k\mathbf{e}_k,$$

establish the relations

$$\mathbf{x}^T\mathbf{x} = c_1^2 + c_2^2 + \cdots + c_n^2 = \sum_{k=1}^{n} c_k^2,$$

$$\mathbf{a}\mathbf{x} = \lambda_1 c_1\mathbf{e}_1 + \lambda_2 c_2\mathbf{e}_2 + \cdots + \lambda_n c_n\mathbf{e}_n = \sum_{k=1}^{n} \lambda_k c_k\mathbf{e}_k,$$

and $$\mathbf{x}^T\mathbf{a}\mathbf{x} = \lambda_1 c_1^2 + \lambda_2 c_2^2 + \cdots + \lambda_n c_n^2 = \sum_{k=1}^{n} \lambda_k c_k^2.$$

(b) Deduce that

$$\frac{\mathbf{x}^T\mathbf{a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\lambda_1 c_1^2 + \lambda_2 c_2^2 + \cdots + \lambda_n c_n^2}{c_1^2 + c_2^2 + \cdots + c_n^2},$$

and hence also that

$$\left|\frac{\mathbf{x}^T\mathbf{a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\right| \leqq |\lambda_i|_{\max}.$$

(c) Prove that

$$\lambda_n - \frac{\mathbf{x}^T\mathbf{a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_{k=1}^{n} (\lambda_n - \lambda_k)c_k^2}{\sum_{k=1}^{n} c_k^2} \geqq 0,$$

for *any* real vector $\mathbf{x}$,

(d) If $\mathbf{x}$ is orthogonal to the characteristic vectors $\mathbf{e}_{r+1}, \mathbf{e}_{r+2}, \cdots, \mathbf{e}_n$, show that

$$\lambda_r - \frac{\mathbf{x}^T\mathbf{a}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_{k=1}^{r} (\lambda_r - \lambda_k)c_k^2}{\sum_{k=1}^{r} c_k^2} \geqq 0.$$

(c) Show that, if $\mathbf{x} = \mathbf{e}_i$, there follows

$$\frac{\mathbf{x}^T \mathbf{a}\, \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \equiv \frac{\mathbf{e}_i{}^T \mathbf{a}\, \mathbf{e}_i}{\mathbf{e}_i{}^T \mathbf{e}_i} = \mathbf{e}_i{}^T \mathbf{a}\, \mathbf{e}_i = \lambda_i \quad (i = 1, 2, \cdots, n).$$

77. Let $\mathbf{a}$ be a real symmetric matrix, with characteristic numbers $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$ and corresponding normalized and orthogonalized characteristic vectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$. Deduce the following results from the results of Problem 76:

(a) The number λ_n is the maximum value of $(\mathbf{x}^T \mathbf{a}\, \mathbf{x})/(\mathbf{x}^T \mathbf{x})$ for all real vectors $\mathbf{x}$, and this maximum value is taken on when $\mathbf{x}$ is identified with a characteristic vector associated with λ_n.

(b) The number λ_r is the maximum value of $(\mathbf{x}^T \mathbf{a}\, \mathbf{x})/(\mathbf{x}^T \mathbf{x})$ for all real vectors $\mathbf{x}$ which are simultaneously orthogonal to the characteristic vectors associated with $\lambda_{r+1}, \lambda_{r+2}, \cdots, \lambda_n$, and this maximum value is taken on when $\mathbf{x}$ is identified with a characteristic vector associated with λ_r.

(c) The number λ_n is the maximum value of $\mathbf{e}^T \mathbf{a}\, \mathbf{e}$ for all real *unit* vectors $\mathbf{e}$, and the number λ_r is the maximum value for all unit vectors simultaneously orthogonal to $\mathbf{e}_{r+1}, \mathbf{e}_{r+2}, \cdots, \mathbf{e}_n$, and these successive maxima are taken on when $\mathbf{e}$ is identified the relevant unit characteristic vector.

78. Suppose that Problems 76 and 77 are modified in such a way that $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$ are the characteristic numbers of the problem $\mathbf{a}\, \mathbf{x} = \lambda\, \mathbf{b}\, \mathbf{x}$, where $\mathbf{a}$ and $\mathbf{b}$ are real and symmetric, and also $\mathbf{b}$ is positive definite, and $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$ comprise an orthonormal set of corresponding characteristic vectors, the orthogonality and normality being relative to the matrix $\mathbf{b}$. Show that the results of those Problems again apply if $\mathbf{x}^T \mathbf{x}$ is replaced by $\mathbf{x}^T \mathbf{b}\, \mathbf{x}$ throughout, and if "unit vectors" are of unit length relative to $\mathbf{b}$.

79. Suppose that the characteristic numbers of a real symmetric matrix $\mathbf{a}$ are arranged in order of increasing *absolute value*.

(a) Deduce from the result of Problem 76(b) that the use of equation (232a), in connection with iterative approximation to characteristic quantities, leads to approximations to λ_n which are not greater than λ_n in absolute value.

(b) Show that the use of equation (232b) amounts to approximating λ_n by a ratio of the form

$$\frac{\lambda_1{}^2 c_1{}^2 + \lambda_2{}^2 c_2{}^2 + \cdots + \lambda_n{}^2 c_n{}^2}{\lambda_1 c_1{}^2 + \lambda_2 c_2{}^2 + \cdots + \lambda_n c_n{}^2},$$

and deduce that such an approximation is conservative if all characteristic numbers λ_i are *positive*.

80. (a) If $\mathbf{a}$ is a real symmetric matrix, all of whose elements are *non-negative* ($a_{ij} \geqq 0$), deduce from preceding results that the characteristic number of largest magnitude is *positive* (although its negative *may* then *also* be a characteristic number), and that all components of the correspond-

ing characteristic vector $\mathbf{e}_n$ are of the same sign, and hence may be taken to be all nonnegative. (Consider the nature of $\mathbf{e}^T\,\mathbf{a}\,\mathbf{e}$.)

(b) Show that the result of part (a) is also true of the dominant characteristic quantities for the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$ if $\mathbf{b}$ is real, symmetric, and positive definite, $\mathbf{a}$ is real and symmetric, and $a_{ij} \geqq 0$.

81. (a) If $\mathbf{a}$ is a real symmetric matrix with characteristic numbers λ_i and corresponding characteristic vectors $\mathbf{e}_i$, show that there follows $\mathbf{e}_i^T\,\mathbf{a}\,\mathbf{x} = \lambda_i \mathbf{e}_i^T\,\mathbf{x}$, and hence also

$$\mathbf{e}_i^T(\mathbf{a}\,\mathbf{x} - \lambda_i \mathbf{x}) = 0 \qquad (i = 1, 2, \cdots, n),$$

for *any* real vector $\mathbf{x}$. [Notice that this result follows also from the results of Problem 76(a).]

(b) With the notation $\mathbf{y} = \mathbf{a}\,\mathbf{x}$, for the "transform" of $\mathbf{x}$, deduce that

$$\mathbf{e}_i^T(\mathbf{y} - \lambda_i \mathbf{x}) = 0 \qquad (i = 1, 2, \cdots, n),$$

for any real vector $\mathbf{x}$.

82. Suppose that $\mathbf{a}$ is a real symmetric matrix with *no negative elements*.

(a) Deduce from the results of Problems 80(a) and (81) that the components of the vector $\mathbf{y} - \lambda_n \mathbf{x}$, where $\mathbf{y} \equiv \mathbf{a}\,\mathbf{x}$, then cannot all be of the same sign (unless they all vanish, so that $\mathbf{x}$ is a multiple of $\mathbf{e}_n$).

(b) Deduce that, in this case, if the input $\mathbf{x}$ of the iterative method of Section 1.23 possesses only nonnegative elements, then the dominant characteristic number λ_n is not greater than the largest ratio y_i/x_i of corresponding elements of the output and input vectors, and not less than the smallest such ratio:

$$\min_i \frac{y_i}{x_i} \leqq \lambda_n \leqq \max_i \frac{y_i}{x_i}.$$

83. Prove that the statement of Problem 82(b) is true also for the application of the iterative method to the determination of the dominant characteristic number of the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$ if $\mathbf{b}$ is real, symmetric, and positive definite, whereas $\mathbf{a}$ is real and symmetric and composed only of nonnegative elements. [Use Problem 80(b) and a generalization of Problem 81.]

84. Suppose that $\mathbf{a}$ is a real, symmetric, positive definite matrix such that, whereas all diagonal elements are positive, all elements off the diagonal are either negative or zero. [See, for example, the matrix of coefficients in (265).]

(a) Show that, if α is any positive constant *larger than the largest diagonal element of* $\mathbf{a}$, then the matrix $\mathbf{m} \equiv \alpha\,\mathbf{I} - \mathbf{a}$ is a symmetric matrix, all of whose elements are nonnegative.

(b) Show that the characteristic numbers μ_i of $\mathbf{m}$ are given by $\mu_i = \alpha - \lambda_i$, where λ_i are the characteristic numbers of $\mathbf{a}$, and that the characteristic vector of $\mathbf{m}$ associated with μ_i is that of $\mathbf{a}$ associated with λ_i.

(c) Use the result of Problem 80(a) to show that the largest μ_i is positive. Deduce that the dominant μ_i is $\mu_1 = \alpha - \lambda_1$, where λ_1 is the smallest characteristic number of $\mathbf{a}$, and that all components of the corresponding characteristic vector have the same sign. Hence show that the *smallest* characteristic number of a matrix $\mathbf{a}$ of the type under consideration is not larger than the largest diagonal element of $\mathbf{a}$, and that all components of the associated characteristic vector may be taken as nonnegative. [Notice that the matrix $\mathbf{m}$ can be obtained more easily than the matrix $\mathbf{a}^{-1}$, for the purpose of determining λ_1 and the corresponding characteristic vector by matrix iteration.]

85. Generalize the results and procedures of Problem 84 to the case of the problem $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{b}\,\mathbf{x}$ where $\mathbf{a}$ is a matrix of the type described in that problem, and $\mathbf{b}$ is a positive definite matrix with no negative elements.

Section 1.28.

86. Prove that the relation

$$\| f(x) - g(x) \| = \| f(x) \| + \| g(x) \|$$

is true, over a prescribed interval (a, b), if and only if f and g are orthogonal over (a, b). Notice that if we think of $\| f(x) \| \equiv \int_a^b f^2\,dx$ as the "square of the length of $f(x)$" in the function space relevant to (a, b), and write $l[f] = +\sqrt{\int_a^b f^2\,dx} \equiv +\sqrt{\| f \|}$, this result becomes

$$(l[f])^2 + (l[g])^2 = (l[f - g])^2,$$

and hence is analogous to the Pythagorean theorem (see Figure 1.2).

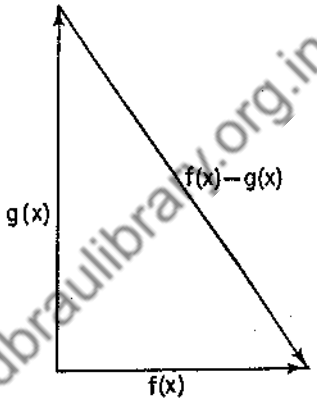


FIGURE 1.2

87. By noticing that, if all integrals are evaluated over an interval (a, b), the quantity

$$\int [f(x) + \lambda\, g(x)]^2\,dx \equiv \int f^2\,dx + 2\lambda \int f\,g\,dx + \lambda^2 \int g^2\,dx$$

is necessarily nonnegative for any real value of λ, deduce that

$$(\int f\,g\,dx)^2 \leqq (\int f^2\,dx)(\int g^2\,dx)$$

and hence

$$|\int f\,g\,dx| \leqq (\sqrt{\int f^2\,dx})(\sqrt{\int g^2\,dx}).$$

This relation is known as the *Schwarz inequality*. Show also that equality holds if and only if $g(x)$ is a constant multiple of $f(x)$.

Deduce that if we define the "angle between $f(x)$ and $g(x)$" in the function space relevant to (a, b) by the equation

$$\cos\theta\,[f, g] = \frac{\int_a^b f\,g\,dx}{\sqrt{\int_a^b f^2\,dx}\,\sqrt{\int_a^b g^2\,dx}} \equiv \frac{(f, g)}{(\mathfrak{l}[f])(\mathfrak{l}[g])},$$

then θ is a *real* angle. Notice that this definition is completely analogous to the geometrical definition of the angle between two vectors (see Figure 1.3).

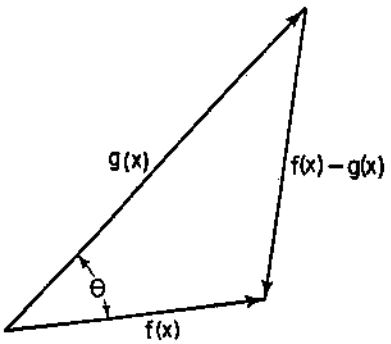


FIGURE 1.3

88. With the terminology of Problems 86 and 87, establish the "law of cosines,"

$$(\mathfrak{l}[f - g])^2 = (\mathfrak{l}[f])^2 + (\mathfrak{l}[g])^2 - 2(\mathfrak{l}[f])(\mathfrak{l}[g])\cos\theta\,[f, g],$$

in function space.

89. Verify the truth of the identity

$$\int (f-g)^2\,dx \equiv \begin{cases} (\sqrt{\int f^2\,dx} - \sqrt{\int g^2\,dx})^2 + 2[\sqrt{\int f^2\,dx}\,\sqrt{\int g^2\,dx} - \int f\,g\,dx] \\ (\sqrt{\int f^2\,dx} + \sqrt{\int g^2\,dx})^2 - 2[\sqrt{\int f^2\,dx}\,\sqrt{\int g^2\,dx} + \int f\,g\,dx] \end{cases}$$

where each integral is evaluated over the interval (a, b). Use the Schwarz inequality (Problem 87) to show that each quantity in square brackets is nonnegative, and hence deduce the relation

$$\left|\sqrt{\int f^2\,dx} - \sqrt{\int g^2\,dx}\right| \leqq \sqrt{\int (f-g)^2\,dx} \leqq \sqrt{\int f^2\,dx} + \sqrt{\int g^2\,dx}$$

or

$$|\mathfrak{l}[f] - \mathfrak{l}[g]| \leqq \mathfrak{l}[f - g] \leqq \mathfrak{l}[f] + \mathfrak{l}[g],$$

where the equality holds if and only if $g(x)$ is a constant multiple of $f(x)$.

To what geometrical relation is this function-theoretical result analogous?

90. As a special case of the Schwarz inequality (Problem 87), deduce that

$$\frac{1}{b-a}\int_a^b f\,dx \leqq \sqrt{\frac{1}{b-a}\int_a^b f^2\,dx}.$$

[The left-hand member is the mean value of $f(x)$ over (a, b), the right-hand member the so-called *root mean square* (rms) value.]

91. Establish the validity of the following statement: "The rms value of the sum of two functions over a given interval is not greater than the sum of the separate rms values, and not less than their difference."

92. Let $f_1(x), f_2(x), \ldots, f_n(x), \ldots$ comprise an infinite set of functions defined over (a, b). Describe a procedure for forming from this set a set of linear combinations $\phi_1(x), \phi_2(x), \ldots, \phi_n(x), \ldots$ which is orthonormal over (a, b). [See Section 1.12.]

93. Show that the functions $f_k(x) = \sin \mu_k x$ $(k = 1, 2, \ldots)$ comprise an orthogonal set over $(0, 1)$ if the constants μ_k satisfy the transcendental equation $\tan \mu_k = \mu_k$.

94. Show that the complex functions $f_k(x) = e^{ikx}$, where k takes on *all* integral values, comprise a set which is orthogonal in the Hermitian sense over any real interval $(a, a + 2\pi)$. Determine the normalizing factors.

Section 1.29.

95. Show that the functions defined in Problem 93 are the characteristic functions of the following Sturm-Liouville problem:

$$\frac{d^2y}{dx^2} + \mu^2 y = 0; \qquad y(0) = 0, \quad y(1) = y'(1).$$

96. Determine the coefficients in the expansion

$$1 = \sum_{k=1}^{\infty} A_k \sin \mu_k x \qquad (0 < x < 1),$$

where $\tan \mu_k = \mu_k$.

97. (a) If a function $F(x)$ possesses the expansion

$$F(x) = \sum_{k=1}^{\infty} A_k \sin \mu_k x \qquad (0 < x < 1),$$

where $\tan \mu_k = \mu_k$, obtain the solution of the problem

$$\frac{d^2y}{dx^2} + \lambda y = F(x); \qquad y(0) = 0, \quad y(1) = y'(1),$$

in the form

$$y(x) = \sum_{k=1}^{\infty} \frac{A_k}{\lambda - \mu_k^2} \sin \mu_k x \qquad (0 < x < 1),$$

when $\lambda \neq \mu_1^2, \mu_2^2, \ldots$

(b) Use the result of Problem 96 to obtain this solution in the special case when $F(x) = 1$.

CHAPTER TWO

Calculus of Variations and Applications

2.1. Maxima and minima. Applications of the calculus of variations are concerned chiefly with determination of maxima and minima of certain expressions involving unknown functions. Certain techniques involved are analogous to procedures in the differential calculus, which are briefly reviewed in this section.

An important problem in the differential calculus is that of determining maximum and minimum values of a function $y = f(x)$ for values of x in a certain interval (a, b). If in that interval $f(x)$ has a continuous derivative, it is recalled that a *necessary* condition for the existence of a maximum or minimum at a point x_0 *inside* (a, b) is that $dy/dx = 0$ at x_0. A *sufficient* condition that y be a maximum (or a minimum) at x_0, relative to values at neighboring points, is that, in addition, $d^2y/dx^2 < 0$ (or $d^2y/dx^2 > 0$) at that point.

If z is a function of two independent variables, say $z = f(x, y)$, in a region R, and if the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ exist and are continuous throughout R, then *necessary* conditions that z possess a relative maximum or minimum at an interior point (x_0, y_0) of R, or that z be *stationary* at that point, are that $\partial z/\partial x = 0$ and $\partial z/\partial y = 0$ simultaneously at (x_0, y_0). These two requirements are equivalent to the single requirement that

$$dz = \frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy = 0$$

at a point (x_0, y_0), for arbitrary values of both dx and dy. *Sufficient* conditions for either a maximum or a minimum involve certain inequalities among the second partial derivatives (see Problem 1).

More generally, a *necessary* condition that a function $f(x_1, x_2, \ldots, x_n)$ of n variables $x_1, \ldots, x_n$ have a stationary value is that

$$df = \frac{\partial f}{\partial x_1}\,dx_1 + \frac{\partial f}{\partial x_2}\,dx_2 + \cdots + \frac{\partial f}{\partial x_n}\,dx_n = 0 \tag{1}$$

for all *permissible* values of the differentials $dx_1, \ldots, dx_n$. If the n variables are all independent, the n differentials can be assigned arbitrarily, and it follows easily that (1) is equivalent to the n conditions

$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial x_2} = \cdots = \frac{\partial f}{\partial x_n} = 0. \tag{2}$$

Sufficient conditions that values of the variables satisfying (1) or (2) actually determine maxima (or minima) involve certain inequalities among the higher partial derivatives (see Problem 1).

Suppose, however, that the n variables are *not* independent, but are related by, say, N conditions each of the form

$$\phi_k(x_1, \cdots, x_n) = 0.$$

Then, at least theoretically, these N equations can generally be solved to express N of the variables in terms of the $n - N$ remaining variables, and hence to express f and df in terms of $n - N$ *independent* variables and their differentials. Alternatively, N linear relations among the n *differentials* can be obtained by differentiation. These conditions permit the expression of N of the *differentials* as linear combinations of the differentials of the $n - N$ independent variables. If (1) is expressed in terms of these differentials, their coefficients must then vanish, giving $n - N$ necessary conditions for stationary values of f which supplement the N constraint conditions.

A procedure which is often still more convenient in this case consists in the introduction of the so-called *Lagrange multipliers*. To illustrate their use, we consider here the problem of obtaining stationary values of $f(x, y, z)$,

$$df \equiv f_x\,dx + f_y\,dy + f_z\,dz = 0, \tag{3}$$

subject to the two constraints

$$\phi_1(x, y, z) = 0, \tag{4a}$$

$$\phi_2(x, y, z) = 0. \tag{4b}$$

Since the three variables x, y, z must satisfy the two auxiliary conditions (4a,b), only one variable can be considered as independent. Equations (4a,b) imply the differential relations

$$\phi_{1x}\,dx + \phi_{1y}\,dy + \phi_{1z}\,dz = 0, \tag{5a}$$

$$\phi_{2x}\,dx + \phi_{2y}\,dy + \phi_{2z}\,dz = 0. \tag{5b}$$

The procedure outlined above would consist in first solving (5a,b) for, say, dx and dy in terms of dz (if this is possible) and in introducing the results into (3), to give a result of the form

$$df \equiv (\cdot\ \cdot\ \cdot)\,dz = 0.$$

Since dz can be assigned arbitrarily, the vanishing of the indicated expression in parentheses in this form is the desired necessary condition that f be a maximum or minimum when (4a, b) are satisfied.

As an alternative procedure, we first multiply (5a) and (5b) respectively by the quantities λ_1 and λ_2, to be specified presently, and add the results to (3). Since the right-hand members are all zeros, there follows

$$(f_x + \lambda_1\phi_{1x} + \lambda_2\phi_{2x})\,dx + (f_y + \lambda_1\phi_{1y} + \lambda_2\phi_{2y})\,dy + (f_z + \lambda_1\phi_{1z} + \lambda_2\phi_{2z})\,dz = 0, \tag{6}$$

for arbitrary values of λ_1 *and* λ_2. Now let λ_1 and λ_2 be determined so that two of the parentheses in (6) vanish. Then the differential multiplying the remaining parenthesis can be arbitrarily assigned, and hence that parenthesis must also vanish. Thus we must have

$$\left.\begin{aligned} \frac{\partial f}{\partial x} + \lambda_1\frac{\partial \phi_1}{\partial x} + \lambda_2\frac{\partial \phi_2}{\partial x} &= 0, \\ \frac{\partial f}{\partial y} + \lambda_1\frac{\partial \phi_1}{\partial y} + \lambda_2\frac{\partial \phi_2}{\partial y} &= 0, \\ \frac{\partial f}{\partial z} + \lambda_1\frac{\partial \phi_1}{\partial z} + \lambda_2\frac{\partial \phi_2}{\partial z} &= 0 \end{aligned}\right\}. \tag{7a,b,c}$$

Equations (7a,b,c) and (4a,b) comprise five equations determining x, y, z and λ_1, λ_2. The quantities λ_1 and λ_2 are known as *Lagrange multipliers*. Their introduction frequently simplifies the relevant algebra in problems of the type just considered. In many applications they are found to have physical significance as well. We

notice that *the conditions* (7) *are necessary conditions that* $f + \lambda_1\phi_1 + \lambda_2\phi_2$ *be stationary when no constraints are present.*

The procedure outlined is applicable without modification to the general case of n variables and $N < n$ constraints.

In illustration of the method, we attempt to determine the point on the curve of intersection of the surfaces

$$z = xy + 5, \qquad x + y + z = 1 \tag{8a,b}$$

which is nearest the origin. Thus, we must minimize the quantity

$$f = x^2 + y^2 + z^2$$

subject to the two constraints (8a,b). With

$$\phi_1 = z - xy - 5, \qquad \phi_2 = x + y + z - 1,$$

equations (7a,b,c) take the form

$$\left.\begin{aligned} 2x - \lambda_1 y + \lambda_2 &= 0, \\ 2y - \lambda_1 x + \lambda_2 &= 0, \\ 2z + \lambda_1 + \lambda_2 &= 0 \end{aligned}\right\}. \tag{9a,b,c}$$

If equations (9a,b) are solved for λ_1 and λ_2, and the results are introduced into (9c), there follows

$$x + y - z + 1 = 0. \tag{10}$$

The simultaneous solution of (8a,b) and (10) leads to the coordinates of the two points (2, −2, 1) and (−2, 2, 1) which are each seen to be three units distant from the origin. Geometrical considerations indicate that there is indeed at least one point nearest the origin; since the two points obtained are *necessarily* the only possible ones, they must accordingly be the points required.

As an illustration closely related to certain topics in Chapter 1, we may seek those points on a central quadric surface

$$\phi \equiv a_{11}x^2 + a_{22}y^2 + a_{33}z^2 + 2a_{12}xy + 2a_{23}yz + 2a_{13}xz = \text{constant}$$

for which distance from the origin is stationary relative to neighboring points. We are thus to render the form

$$f \equiv x^2 + y^2 + z^2$$

stationary, subject to the constraint $\phi =$ constant. Here, if we denote the Lagrange multiplier by $-1/\lambda$, the requirement that $\phi - \lambda f$ be stationary leads to the conditions

$$\left.\begin{aligned} a_{11}x + a_{12}y + a_{13}z &= \lambda x, \\ a_{12}x + a_{22}y + a_{23}z &= \lambda y, \\ a_{13}x + a_{23}y + a_{33}z &= \lambda z \end{aligned}\right\}.$$

This set of equations comprises a characteristic-value problem of the type discussed in Section 1.11. Each "characteristic value" of λ, for which a nontrivial solution exists, leads to the three coordinates of one or more *points P:* (x, y, z), determined within a common arbitrary multiplicative factor which is available for the satisfaction of the equation of the surface. Sections 1.19 and 1.20 show that it is always possible to rotate the coordinate axes in such a way that each new axis coincides with the direction from the origin to such a point, and that the equation of the surface, referred to the new axes, then involves only squares of the new coordinates. That is, the new axes (which coincide with the "characteristic vectors" of the problem) are the *principal axes* of the quadric surface. The characteristic values of λ are inversely proportional to the squares of the semiaxes. Repeated roots of the characteristic equation correspond to surfaces of revolution, in which cases the new axes can be so chosen in infinitely many ways, while zero roots correspond to surfaces which extend infinitely far from the origin.

The basic problem in the *calculus of variations* is to determine a *function* such that a certain definite integral involving that function and certain of its derivatives takes on a maximum or minimum value. The elementary part of the theory is concerned with a *necessary* condition (generally in the form of a differential equation with boundary conditions) which the required function must satisfy. To show mathematically that the function obtained actually maximizes (or minimizes) the integral is much more difficult than in the corresponding problems of differential calculus. *Sufficient* conditions are developed in more advanced works. In physically motivated problems, such additional considerations may frequently be avoided.

As an example of a problem of this sort, we notice that in order to determine the surface of revolution, obtained by rotating about

the x-axis a curve passing through two given points (x_1, y_1) and (x_2, y_2), which has minimum surface area, we must determine the function $y(x)$ which specifies the curve to be revolved, in such a way that the integral

$$I = 2\pi \int_{x_1}^{x_2} y(1 + y'^2)^{\frac{1}{2}}\, dx$$

is a minimum, and also so that $y(x_1) = y_1$ and $y(x_2) = y_2$.

In most cases it is to be required that the function and the derivatives explicitly involved be continuous in the region of definition.

2.2. The simplest case. We now consider the problem of determining a function $y(x)$ which makes the integral

$$I = \int_{x_1}^{x_2} F(x, y, y')\, dx \tag{11}$$

stationary,* and which satisfies the prescribed end conditions

$$y(x_1) = y_1, \qquad y(x_2) = y_2.$$

To fix ideas, we may suppose that I is to be *minimized.*

Suppose that $y(x)$ is the actual minimizing function, and choose *any* continuously differentiable function $\eta(x)$ which *vanishes* at the end points $x = x_1$ and $x = x_2$. Then for any constant ϵ the function $y(x) + \epsilon\,\eta(x)$ will satisfy the end conditions (Figure 2.1). The integral

$$I(\epsilon) = \int_{x_1}^{x_2} F(x, y + \epsilon\,\eta, y' + \epsilon\,\eta')\, dx, \tag{12}$$

obtained by replacing y by $y + \epsilon\,\eta$ in (11), is then a function of ϵ, once y and η are assigned, which takes on its minimum value when $\epsilon = 0$. But this is possible only if

$$\frac{dI(\epsilon)}{d\epsilon} = 0 \quad \text{when} \quad \epsilon = 0. \tag{13}$$

If we denote the integrand in (12) by F_ϵ,

$$F_\epsilon = F(x, y + \epsilon\,\eta, y' + \epsilon\,\eta'),$$

and notice that then

$$\frac{dF_\epsilon}{d\epsilon} = \frac{\partial F_\epsilon}{\partial y}\,\eta + \frac{\partial F_\epsilon}{\partial y'}\,\eta',$$

* We suppose that F has continuous second partial derivatives with respect to its three arguments.

we obtain from (12) the result

$$\frac{dI(\epsilon)}{d\epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial F_\epsilon}{\partial y} \eta + \frac{\partial F_\epsilon}{\partial y'} \frac{d\eta}{dx} \right) dx,$$

by differentiating under the integral sign. Finally, since $F_\epsilon \to F$ when $\epsilon \to 0$, and the same is true of the partial derivatives, the

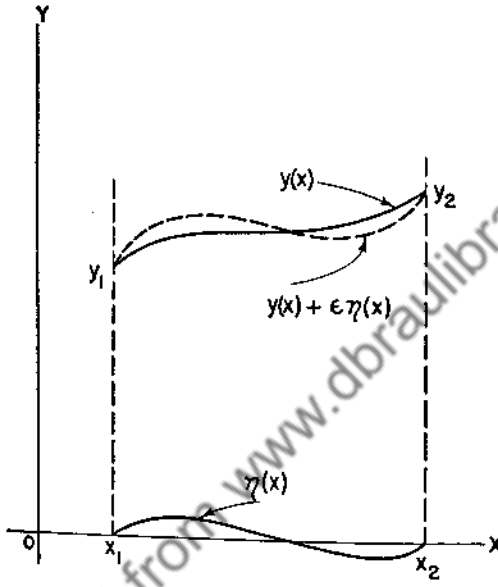


FIGURE 2.1

necessary condition (13) takes the form

$$\int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y} \eta + \frac{\partial F}{\partial y'} \frac{d\eta}{dx} \right) dx = 0. \tag{14}$$

The next step in the development consists in integrating the second term by parts, to transform (14) to the condition

$$\int_{x_1}^{x_2} \left[\frac{\partial F}{\partial y} \eta - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \eta \right] dx + \left[\frac{\partial F}{\partial y'} \eta(x) \right]_{x_1}^{x_2} = 0. \tag{15}$$

But since $\eta(x)$ vanishes at the end points, by assumption, the integrated terms vanish and (15) becomes

$$\int_{x_1}^{x_2} \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \right] \eta \, dx = 0. \tag{16}$$

Finally, since $\eta(x)$ is arbitrary, we conclude that its coefficient in (16) must vanish identically over (x_1, x_2). For if this were not so we could choose a continuously differentiable function $\eta(x)$ in such a way that the (continuous) integrand in (16) is positive whenever it is not zero,* and a contradiction would be obtained.

The end result is that if $y(x)$ minimizes (or maximizes) the integral (11), it must satisfy the *Euler equation*

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) - \frac{\partial F}{\partial y} = 0. \tag{17a}$$

Here the partial derivatives $\partial F/\partial y$ and $\partial F/\partial y'$ have been formed by treating x, y, and y' as independent variables. Remembering that $\partial F/\partial y'$ is, in general, a function of x explicitly and also implicitly through y and $y' = dy/dx$, the first term in (17a) can be written in the expanded form

$$\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial y'}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial y'}\right)\frac{dy}{dx} + \frac{\partial}{\partial y'}\left(\frac{\partial F}{\partial y'}\right)\frac{dy'}{dx}.$$

Thus (17a) is equivalent to the equation

$$F_{y'y'}\frac{d^2y}{dx^2} + F_{y'y}\frac{dy}{dx} + (F_{y'x} - F_y) = 0. \tag{17b}$$

This equation is of second order in y unless $F_{y'y'} = \partial^2 F/\partial y'^2 \equiv 0$, so that in general two constants are available for the satisfaction of the end conditions.

It is useful to notice that (17b) is equivalent to the form

$$\frac{1}{y'}\left[\frac{d}{dx}\left(F - \frac{\partial F}{\partial y'}\frac{dy}{dx}\right) - \frac{\partial F}{\partial x}\right] = 0, \tag{17c}$$

as can be verified by expansion. From this result it follows that if F does not involve x *explicitly* a first integral of Euler's equation is

$$F - y'\frac{\partial F}{\partial y'} = C \quad \text{if} \quad \frac{\partial F}{\partial x} \equiv 0, \tag{18a}$$

while (17a) shows that if F does not involve y explicitly a first integral is

$$\frac{\partial F}{\partial y'} = C \quad \text{if} \quad \frac{\partial F}{\partial y} \equiv 0. \tag{18b}$$

* This fact, which is intuitively plausible, can be proved analytically.

Solutions of Euler's equation are known as *extremals* of the problem considered. In general, they comprise a two-parameter family of functions in the case just considered.

We may notice also that if at one (or both) of the end points $y(x)$ is *not* prescribed, then the function $\eta(x)$ need not vanish at this point. Reference to (15) then shows that the Euler equation still follows if at that point the condition

$$\frac{\partial F}{\partial y'} = 0 \quad \text{at} \quad x = x_1 \quad \text{or} \quad x = x_2 \tag{19}$$

is imposed instead. This condition is known as a *natural boundary condition.*

2.3. Illustrative examples. In Section 2.1 it was pointed out that to find the minimal surface of revolution passing through two given points it is necessary to minimize the integral

$$\frac{I}{2\pi} = \int_{x_1}^{x_2} y(1 + y'^2)^{\frac{1}{2}}\, dx. \tag{20}$$

With $F = y(1 + y'^2)^{\frac{1}{2}}$, the Euler equation (17a) becomes

$$\frac{d}{dx}\left[\frac{y\,y'}{(1 + y'^2)^{\frac{1}{2}}}\right] - (1 + y'^2)^{\frac{1}{2}} = 0$$

or, after a reduction or use of (17b),

$$y\,y'' - y'^2 - 1 = 0. \tag{21}$$

Following the usual procedure for solving equations of this type, we set

$$y' = p, \qquad y'' = \frac{dp}{dx} = p\,\frac{dp}{dy},$$

so that (21) becomes

$$p\,y\,\frac{dp}{dy} = p^2 + 1.$$

This equation is separable, and is integrated to give

$$y = c_1(1 + p^2)^{\frac{1}{2}} \equiv c_1\left[1 + \left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}},$$

as would be obtained more directly by use of (18a), since here F does not explicitly involve x. There follows

$$\frac{dy}{dx} = \left(\frac{y^2}{c_1^2} - 1\right)^{1/2},$$

and hence finally

$$y = c_1 \cosh\left(\frac{x}{c_1} + c_2\right). \tag{22}$$

Thus, as is well known, the required minimal surface (if it exists) must be obtained by revolving a catenary. It then remains to be seen whether the arbitrary constants c_1 and c_2 can indeed be so chosen that the curve (22) passes through any two assigned points in the upper half plane.

The determination of these constants is found to involve the solution of a transcendental equation which possesses two, one, or no solutions, depending upon the prescribed values $y(x_1)$ and $y(x_2)$.

The classical "elementary" application of the calculus of variations consists in *proving* mathematically that the shortest distance between two points in a plane is a straight line. If the points, in the xy-plane, are (x_1, y_1) and (x_2, y_2) and the equation of the minimizing curve is $y = y(x)$, we are then to minimize

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

Since here $F = (1 + y'^2)^{1/2}$ does not involve either x or y explicitly, either of the forms (18a,b) can be used to give a first integral of Euler's equation directly. If form (18b) is used, there follows

$$y' = C(1 + y'^2)^{1/2},$$

and consequently $y' = c_1$ or $y = c_1 x + c_2$. From this result we can conclude that *if a minimizing curve exists and if it can be specified by an equation of the form $y = y(x)$, then that curve must necessarily be a straight line.* It is clear that the case in which $x_1 = x_2$ is exceptional, and must be treated separately.

In the preceding examples no proof was given that the curve obtained actually possesses the required minimizing property. Such considerations comprise most of the less elementary theory of the calculus of variations. In a great number of physically motivated problems it is intuitively clear that a minimizing function does indeed *exist*. Then if the present methods and their extensions show that only the particular function obtained could

possibly be the minimizing function, the problem can be considered as solved for practical purposes. If several alternatives are determined, direct calculation will show which one actually leads to the smaller value of the quantity to be minimized.

Further applications are deferred until the more general theory has been established.

2.4. The variational notation. We next introduce the notation of "variations" in order to establish more clearly the analogy between the calculus of variations and the differential calculus.

Suppose that we consider a set S of functions satisfying certain conditions. For example, we might define S to be the set of all functions of a single variable x which possess a continuous first derivative at all points in an interval $a \leqq x \leqq b$. Then any quantity which takes on a specific numerical value corresponding to each function in S is said to be a *functional*.

In illustration, we may speak of the quantities

$$I_1 = \int_a^b y(x)\,dx, \quad I_2 = \int_a^b \{y(x)y''(x) - [y'(x)]^2\}\,dx$$

as functionals, since corresponding to any function $y(x)$ for which the indicated operations are defined each quantity has a definite numerical value.

With the above definition, it is proper also to speak of such quantities as $f[y(x)]$ and $g[x, y(x), y'(x), \ldots, y^{(n)}(x)]$ as functionals in those cases when the variable x is considered as fixed in a given discussion and the function $y(x)$ is varied.

In Section 2.2, we considered an integrand of the form

$$F = F(x, y, y')$$

which *for a fixed value of* x depends upon the *function* $y(x)$ and its derivative. We then changed the function $y(x)$, to be determined, into a new function $y(x) + \epsilon\,\eta(x)$. The change $\epsilon\,\eta(x)$ in $y(x)$ is called the *variation* of y and is conventionally denoted by δy,

$$\delta y \equiv \epsilon\,\eta(x). \tag{23}$$

Corresponding to this change in $y(x)$, for a fixed value of x, the functional F changes by an amount ΔF, where

$$\Delta F = F(x, y + \epsilon\eta, y' + \epsilon\eta') - F(x, y, y'). \tag{24}$$

If the right-hand member is expanded in powers of ϵ, there follows

$$\Delta F = \frac{\partial F}{\partial y}\epsilon\eta + \frac{\partial F}{\partial y'}\epsilon\eta' + (\text{terms involving higher powers of } \epsilon). \tag{25}$$

In analogy with the definition of the *differential*, the first two terms in the right-hand member of (25) are *defined* to be the *variation of* F,

$$\delta F = \frac{\partial F}{\partial y}\epsilon\eta + \frac{\partial F}{\partial y'}\epsilon\eta',$$

or, using (23),

$$\delta F = \frac{\partial F}{\partial y}\delta y + \frac{\partial F}{\partial y'}\delta y'. \tag{26}$$

For a complete analogy with the definition of the differential, we would perhaps anticipate the definition

$$\delta F = \frac{\partial F}{\partial x}\delta x + \frac{\partial F}{\partial y}\delta y + \frac{\partial F}{\partial y'}\delta y'.$$

But here x *is not varied*, so that we have

$$\delta x \equiv 0, \tag{27}$$

and hence the analogy is indeed complete.

We notice that the *differential* of a function is a first-order approximation to the change in that function *along a particular curve*, while the *variation* of a functional is a first-order approximation to the change *from curve to curve*.

It is easily verified directly, from the definition, that the laws of variation of sums, products, ratios, powers, and so forth, are completely analogous to the corresponding laws of differentiation. Thus, for example, there follows

$$\delta(F_1F_2) = F_1\,\delta F_2 + F_2\,\delta F_1,$$

$$\delta\left(\frac{F_1}{F_2}\right) = \frac{F_2\,\delta F_1 - F_1\,\delta F_2}{F_2^2},$$

and so forth.

Analogous definitions are introduced in the more general case. Thus, for example, if x and y are independent variables, and u and v are dependent variables, we may consider a functional

$$F = F(x, y, u, v, u_x, u_y, v_x, v_y).$$

We now vary both u and v, holding x and y fixed, into new functions $u + \epsilon\,\xi$ and $v + \epsilon\,\eta$, and define the variations of u and v as follows:

$$\delta u = \epsilon\,\xi(x, y), \qquad \delta v = \epsilon\,\eta(x, y). \tag{28}$$

The change in F is then found (by expansion in powers of ϵ) to be

$$\Delta F = \frac{\partial F}{\partial u}\epsilon\,\xi + \frac{\partial F}{\partial v}\epsilon\,\eta + \frac{\partial F}{\partial u_x}\epsilon\,\xi_x + \frac{\partial F}{\partial u_y}\epsilon\,\xi_y + \frac{\partial F}{\partial v_x}\epsilon\,\eta_x + \frac{\partial F}{\partial v_y}\epsilon\,\eta_y$$
$$+ (\text{terms involving higher powers of } \epsilon). \tag{29}$$

The first-order terms are *defined* to comprise the variation of F. Hence, using (28), we have the definition

$$\delta F = \frac{\partial F}{\partial u}\delta u + \frac{\partial F}{\partial v}\delta v + \frac{\partial F}{\partial u_x}\delta u_x + \frac{\partial F}{\partial u_y}\delta u_y + \frac{\partial F}{\partial v_x}\delta v_x + \frac{\partial F}{\partial v_y}\delta v_y. \tag{30}$$

Since the independent variables x and y are held fixed, there follows also

$$\delta x = \delta y \equiv 0. \tag{31}$$

From (23) we obtain the result

$$\frac{d}{dx}(\delta y) = \epsilon\frac{d\eta}{dx} = \delta\frac{dy}{dx}.$$

Hence, *if x is the independent variable* (and, accordingly, $\delta x \equiv 0$) *the operators δ and d/dx are commutative:*

$$\frac{d}{dx}\delta y = \delta\frac{dy}{dx}. \tag{32}$$

Similarly, from (28) we find that *if x and y are independent variables* (and hence $\delta x = \delta y \equiv 0$) *the operators δ and $\partial/\partial x$ or $\partial/\partial y$ are commutative:*

$$\frac{\partial}{\partial x}\delta u = \delta\frac{\partial u}{\partial x}, \qquad \frac{\partial}{\partial y}\delta u = \delta\frac{\partial u}{\partial y}. \tag{33}$$

That is, *the derivative of the variation with respect to an independent variable is the same as the variation of the derivative.*

It should be noticed that this is not generally true unless the differentiation is with respect to an independent variable. For if

x and y are both functions of an independent variable t we may write

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{y'}{x'}$$

where a prime now denotes t-differentiation. Thus we then have

$$\delta \frac{dy}{dx} = \delta\left(\frac{y'}{x'}\right) = \frac{x'\,\delta y' - y'\,\delta x'}{x'^2}.$$

But now δ and d/dt *are* commutative, so that

$$\delta \frac{dy}{dx} = \frac{\dfrac{dx}{dt}\dfrac{d}{dt}(\delta y) - \dfrac{dy}{dt}\dfrac{d}{dt}(\delta x)}{\left(\dfrac{dx}{dt}\right)^2} = \frac{\dfrac{d}{dt}(\delta y)}{\dfrac{dx}{dt}} - \frac{\dfrac{dy}{dt}\dfrac{d}{dt}(\delta x)}{\dfrac{dx}{dt}\dfrac{dx}{dt}}$$

or, finally,

$$\delta \frac{dy}{dx} = \frac{d}{dx}\,\delta y - \frac{dy}{dx}\frac{d}{dx}\,\delta x. \tag{32a}$$

If $\delta x \equiv 0$, equation (32a) reduces to (32).

The quantity δF is sometimes called the *first* variation of F, the *second* variation then being defined as the group of second-order terms in ϵ in (25) or (29). However, when the term "variation" is used alone the *first* variation is generally implied.

For a functional expressed as a definite integral,

$$I = \int_{x_1}^{x_2} F(x, y, y')\,dx, \tag{34a}$$

where x is the independent variable, there follows from the definition

$$\delta I = \delta \int_{x_1}^{x_2} F\,dx = \int_{x_1}^{x_2} \delta F\,dx, \tag{34b}$$

so that variation and integration between limits (which do not involve the dependent variable) are commutative processes.

We now show that *a necessary condition that the integral*

$$I = \int_{x_1}^{x_2} F(x, y, y')\,dx \tag{35}$$

be stationary is that its (first) variation vanish:

$$\delta I \equiv \delta \int_{x_1}^{x_2} F(x, y, y')\,dx = 0. \tag{36}$$

According to (34) and (26), equation (36) is equivalent to

$$\delta I = \int_{x_1}^{x_2} \delta F(x, y, y')\, dx = \int_{x_1}^{x_2} \left[\frac{\partial F}{\partial y}\, \delta y + \frac{\partial F}{\partial y'}\, \delta y' \right] dx = 0.$$

If we replace $\delta y'$ by $d(\delta y)/dx$, in accordance with (32), and integrate by parts, there follows

$$\delta I = \int_{x_1}^{x_2} \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'} \right) \right] \delta y\, dx + \left[\frac{\partial F}{\partial y'}\, \delta y \right]_{x_1}^{x_2}. \tag{37}$$

But the right-hand member of (37) is proportional to the left-hand member of (15), the vanishing of which was shown to be necessary if I is to be stationary, and by retracing steps we see that (15) or (37) implies (36), as was to be shown.

The use of the variational notation leads to concise derivations and computations. This notation will be used in the remainder of this chapter; its justification in any particular case follows the lines of the preceding argument.

2.5. The more general case. We consider next the case when the integral to be made stationary is of the form

$$I = \iint_R F(x, y, u, v, u_x, u_y, v_x, v_y)\, dx\, dy. \tag{38}$$

Here x and y are independent variables, u and v are functions of x and y to be determined, and the integration is carried out over a simple two-dimensional region R of the xy-plane. We suppose that the values of $u(x, y)$ and $v(x, y)$ are prescribed along the boundary C of the region R.

The condition

$$\delta I = 0 \tag{39}$$

then becomes

$$\delta I = \iint_R \left[\left(\frac{\partial F}{\partial u}\, \delta u + \frac{\partial F}{\partial u_x}\, \delta u_x + \frac{\partial F}{\partial u_y}\, \delta u_y \right) + \left(\frac{\partial F}{\partial v}\, \delta v + \frac{\partial F}{\partial v_x}\, \delta v_x + \frac{\partial F}{\partial v_y}\, \delta v_y \right) \right] dx\, dy = 0. \tag{40}$$

Here the variations δu and δv are to be continuously differentiable over R and are to *vanish* on the boundary C, but are otherwise completely arbitrary.

The terms involving variations of derivatives are next to be integrated by parts. The general procedure may be illustrated by considering the treatment of a typical term:

$$\iint_R \frac{\partial F}{\partial u_x}\,\delta\frac{\partial u}{\partial x}\,dx\,dy = \int_a^b \left[\int_{x_1(y)}^{x_2(y)} \frac{\partial F}{\partial u_x}\frac{\partial}{\partial x}\,\delta u\,dx\right] dy.$$

(See Figure 2.2.) If the inner integral is integrated by parts, this term becomes

$$\int_a^b \left\{\left[\frac{\partial F}{\partial u_x}\,\delta u\right]_{x=x_1}^{x=x_2} - \int_{x_1}^{x_2}\left[\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right)\delta u\right] dx\right\} dy,$$

where x_1 and x_2 are the extreme (boundary) values of x corresponding to the value of y held constant in the x-integration. But since

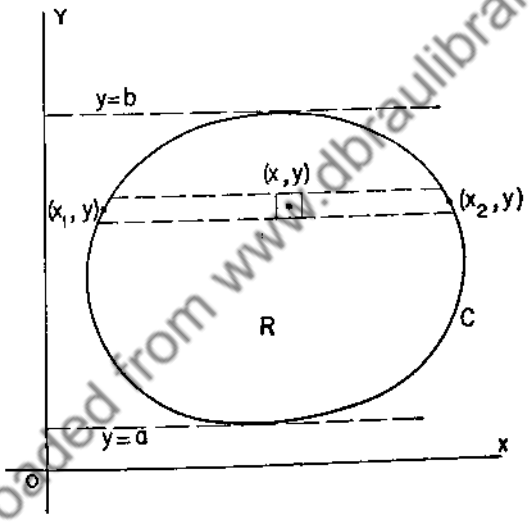


FIGURE 2.2

δu is required to vanish along the boundary, the partially integrated terms vanish and there follows*

$$\iint_R \frac{\partial F}{\partial u_x}\,\delta\frac{\partial u}{\partial x}\,dx\,dy = -\iint_R \left[\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right)\right]\delta u\,dx\,dy.$$

If the other terms of this type in (40) are treated similarly, (40) takes the form

* If lines parallel to the axes intersect C in more than two points, the region R must be appropriately subdivided in this derivation.

$$\delta I = \iint_R \left\{ \left[\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) \right] \delta u + \left[\frac{\partial F}{\partial v} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial v_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial v_y}\right) \right] \delta v \right\} dx\, dy = 0. \tag{41}$$

If the variations δu and δv are independent of each other, that is, if u and v can be varied independently, then as in an earlier argument it follows that the coefficients of δu and δv in (41) must *each* vanish identically in R. Thus two Euler equations are obtained in the form

$$\left.\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) - \frac{\partial F}{\partial u} = 0, \\ \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial v_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial v_y}\right) - \frac{\partial F}{\partial v} = 0 \end{aligned}\right\}. \tag{42a,b}$$

These conditions comprise two partial differential equations in u and v and are, in general, linear or quasi-linear of second order in u and v, as can be shown by expansion. We notice that in the differentiations with respect to u, v, u_x, u_y, v_x, and v_y all of the *eight* variables listed in (38) are treated as though they were independent. Thus $\partial F/\partial u_x$ is formed by holding x, y, u, v, u_y, v_x, and v_y constant in the expression for F. However, in the differentiations with respect to x or y only these *two* variables are treated as independent. Since, in general, $\partial F/\partial u_x$ will involve x (and y) both explicitly and implicitly, the *first term* in (42a) becomes, on expansion,

$$(F_{u_x})_x + (F_{u_x})_u \frac{\partial u}{\partial x} + (F_{u_x})_v \frac{\partial v}{\partial x} + (F_{u_x})_{u_x} \frac{\partial u_x}{\partial x} + (F_{u_x})_{u_y} \frac{\partial u_y}{\partial x} + (F_{u_x})_{v_x} \frac{\partial v_x}{\partial x} + (F_{u_x})_{v_y} \frac{\partial v_y}{\partial x}$$

or

$$F_{u_x u_x} \frac{\partial^2 u}{\partial x^2} + F_{u_x u_y} \frac{\partial^2 u}{\partial x\, \partial y} + F_{u u_x} \frac{\partial u}{\partial x} + F_{u_x v_x} \frac{\partial^2 v}{\partial x^2} + F_{u_x v_y} \frac{\partial^2 v}{\partial x\, \partial y} + F_{v u_x} \frac{\partial v}{\partial x} + F_{x u_x}.$$

Equations (42a,b) represent *necessary* conditions that I of equation (38) be stationary. They are subject, of course, to the

prescribed boundary conditions along C. The formulation of *sufficient* conditions is again extremely involved, and is omitted here.

Completely analogous equations are obtained in the general case of m dependent and n independent variables. Here m equations analogous to (42) are obtained, each having $n + 1$ terms. Still more generally, partial derivatives of higher order than the first may be involved in the integral to be made stationary. The extension of the present methods to such cases is straightforward.

If the integrand F involves n independent variables $x, y, \ldots,$ and m dependent variables $u, v, \ldots,$ together with partial derivatives, of various orders, of $u, v, \ldots$ with respect to $x, y, \ldots,$ one obtains one Euler equation for each of the m dependent variables. The equation corresponding to u is then of the form

$$F_u - \left(\frac{\partial}{\partial x} F_{u_x} + \frac{\partial}{\partial y} F_{u_y} + \cdots\right) + \left(\frac{\partial^2}{\partial x^2} F_{u_{xx}} + \frac{\partial^2}{\partial x\,\partial y} F_{u_{xy}} + \frac{\partial^2}{\partial y^2} F_{u_{yy}} + \cdots\right) - \left(\frac{\partial^3}{\partial x^3} F_{u_{xxx}} + \cdots\right) + \left(\frac{\partial^4}{\partial x^4} F_{u_{xxxx}} + \cdots\right) - \cdots = 0. \tag{43}$$

As an application of the preceding results, we obtain the partial differential equation satisfied by the equation of a *minimal surface*, that is, the surface passing through a given simple closed curve c in space and having minimum surface area bounded by c. If the equation of the surface is assumed to be expressible in the form $z = z(x, y)$, the area to be minimized is then given by the integral

$$S = \iint_R (1 + z_x^2 + z_y^2)^{\frac{1}{2}}\, dx\, dy, \tag{44}$$

where R is the region in the xy-plane bounded by the projection C of c onto the xy-plane. With $F = (1 + z_x^2 + z_y^2)^{\frac{1}{2}}$, the Euler equation is

$$\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial z_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial z_y}\right) - \frac{\partial F}{\partial z} = 0$$

or

$$\frac{\partial}{\partial x}\left[\frac{z_x}{(1 + z_x^2 + z_y^2)^{\frac{1}{2}}}\right] + \frac{\partial}{\partial y}\left[\frac{z_y}{(1 + z_x^2 + z_y^2)^{\frac{1}{2}}}\right] = 0.$$

After some reduction, this equation takes the form

$$(1 + z_y^2)z_{xx} - 2z_x z_y z_{xy} + (1 + z_x^2)z_{yy} = 0,$$

or, with the conventional abbreviations for the partial derivatives,

$$p = z_x, \quad q = z_y, \quad r = z_{xx}, \quad s = z_{xy}, \quad t = z_{yy},$$

the differential equation of minimal surfaces becomes

$$(1 + q^2)r - 2pq\,s + (1 + p^2)t = 0. \tag{45}$$

As a second example, we seek the function $\phi(x, y, z)$ for which the mean square value of the gradient over a certain region R of space is minimum. This problem is closely related to many physical considerations, as will be seen. A necessary condition for the determination of ϕ is then

$$\delta \iiint_R (\phi_x^2 + \phi_y^2 + \phi_z^2)\,dx\,dy\,dz = 0. \tag{46}$$

With $F = \phi_x^2 + \phi_y^2 + \phi_z^2$, the Euler equation can be obtained by reference to (43), in the form

$$\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \phi_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial \phi_y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial F}{\partial \phi_z}\right) = 0,$$

since F does not involve ϕ explicitly. This equation reduces to

$$\phi_{xx} + \phi_{yy} + \phi_{zz} \equiv \nabla^2\phi = 0, \tag{47}$$

so that ϕ must satisfy *Laplace's equation*.

Conversely, suppose that ϕ satisfies (47) everywhere in a region R, and takes on prescribed values on the boundary of that region. We may then multiply both sides of (47) by any continuously differentiable variation $\delta\phi$ which vanishes on the boundary of R, and integrate the results over R to obtain

$$\iiint_R (\phi_{xx} + \phi_{yy} + \phi_{zz})\,\delta\phi\,dx\,dy\,dz = 0.$$

If the first term is transformed by integration by parts, there follows

$$\iint \left\{\int_{x_1}^{x_2} \phi_{xx}\,\delta\phi\,dx\right\} dy\,dz = \iint \left\{\Big[\phi_x\,\delta\phi\Big]_{x_1}^{x_2} - \int_{x_1}^{x_2} \phi_x\,\delta\phi_x\,dx\right\} dy\,dz$$

$$= -\iiint \phi_x \, \delta\phi_x \, dx \, dy \, dz$$

$$= -\frac{1}{2} \, \delta \iiint \phi_x{}^2 \, dx \, dy \, dz.$$

By treating the other terms similarly, we thus recover (46) from (47) if ϕ is prescribed on the boundary of R. Thus, in this sense, the two problems are *equivalent.*

Hence, the so-called *Dirichlet problem,* in which we seek a function which satisfies Laplace's equation in a region R and which takes on prescribed values along the boundary of R, can be expressed as a variational problem. As will be shown, the variational problem (46) can often be treated conveniently by approximate methods, to yield an approximate solution to the corresponding Dirichlet problem.

2.6. Constraints and Lagrange multipliers. In certain cases in which one or more functions are to be determined by a variational procedure, the variations cannot all be arbitrarily assigned, but are governed by one or more auxiliary conditions or *constraints.* Methods analogous to those described in Section 2.1 are available for the treatment of such cases.

We illustrate the procedure first in the special case of two dependent variables u and v, and one independent variable x,

$$\delta \int_{x_1}^{x_2} F(x, u, v, u_x, v_x) \, dx = 0, \tag{48}$$

in which the constraint is of the form

$$\phi(u, v) = 0. \tag{49}$$

We again require that the variations of u and v vanish at the end points. Then, by proceding as in Section 2.5, equation (48) is transformed into the condition

$$\int_{x_1}^{x_2} \left\{ \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) \right] \delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right) \right] \delta v \right\} dx = 0, \tag{50}$$

Since u and v must satisfy (49), the variations δu and δv cannot *both* be assigned arbitrarily inside (x_1, x_2) so that their coefficients in (50) need not vanish separately. However, from (49) there follows $\delta\phi = 0$, or

$$\phi_u\,\delta u + \phi_v\,\delta v = 0. \tag{51}$$

If we multiply (51) by a quantity λ (a *Lagrange multiplier*), which may be a function of x, and integrate the result with respect to x over (x_1, x_2), there follows

$$\int_{x_1}^{x_2} (\lambda\,\phi_u\,\delta u + \lambda\,\phi_v\,\delta v)\,dx = 0 \tag{52}$$

for any λ. The result of adding (50) and (52),

$$\int_{x_1}^{x_2} \left\{ \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) + \lambda\,\phi_u \right] \delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right) + \lambda\,\phi_v \right] \delta v \right\} dx = 0, \tag{53}$$

must then also be true for any λ. Let λ be chosen so that, say, the coefficient of δu in (53) vanishes. Then, since the single variation δv *can* be arbitrarily assigned inside (x_1, x_2), its coefficient must also vanish.* Thus we must have

$$\left.\begin{aligned} \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial F}{\partial u} - \lambda\,\phi_u = 0, \\ \frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right) - \frac{\partial F}{\partial v} - \lambda\,\phi_v = 0 \end{aligned}\right\}. \tag{54a,b}$$

Equation (54a,b) and (49) comprise three equations in the three functions u, v, and λ.

If λ is eliminated between (54a) and (54b), the result

$$\phi_v\left[\frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial F}{\partial u}\right] - \phi_u\left[\frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right) - \frac{\partial F}{\partial v}\right] = 0, \tag{54c}$$

together with (49), gives two conditions governing u and v. It may be noticed that this same relation would be obtained more directly by solving (51) for, say, δu as a multiple of δv, introducing the result into (50), and equating to zero the resultant coefficient of δv in (50). In more involved cases, the use of Lagrange multipliers is frequently more advantageous.

* If ϕ_u vanishes identically, λ is to be chosen such that the coefficient of δv vanishes. Clearly, ϕ_u and ϕ_v cannot *both* vanish identically.

The extension to the more general case is perfectly straightforward.

In some cases the constraint condition is prescribed directly in the variational form

$$f\,\delta u + g\,\delta v = 0,$$

rather than in the form of (49). Whether or not a function ϕ can be found whose variation is given by the left-hand member, the preceding derivation shows that the required necessary conditions are given by replacing ϕ_u and ϕ_v by f and g, respectively, in (54a,b) or (54c).

Also, a constraint condition may be expressed by the requirement that a certain definite *integral* involving the unknown function or functions take on a prescribed value. We illustrate a procedure which may be used in such cases by supposing that the minimal condition is of the form

$$\delta I = \delta \int_{x_1}^{x_2} F(x, y, y')\,dx = 0 \tag{55}$$

where y is to take on prescribed values at the ends of the interval of definition, and also is to satisfy the constraint condition

$$\int_{x_1}^{x_2} G(x, y, y')\,dx = K \tag{56}$$

where K is a prescribed constant. As before, if y is the minimizing function, the condition (55) leads to the requirement that

$$\int_{x_1}^{x_2} \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right] \delta y\,dx = 0, \tag{57}$$

for any *admissible* continuously differentiable δy which vanishes at the end points. But now δy is *not* completely arbitrary inside the interval (x_1, x_2), since both y and $y + \delta y$ must satisfy the constraint (56).

In order to specify a set of *admissible* forms of δy, we write

$$\delta y = \epsilon f(x) + \alpha g(x), \tag{58}$$

where $f(x)$ and $g(x)$ are continuously differentiable functions which vanish at the end points, and ϵ and α are constants. The require-

ment that $y + \delta y$ then satisfy the constraint condition is of the form

$$J(\epsilon, \alpha) \equiv \int_{x_1}^{x_2} G(x, y + \epsilon f + \alpha g, y' + \epsilon f' + \alpha g')\, dx = K. \quad (59)$$

We now propose first to *fix* the function $g(x)$ and, for any subsequently chosen $f(x)$, then to determine α as a function of ϵ in such a way that (59) is satisfied. It is recalled that an equation of the form $J(\epsilon, \alpha) = K$, which is satisfied by $\epsilon = \alpha = 0$, determines α as a function of ϵ for small values of ϵ and α if $\partial J/\partial \alpha \neq 0$ when $\epsilon = \alpha = 0$. If we denote the integrand in (59) by $\bar{G}$, noticing that $\bar{G}$ tends to G as ϵ and α tend to zero, this requirement takes the form

$$\frac{\partial J(0, 0)}{\partial \alpha} = \int_{x_1}^{x_2} \left(\frac{\partial \bar{G}}{\partial y} g + \frac{\partial \bar{G}}{\partial y'} g' \right) dx \Bigg|_{(0,0)}$$

$$= \int_{x_1}^{x_2} \left(\frac{\partial G}{\partial y} - \frac{d}{dx} \frac{\partial G}{\partial y'} \right) g\, dx \neq 0.$$

Unless the coefficient of g in the last integrand vanishes identically when $y(x)$ is identified with the minimizing function,* $g(x)$ certainly can be so chosen that the last integral does not vanish. Let $g(x)$ be so chosen. Then for any subsequent choice of $f(x)$, α can be determined as a function of ϵ in such a way that (59) is satisfied.

The requirement that the function

$$I(\epsilon) \equiv \int_{x_1}^{x_2} F(x, y + \epsilon f + \alpha g, y' + \epsilon f' + \alpha g')\, dx$$

take on a minimum value when $\epsilon = 0$, and also $\alpha(\epsilon) = 0$, then takes the form

$$\frac{dI(0)}{d\epsilon} = \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} \right) f\, dx + \frac{d\alpha(0)}{d\epsilon} \int_{x_1}^{x_2} \left(\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} \right) g\, dx = 0. \quad (60)$$

If we recall that, when $J(\epsilon, \alpha) = K$, there follows

$$\frac{d\alpha}{d\epsilon} = -\left(\frac{\partial J}{\partial \epsilon} \right) \Big/ \left(\frac{\partial J}{\partial \alpha} \right),$$

* It can be shown that those cases in which this situation exists are of no interest.

we may obtain the result

$$\frac{d\alpha(0)}{d\epsilon} = -\left.\frac{\int_{x_1}^{x_2}\left(\frac{\partial \bar{G}}{\partial y} - \frac{d}{dx}\frac{\partial \bar{G}}{\partial y'}\right) f\,dx}{\int_{x_1}^{x_2}\left(\frac{\partial \bar{G}}{\partial y} - \frac{d}{dx}\frac{\partial \bar{G}}{\partial y'}\right) g\,dx}\right|_{(0,0)}$$

$$= -\frac{\int_{x_1}^{x_2}\left(\frac{\partial G}{\partial y} - \frac{d}{dx}\frac{\partial G}{\partial y'}\right) f\,dx}{\int_{x_1}^{x_2}\left(\frac{\partial G}{\partial y} - \frac{d}{dx}\frac{\partial G}{\partial y'}\right) g\,dx}. \tag{61}$$

The result of introducing this expression into (60) can be written in the form

$$\int_{x_1}^{x_2}\left[\left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'}\right) + \lambda\left(\frac{\partial G}{\partial y} - \frac{d}{dx}\frac{\partial G}{\partial y'}\right)\right] f\,dx = 0, \tag{62}$$

where λ is a constant, defined as the ratio of two definite integrals involving the arbitrarily *fixed* function $g(x)$:

$$\lambda = -\frac{\int_{x_1}^{x_2}\left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'}\right) g\,dx}{\int_{x_1}^{x_2}\left(\frac{\partial G}{\partial y} - \frac{d}{dx}\frac{\partial G}{\partial y'}\right) g\,dx}.$$

Since the function f *can* now be prescribed arbitrarily inside (x_1, x_2), its coefficient in (62) must vanish, giving the desired necessary condition

$$\left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right] + \lambda\left[\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right] = 0. \tag{63}$$

The result established may be summarized as follows: *In order to minimize (or maximize) an integral* $\int_a^b F\,dx$ *subject to a constraint* $\int_a^b G\,dx = K$, *first write* $H = F + \lambda G$, *and minimize (or maximize)* $\int_a^b H\,dx$ *subject to no constraint. Carry the Lagrange multiplier* λ *through the calculation, and determine it, together with the constants of integration arising in the solution of Euler's equation, so that the constraint* $\int_a^b G\,dx = K$ *is satisfied, and so that the end conditions are satisfied.*

The above statement applies in the more general case, in which two or more independent and dependent variables are involved.

To illustrate the procedure, we determine the curve of length L which passes through the points (0, 0) and (1, 0) and for which the area I between the curve and the x-axis is a maximum. We are thus to maximize the integral

$$I = \int_0^1 y\,dx, \tag{64a}$$

subject to the end conditions

$$y(0) = y(1) = 0 \tag{64b}$$

and to the constraint

$$\int_0^1 (1 + y'^2)^{1/2}\,dx = L, \tag{64c}$$

where L is a prescribed constant greater than unity.

The Euler equation corresponding to the maximization of the integral of $H = y + \lambda(1 + y'^2)^{1/2}$ is then of the form

$$\lambda \frac{d}{dx}\left[\frac{y'}{(1 + y'^2)^{1/2}}\right] - 1 = 0$$

or, after integration and simplification,

$$[\lambda^2 - (x - c_1)^2]y'^2 = (x - c_1)^2.$$

By solving for y' and integrating again, we find that the extremals are of the form

$$y = \pm[\lambda^2 - (x - c_1)^2]^{1/2} + c_2,$$

and hence (as might have been expected) are arcs of the circles

$$(x - c_1)^2 + (y - c_2)^2 = \lambda^2. \tag{65}$$

The three constants are to be determined so that the circle passes through the end points, and so that the relevant arc length is L.

We may notice that, *if* $L > \pi/2$, the "solution" obtained does not satisfy the requirement that y be a single-valued function of x.

2.7. Sturm-Liouville problems. To illustrate an application of a closely related procedure, we consider next the problem of determining stationary values of the quantity λ defined by the ratio

$$\lambda = \frac{\int_a^b (p\,y'^2 - q\,y^2)\,dx}{\int_a^b r\,y^2\,dx} \equiv \frac{I_1}{I_2} \tag{66}$$

where p, q, and r are given functions of the independent variable x. The variation of the ratio is of the form

$$\delta\lambda = \frac{I_2\,\delta I_1 - I_1\,\delta I_2}{I_2^2} = \frac{1}{I_2}(\delta I_1 - \lambda\,\delta I_2) = 0 \tag{67}$$

where, if δy vanishes at the end points,

$$\left.\begin{aligned} \delta I_1 &= -2\int_a^b [q\,y + (p\,y')']\,\delta y\,dx, \\ \delta I_2 &= 2\int_a^b r\,y\,\delta y\,dx \end{aligned}\right\}. \tag{68a,b}$$

If (68a,b) are introduced into (67), there follows

$$\delta\lambda = \frac{-2\int_a^b [(p\,y')' + q\,y + \lambda\,r\,y]\,\delta y\,dx}{\int_a^b r\,y^2\,dx}. \tag{69}$$

Thus, from the arbitrariness of δy, the condition $\delta\lambda = 0$ leads to the Euler equation in the form

$$\frac{d}{dx}\left(p\frac{dy}{dx}\right) + q\,y + \lambda\,r\,y = 0. \tag{70}$$

Suppose that the boundary conditions are of the form

$$y(a) = 0, \qquad y(b) = 0. \tag{71}$$

Then the problem is one of *characteristic values*, and is a particular case of the general *Sturm-Liouville problem*. (See Section 1.29.) It follows that the problem of determining characteristic functions of (70), subject to (71), is equivalent to the problem of determining functions satisfying (71) which render (66) stationary.

Stationary values of λ must then be characteristic numbers of the problem. To verify this fact directly, suppose that λ_k and $\phi_k(x)$ are corresponding characteristic quantities, so that

$$(p\,\phi_k')' + q\,\phi_k + \lambda_k r\,\phi_k = 0. \tag{72}$$

Then if y is replaced by ϕ_k in (66), λ should reduce to λ_k. Before making the substitution, we transform I_1 by integrating the first

term by parts, using (71), to rewrite (66) in the form

$$\lambda = -\frac{\int_a^b [(p\,y')' + q\,y]y\,dx}{\int_a^b r\,y^2\,dx}. \tag{66'}$$

Now, by replacing y by ϕ_k, there follows

$$\lambda = \frac{\int_a^b [-(p\,\phi_k')' - q\,\phi_k]\phi_k\,dx}{\int_a^b r\,\phi_k{}^2\,dx} = \frac{\int_a^b [\lambda_k r\,\phi_k]\phi_k\,dx}{\int_a^b r\,\phi_k{}^2\,dx} = \lambda_k, \tag{73}$$

as was to be shown. The equality of the two square brackets in (73) follows from (72).

If we artificially impose the condition ("constraint")

$$\int_a^b r\,y^2\,dx = 1, \tag{74}$$

it follows from (66) that the minimal condition takes the form

$$\delta\lambda \equiv \delta \int_a^b (p\,y'^2 - q\,y^2)\,dx = 0, \tag{75}$$

where y is to satisfy (74) and the prescribed end conditions. From (67) it follows also that the condition

$$\delta(I_1 - \lambda\,I_2) = 0, \tag{76}$$

with the provision $y \not\equiv 0$, is equivalent to either (67) or the combination of (75) and (74). In this last form, the constant λ plays the role of a Lagrange multiplier, and is to be determined together with the function y so that $I_1 - \lambda\,I_2$ is stationary and $y(x) \not\equiv 0$. The condition (74) is recognized as a *normalizing* condition, the weighting function $r(x)$ being that function with respect to which the distinct characteristic functions of the problem are orthogonal. (See Section 1.29.)

If the constraint (74) were suppressed, the problem (75) would in general determine only one extremal ($y \equiv 0$). However, when the condition (74) is added, the problem has in general an infinite set of extremals, for each of which λ is stationary for small variations in y.

The requirement that (71) be satisfied was needed only in establishing (68a) and (66'). It is readily verified that these equations are still valid if, instead of requiring that y vanish at an end point,

we impose the *natural boundary condition* that $p\,dy/dx$ vanish at that point.

For a physical interpretation of these results, we recall that, for free vibration of an elastic string of length L under tension $F(x)$, the amplitude y satisfies the equation

$$\frac{d}{dx}\left(F\frac{dy}{dx}\right) + \omega^2\rho\, y = 0 \tag{77}$$

where $\rho(x)$ is the linear mass density and ω the circular frequency. This equation is identified with (70) if we set

$$p = F, \quad q = 0, \quad r = \rho, \quad \lambda = \omega^2. \tag{78}$$

Hence the vibration modes are extremals of the problem

$$\delta\omega^2 = \delta\left[\frac{\int_0^L F\,y'^2\,dx}{\int_0^L \rho\,y^2\,dx}\right] = 0, \tag{79}$$

and stationary values of the ratio are squares of the natural circular frequencies. Alternatively, from (76), the variational problem can be taken in the form

$$\delta\int_0^L (F\,y'^2 - \omega^2\rho\,y^2)\,dx = 0. \tag{80}$$

The statement of (79) is a special case of *Rayleigh's principle* which applies to more general elastic systems.* *It can be shown that the smallest stationary value of* ω^2 *is truly the minimum value of the ratio in* (79), for all continuously differentiable functions $y(x)$ which vanish at $x = 0$ and $x = L$ (see Problem 37).

Equation (80) is closely connected with *Hamilton's principle*, which is treated in the following section.

Methods for obtaining *approximations* to the extremals of such problems are to be considered in Sections 2.17 and 2.18.

2.8. Hamilton's principle. One of the most basic and important principles of mathematical physics bears the name of Hamilton.† From it can be deduced the fundamental equations governing a large number of physical phenomena. It is formulated here

* See Reference 3.

† Sir William Rowan Hamilton (1805–1865), an Irish mathematician, is also known for his invention of *quaternions*.

in terms of the dynamics of a system of particles, and is readily extended by analogy to other considerations.

We consider first a single particle of mass m, moving subject to a force field. If the vector from a fixed origin to the particle at time t is denoted by $\mathbf{r}$, then, according to Newton's laws of motion, the actual path followed is governed by the vector equation

$$m\frac{d^2\mathbf{r}}{dt^2} - \mathbf{F} = \mathbf{0}, \tag{81}$$

where $\mathbf{F}$ is the force acting on the particle. Now consider any *other* path $\mathbf{r} + \delta\mathbf{r}$. We require only that the *true* path and the *varied* path coincide at two distinct instants $t = t_1$ and $t = t_2$, that is, that the *variation* $\delta\mathbf{r}$ vanish at those two instants:

$$\delta\mathbf{r}\Big|_{t_1} = \delta\mathbf{r}\Big|_{t_2} = \mathbf{0}. \tag{82}$$

At any intermediate time t we then have to consider the true path $\mathbf{r}$ and the varied path $\mathbf{r} + \delta\mathbf{r}$.

The first step in the derivation consists in taking the scalar (dot) product of the variation $\delta\mathbf{r}$ into (81), and in integrating the result with respect to time over (t_1, t_2), to obtain the relation

$$\int_{t_1}^{t_2}\left(m\frac{d^2\mathbf{r}}{dt^2}\cdot\delta\mathbf{r} - \mathbf{F}\cdot\delta\mathbf{r}\right)dt = 0. \tag{83}$$

If the first term is integrated by parts, it takes the form

$$m\int_{t_1}^{t_2}\frac{d^2\mathbf{r}}{dt^2}\cdot\delta\mathbf{r}\,dt = m\left\{\left[\frac{d\mathbf{r}}{dt}\cdot\delta\mathbf{r}\right]_{t_1}^{t_2} - \int_{t_1}^{t_2}\frac{d\mathbf{r}}{dt}\cdot\delta\frac{d\mathbf{r}}{dt}\,dt\right\}.$$

Since the variation $\delta\mathbf{r}$ vanishes at the ends of the interval, the integrated terms vanish. Also, we have the relation

$$\frac{d\mathbf{r}}{dt}\cdot\delta\frac{d\mathbf{r}}{dt} = \frac{1}{2}\delta\left(\frac{d\mathbf{r}}{dt}\right)^2.$$

Hence, the first term in (83) is equivalent to

$$-\delta\left[\frac{1}{2}m\left(\frac{d\mathbf{r}}{dt}\right)^2\right] = -\delta T, \tag{84}$$

where T is the *kinetic energy* ($\frac{1}{2}mv^2$) of the particle, and (83) becomes

$$\int_{t_1}^{t_2}(\delta T + \mathbf{F}\cdot\delta\mathbf{r})\,dt = 0. \tag{85}$$

This is *Hamilton's principle* in its most general form, as applied to the motion of a single particle. However, if the force field is *conservative* it can be put in a more concise form.

To fix ideas, suppose that $\mathbf{F}$ is specified by its components X, Y, Z in the directions of the rectangular xyz-coordinates. We recall that a force field $\mathbf{F}$ is conservative if and only if

$$\mathbf{F} \cdot d\mathbf{r} \equiv X\,dx + Y\,dy + Z\,dz$$

is the differential $d\Phi$ of a single-valued function Φ. The force $\mathbf{F}$ is then the gradient of Φ. The function Φ is called the *force potential* and its *negative*, say V, is called the *potential energy*. Clearly, Φ and V each involve an irrelevant arbitrary additive constant.

It follows that $\mathbf{F}$ is conservative if there exists a single-valued function Φ such that

$$\mathbf{F} \cdot \delta\mathbf{r} = \delta\Phi. \tag{86}$$

In terms of the xyz-components of $\mathbf{F}$, this means that

$$X\,\delta x + Y\,\delta y + Z\,\delta z = \delta\Phi \tag{87}$$

where

$$X = \frac{\partial\Phi}{\partial x}, \quad Y = \frac{\partial\Phi}{\partial y}, \quad Z = \frac{\partial\Phi}{\partial z}. \tag{88}$$

Thus if Φ is the potential function, equation (85) becomes

$$\delta \int_{t_1}^{t_2} (T + \Phi)\,dt = 0. \tag{89}$$

In place of the potential function Φ, it is more customary to use the *potential energy* function V,

$$V = -\Phi, \tag{90}$$

so that Hamilton's principle takes the form

$$\delta \int_{t_1}^{t_2} (T - V)\,dt = 0 \tag{91}$$

when a potential function exists, that is, when the forces acting are conservative.

For such a problem Hamilton's principle states that the motion is such that the integral of the difference between the kinetic and potential energies is stationary for the true path. It can be shown further that actually this integral is a *minimum* when compared

with that corresponding to any neighboring path having the same terminal configurations. Thus we may say that "nature tends to equalize the kinetic and potential energies over the motion."

The energy difference

$$L = T - V$$

is sometimes called the *kinetic potential* or the *Lagrangian function*. In terms of this function, (91) becomes merely

$$\delta \int_{t_1}^{t_2} L\, dt = 0. \tag{92}$$

If nonconservative forces are present, the potential energy function generally does not exist, and recourse must be had to (85). We may notice, however, that in any case $\mathbf{F} \cdot \delta\mathbf{r}$ is the element of *work* done *by* the force $\mathbf{F}$ in a small *displacement* $\delta\mathbf{r}$. In particular, when the force *is* conservative this element of work is equivalent to $\delta\Phi = -\delta V$.

The above derivation is extended to a *system* of N particles by summation, and to a *continuous* system by integration. Thus if the kth particle is of mass m_k, is specified by the vector $\mathbf{r}_k$, and is subject to a force $\mathbf{F}_k$, the total kinetic energy is given by

$$T = \sum_{k=1}^{N} \frac{1}{2} m_k \left(\frac{d\mathbf{r}_k}{dt}\right)^2 \equiv \sum_{k=1}^{N} \frac{1}{2} m_k v_k^2 \tag{93a}$$

while the total work done by the forces acting is given by

$$\sum_{k=1}^{N} \mathbf{F}_k \cdot \delta\mathbf{r}_k. \tag{93b}$$

Finally, the principle applies equally well to a general dynamical system consisting of particles and rigid bodies subject to interconnections and constraints. We notice that the derivation is independent of the coordinates specifying the system.

2.9. Lagrange's equations. In a dynamical system with n degrees of freedom it is usually possible to choose n independent geometrical quantities which uniquely specify the position of all components of the system. These quantities are known as *generalized coordinates*. For example, in the case of a pendulum consisting

of a point mass m suspended by an inextensible string of length L, the position of the mass is completely determined by the angle θ between the deflected and equilibrium positions of the string (Figure 2.3). If xy-coordinates were used, x and y would not be independent since the constraint equation $x^2 + y^2 = L^2$ would have to be imposed. Similarly, the compound pendulum of Figure 2.4 has two degrees of freedom and the indicated angles θ_1 and θ_2 are suitable generalized coordinates. In rectangular coordinates, if the quantities x_1, y_1 and x_2, y_2 representing the positions of m_1 and m_2 were used, two equations of constraint would be needed.

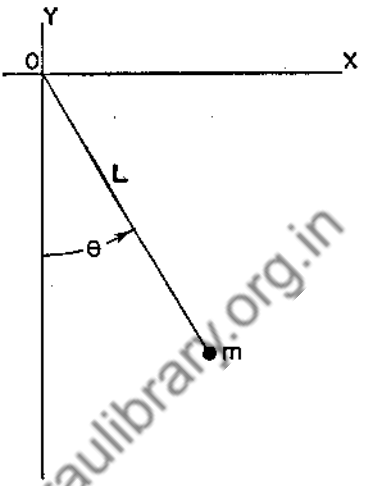


FIGURE 2.3

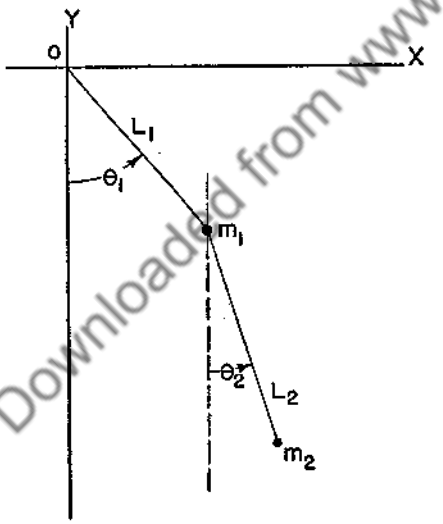


FIGURE 2.4

In the general case, the total *kinetic energy* T may depend upon the generalized coordinates, say q_1, q_2, . . . , q_n, as well as upon their time rates of change or so-called *generalized velocities* $\dot{q}_1$, $\dot{q}_2$, . . . , $\dot{q}_n$.* For a *conservative* system the total *potential energy* V is a function only of position and hence does *not* depend upon the generalized *velocities*.

Also, the work done *by* the force system involved when the q's are given small displacements is

$$-\delta V = +\delta\Phi = Q_1\,\delta q_1 + Q_2\,\delta q_2 + \cdots + Q_n\,\delta q_n, \qquad (94)$$

* Here and elsewhere, a dot indicates *time* differentiation.

where

$$Q_1 = -\frac{\partial V}{\partial q_1} = \frac{\partial \Phi}{\partial \dot{q}_1}, \quad \cdots, \quad Q_n = -\frac{\partial V}{\partial q_n} = \frac{\partial \Phi}{\partial \dot{q}_n}. \tag{95}$$

The quantity $Q_i\,\delta q_i$ is the work done *by* the forces in a displacement δq_i. Since the Q's may or may not have the dimension of true force they are called *generalized forces*. Thus if q_i is a linear displacement, then Q_i is truly a force, while if q_i is an *angular* displacement, then Q_i is a *torque*. In other applications the q's may represent electric charges, currents, areas, volumes, and so forth, and the nature of the Q's is determined accordingly, in such a way that $Q\,\delta q$ has the dimensions of *work*.

In applications of Hamilton's principle,

$$\delta \int_{t_1}^{t_2} (T - V)\,dt = 0, \tag{96}$$

to *conservative* systems we may thus suppose that the total kinetic energy T is expressed in terms of the n q's and the n $\dot{q}$'s, while the total potential energy V is expressed in terms of the q's only. The associated Euler equations then become

$$\frac{d}{dt}\left[\frac{\partial(T - V)}{\partial \dot{q}_i}\right] - \frac{\partial(T - V)}{\partial q_i} = 0 \qquad (i = 1, 2, \cdots, n). \tag{97a}$$

Since $\partial V/\partial \dot{q}_i \equiv 0$, we may write each equation alternatively in the form

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i} = 0 \tag{97b}$$

or, using (95), in the form

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_i} - \frac{\partial T}{\partial q_i} = Q_i. \tag{97c}$$

The three forms (97a,b,c) are completely equivalent for a conservative system, and are usually called *Lagrange's equations*. One equation is obtained for each independent q.

In illustration, for the simple pendulum of Figure 2.3 the kinetic energy is given by

$$T = \tfrac{1}{2}m(L\,\dot{\theta})^2. \tag{98a}$$

If damping is neglected, the work done *by* gravity in lifting the mass from its equilibrium position to the position θ is *negative* and

is given by $-m\,g\,L(1 - \cos\theta)$, so that the potential energy is of the form

$$V = +m\,g\,L(1 - \cos\theta) + \text{constant}. \tag{98b}$$

Thus, with $q_1 = \theta$, equation (97b) becomes

$$\frac{d}{dt}(m\,L^2\dot{\theta}) - 0 + m\,g\,L\sin\theta = 0$$

or

$$\ddot{\theta} + \frac{g}{L}\sin\theta = 0. \tag{99}$$

This is the well-known equation of motion for such a pendulum.

For the compound pendulum of Figure 2.4 we may proceed as follows. If the rectangular coordinates of m_1 and m_2 are taken as (x_1, y_1) and (x_2, y_2) there follows

$$x_1 = L_1\sin\theta_1, \quad y_1 = -L_1\cos\theta_1;$$

$$x_2 = L_1\sin\theta_1 + L_2\sin\theta_2, \quad y_2 = -L_1\cos\theta_1 - L_2\cos\theta_2.$$

The total kinetic energy T is then

$$T = \tfrac{1}{2}m_1(\dot{x}_1^2 + \dot{y}_1^2) + \tfrac{1}{2}m_2(\dot{x}_2^2 + \dot{y}_2^2).$$

Hence there follows, by substitution and simplification,

$$T = \tfrac{1}{2}(m_1 + m_2)L_1^2\dot{\theta}_1^2 + m_2L_1L_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1 - \theta_2) + \tfrac{1}{2}m_2L_2^2\dot{\theta}_2^2. \tag{100a}$$

The total potential energy is given by

$$V = m_1g\,y_1 + m_2g\,y_2 + \text{constant}$$

or

$$V = -(m_1 + m_2)g\,L_1\cos\theta_1 - m_2g\,L_2\cos\theta_2 + \text{constant}. \tag{100b}$$

Use of equation (97b) then leads to the two equations of motion

$$(m_1 + m_2)L_1\ddot{\theta}_1 + m_2L_2(\ddot{\theta}_2\cos\alpha + \dot{\theta}_2^2\sin\alpha) + (m_1 + m_2)g\sin\theta_1 = 0 \tag{101a}$$

and

$$L_1\ddot{\theta}_1\cos\alpha + L_2\ddot{\theta}_2 - L_1\dot{\theta}_1^2\sin\alpha + g\sin\theta_2 = 0, \tag{101b}$$

where

$$\alpha = \theta_1 - \theta_2. \tag{102}$$

We notice that it is not necessary to evaluate the constraints exerted by tensions in the strings supporting the masses, since they do no *work*.

For a *nonconservative* force field, Lagrange's equations must be based on the form (85), rather than (91). In this case it is still possible to express the work done by the force system in small displacements $\delta q_1, \ldots, \delta q_n$, in the form

$$\sum_{k=1}^{N} \mathbf{F}_k \cdot \delta \mathbf{r}_k = Q_1\,\delta q_1 + Q_2\,\delta q_2 + \cdots + Q_n\,\delta q_n, \tag{103}$$

as in (94). However, *the generalized forces Q_i are now in general not derivable from a potential function* as in (95). (In certain nonconservative systems such a function *may* exist, depending upon *time* as well as position.) To determine the Q's by *physical* considerations, we need only notice that, as before, $Q_i\,\delta q_i$ is the work done *by* the force system when q_i is changed to $q_i + \delta q_i$ *and the other q's are held fixed.*

For an *analytical* determination, we may suppose that the components X_k, Y_k, and Z_k of the force $\mathbf{F}_k$ acting on the kth particle of the system are known in the directions of the x-, y-, and z-axes. Then (103) gives the relation

$$Q_1\,\delta q_1 + Q_2\,\delta q_2 + \cdots + Q_n\,\delta q_n = \sum_{k=1}^{N} (X_k\,\delta x_k + Y_k\,\delta y_k + Z_k\,\delta z_k). \tag{104}$$

Since x_k, y_k, and z_k are functions of the coordinates $q_1, q_2, \ldots, q_n$, there follows also

$$\left.\begin{aligned} \delta x_k &= \frac{\partial x_k}{\partial q_1}\,\delta q_1 + \cdots + \frac{\partial x_k}{\partial q_n}\,\delta q_n, \\ \delta y_k &= \frac{\partial y_k}{\partial q_1}\,\delta q_1 + \cdots + \frac{\partial y_k}{\partial q_n}\,\delta q_n, \\ \delta z_k &= \frac{\partial z_k}{\partial q_1}\,\delta q_1 + \cdots + \frac{\partial z_k}{\partial q_n}\,\delta q_n \end{aligned}\right\}. \tag{105}$$

Equation (104) must hold for arbitrary choices of the δq's. In particular, if we require that *all* the δq's *except* δq_i vanish, (105) then gives

$$\delta x_k = \frac{\partial x_k}{\partial q_i}\,\delta q_i, \quad \delta y_k = \frac{\partial y_k}{\partial q_i}\,\delta q_i, \quad \delta z_k = \frac{\partial z_k}{\partial q_i}\,\delta q_i \tag{106a}$$

and (104) becomes

$$Q_i\,\delta q_i = \sum_{k=1}^{N} (X_k\,\delta x_k + Y_k\,\delta y_k + Z_k\,\delta z_k), \tag{106b}$$

in this case. By introducing (106a) into (106b), we then obtain the desired relation

$$Q_i = \sum_{k=1}^{N} \left(X_k \frac{\partial x_k}{\partial q_i} + Y_k \frac{\partial y_k}{\partial q_i} + Z_k \frac{\partial z_k}{\partial q_i} \right). \tag{107}$$

This result is clearly valid whether or not the system is conservative.

Hamilton's principle (85) then states that

$$\delta \int_{t_1}^{t_2} T\,dt + \int_{t_1}^{t_2} (Q_1\,\delta q_1 + Q_2\,\delta q_2 + \cdots + Q_n\,\delta q_n)\,dt = 0.$$

By calculating the variation of the first integral in the usual way, we obtain the condition

$$\int_{t_1}^{t_2} \left\{ \left[\frac{\partial T}{\partial q_1} - \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_1}\right) + Q_1 \right] \delta q_1 + \cdots + \left[\frac{\partial T}{\partial q_n} - \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_n}\right) + Q_n \right] \delta q_n \right\} dt = 0. \tag{108}$$

The vanishing of the coefficients of the independent variations leads again to the equations (97c).

Thus *these equations are valid whenever the variations of the n q's are independent:*

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} = Q_i \qquad (i = 1, 2, \cdots, n). \tag{109}$$

For a *conservative* system the Q's are derivable from a potential function and (97a) or (97b) can be used alternatively.

2.10. Generalized dynamical entities. Before considering the definition of additional generalized dynamical entities, it is desirable to emphasize the fact that the so-called *generalized velocity* $\dot{q}_i$, associated with a generalized coordinate q_i, is merely the *time rate of change of that coordinate.* Thus, for example, in polar coordi-

nates (r, θ) the generalized velocities associated with r and θ are merely $\dot{r}$ and $\dot{\theta}$, respectively. We notice that $\dot{\theta}$ is *not* the component of the velocity vector in the circumferential direction $(r\theta)$. Similarly, the so-called *generalized accelerations* $\ddot{r}$ and $\ddot{\theta}$, associated with r and θ, are *not* the respective components of the acceleration vector in the radial and circumferential directions. It will be recalled that these latter quantities are of the forms $\ddot{r} - r\dot{\theta}^2$ and $r\ddot{\theta} + 2\dot{r}\dot{\theta}$.

In the remainder of this section we deal *always* with *generalized* forces, velocities, accelerations, and momenta. For brevity, the adjective "generalized" will frequently be omitted.

In rectangular coordinates (x, y, z) the quantities

$$p_x = m\,\dot{x}, \quad p_y = m\,\dot{y}, \quad p_z = m\,\dot{z}$$

are called the components of *momentum*. Since we have the relation

$$T = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2),$$

it follows that

$$\frac{\partial T}{\partial \dot{x}} = p_x, \quad \frac{\partial T}{\partial \dot{y}} = p_y, \quad \frac{\partial T}{\partial \dot{z}} = p_z.$$

In generalized coordinates, we call the quantity $\partial T/\partial \dot{q}_i$ the *generalized momentum* associated with q_i, and write

$$p_i = \frac{\partial T}{\partial \dot{q}_i}. \tag{110}$$

The associated equation of motion, (109), then becomes

$$\frac{dp_i}{dt} = Q_i + \frac{\partial T}{\partial q_i}. \tag{111}$$

Hence, *the rate of change of the* i*th generalized momentum is equal to the sum of the* i*th generalized force* Q_i *and the quantity* $\partial T/\partial q_i$. In *rectangular* coordinates the "corrective terms" $\partial T/\partial q_i$ are absent.

In motion specified by plane polar coordinates (r, θ) we have

$$T \equiv \frac{1}{2}\,m\left(\frac{ds}{dt}\right)^2 = \frac{1}{2}\,m(\dot{r}^2 + r^2\dot{\theta}^2)$$

and hence

$$p_r = m\,\dot{r}, \qquad p_\theta = m\,r^2\dot{\theta}; \qquad \frac{\partial T}{\partial r} = m\,r\,\dot{\theta}^2, \qquad \frac{\partial T}{\partial \theta} = 0. \tag{112}$$

The equations of motion (111) are then of the form

$$\frac{dp_r}{dt} = Q_r + m\,r\,\dot{\theta}^2, \tag{113a}$$

$$\frac{dp_\theta}{dt} = Q_\theta. \tag{113b}$$

Here Q_r is the impressed radial *force,* while the generalized θ-force Q_θ is a *torque.* If a particle moves in such a way that the (generalized) momentum associated with r is constant and hence, from (112), dr/dt is constant, (113a) shows that a net force $Q_r = -m\,r\,\dot{\theta}^2$ must then be exerted externally (e.g., by a spring) in the r-direction. More generally, in so far as change in the r-momentum is involved, the mass behaves as though a force $+m\,r\,\dot{\theta}^2 \equiv \partial T'/\partial r$ were acting in the r-direction in addition to the *actual* external force Q_r. The fictitious force is recognized as the so-called *centrifugal force.*

Since such quantities are not true physical forces, they are often called *inertia forces.* Their presence or absence depends, not upon the particular problem at hand, but *upon the coordinate system chosen.*

In general, we see that if T involves the coordinate q_i explicitly, the quantity $\partial T/\partial q_i$ can be considered as an associated inertia force. Thus, if we denote this quantity by P_i,

$$P_i = \frac{\partial T}{\partial q_i}, \tag{114}$$

the ith equation of motion (111) becomes

$$\frac{dp_i}{dt} = Q_i + P_i. \tag{115}$$

We shall refer to the quantity P_i as a *momental* inertia force.*

We may notice next that while the quantity dp_i/dt will in general contain terms involving the *generalized acceleration* $\ddot{q}_i$, its expansion may also involve nonaccelerational terms. Thus, if *accelerations* associated with generalized coordinates are to be of prime interest (as is usually the case), these latter terms may be conveniently transferred to the right in (115) and considered as addi-

* Various terminologies, some of which are at variance with this one, are in use.

tional (generalized) inertia forces. Such inertia forces are often said to be of the *Coriolis* type.

Thus a Coriolis inertia force is equivalent to an impressed force associated with q_i which tends to change the generalized velocity $\dot{q}_i$, but which does *not* tend to change the generalized momentum, when *actual* external forces are omitted. On the other hand, an inertia force of the "momental" type (e.g., a centrifugal force) is equivalent to an impressed force which tends to change both p_i and $\dot{q}_i$ in the absence of true external forces.

Since, from (112), we have $dp_r/dt \equiv m\,\ddot{r}$, no such terms are present in (113a) and we have

$$m\,\ddot{r} = Q_r + m\,r\,\dot{\theta}^2. \tag{116a}$$

However, since $dp_\theta/dt \equiv m\,r^2\ddot{\theta} + 2m\,r\,\dot{r}\,\dot{\theta}$, the second term can be conveniently transferred to the right in (113b) to give

$$m\,r^2\ddot{\theta} = Q_\theta - 2m\,r\,\dot{r}\,\dot{\theta}. \tag{116b}$$

The generalized "Coriolis force" in (116b) is clearly a *torque*.

We notice that the momental (centrifugal) term $m\,r\,\dot{\theta}^2$ is an inertia force with regard to change in both r-momentum and r-velocity, while the Coriolis term $-2m\,r\,\dot{r}\,\dot{\theta}$ is an inertia "force" (torque) only with regard to change in θ-velocity. That is, one may say that "a θ-velocity tends to change the r-velocity and the r-momentum, whereas simultaneous r- and θ-velocities tend to change the θ-velocity, in the absence of actual impressed forces."

As a further example, we consider motion specified by spherical coordinates $(q_1, q_2, q_3) \equiv (r, \theta, \phi)$ in space, where r is distance from the origin, θ is polar angle, and ϕ is "cone angle" (Figure 2.5), so that

$$x = r\cos\theta\sin\phi,$$

$$y = r\sin\theta\sin\phi,$$

$$z = r\cos\phi.$$

Since the element of arc length is $ds^2 = dr^2 + r^2\sin^2\phi\,d\theta^2 + r^2\,d\phi^2$, there follows from $T = \dfrac{1}{2}\,m\left(\dfrac{ds}{dt}\right)^2$ the result

$$T = \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2\sin^2\phi + r^2\dot{\phi}^2).$$

Thus, with the notation of (110) and (114), we have

$$p_r = m\,\dot{r}, \quad p_\theta = m\,r^2\dot{\theta}\,\sin^2\phi, \quad p_\phi = m\,r^2\dot{\phi} \tag{117a}$$

and

$$P_r = m\,r\,\dot{\theta}^2\sin^2\phi + m\,r\,\dot{\phi}^2, \quad P_\theta = 0, \quad P_\phi = m\,r^2\dot{\theta}^2\sin\phi\cos\phi. \tag{117b}$$

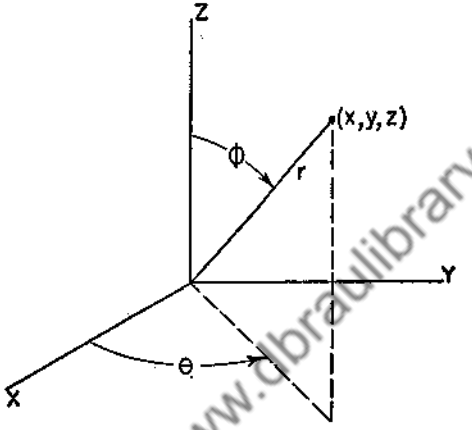


FIGURE 2.5

The equations of motion in the form (115) then become

$$\left.\begin{aligned} \frac{d}{dt}(m\,\dot{r}) &= Q_r + m\,r\,\dot{\theta}^2\sin^2\phi + m\,r\,\dot{\phi}^2, \\ \frac{d}{dt}(m\,r^2\dot{\theta}\sin^2\phi) &= Q_\theta, \\ \frac{d}{dt}(m\,r^2\dot{\phi}) &= Q_\phi + m\,r^2\dot{\theta}^2\sin\phi\cos\phi \end{aligned}\right\}. \tag{118a,b,c}$$

In particular, if a particle is constrained to move *on the surface of a sphere of radius a*, there follows $r = a$, $\dot{r} = \ddot{r} = 0$, and hence, from (118a), the necessary *physical constraint* normal to the sphere surface is given by

$$Q_r = -m\,a\,\dot{\theta}^2\sin^2\phi - m\,a\,\dot{\phi}^2. \tag{119}$$

That is, the total centrifugal inertia force $(P_r)_{r=a}$ must be balanced by a physical constraint equal to the negative of that quantity. Equations (118b,c) then give the equations of motion

$$m a^2 \frac{d}{dt}(\dot\theta \sin^2 \phi) = Q_\theta$$

or
$$m a^2 \sin^2 \phi\, \ddot\theta = Q_\theta - 2m a^2 \dot\theta \dot\phi \sin \phi \cos \phi \tag{120a}$$

and
$$m a^2 \ddot\phi = Q_\phi + m a^2 \dot\theta^2 \sin \phi \cos \phi. \tag{120b}$$

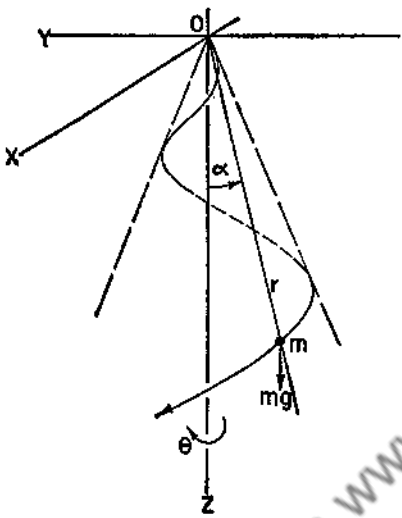


FIGURE 2.6

The inertia force (torque) associated with θ is of Coriolis type, while that associated with ϕ is of momental type.

Similarly, if the particle is constrained to move *on the surface of the cone* $\phi = \alpha$ the constraint is given by (118c) in the form

$$Q_\phi = -m r^2 \dot\theta^2 \sin \alpha \cos \alpha, \tag{121}$$

and the equations of motion are

$$m \ddot r = Q_r + m r \dot\theta^2 \sin^2 \alpha \tag{122a}$$

and
$$m \sin^2 \alpha \frac{d}{dt}(r^2 \dot\theta) = Q_\theta,$$

or
$$m r^2 \sin^2 \alpha\, \ddot\theta = Q_\theta - 2m r \dot r \dot\theta \sin^2 \alpha. \tag{122b}$$

The inertia r-force is momental, whereas the inertia θ-torque is of Coriolis type.

In illustration, suppose that a bead of mass m is sliding without friction under gravity along a wire, inclined at an angle α to the vertical and rotating with constant angular velocity ω (Figure 2.6). Then m must move on the surface of the cone $\phi = \alpha$, in such a way that

$$\theta = \omega t, \tag{123}$$

if we measure t from a time when $\theta = 0$. Also we have

$$Q_r = m g \cos \alpha.$$

Hence, with $\dot{\theta} = \omega$, equation (122a) is the equation of motion determining the coordinate r which, together with (123) and the relation $\phi = \alpha$, specifies the position of m. This equation takes the form

$$m\ddot{r} - m r \omega^2 \sin^2 \alpha = m g \cos \alpha, \tag{124}$$

from which there follows

$$r = c_1 \sinh (\omega t \sin \alpha) + c_2 \cosh (\omega t \sin \alpha) - \frac{g \cos \alpha}{\omega^2 \sin^2 \alpha}.$$

If the bead is released from rest at the origin at the instant $t = 0$, the evaluation of the constants gives

$$r = \frac{g \cos \alpha}{\omega^2 \sin^2 \alpha} [\cosh (\omega t \sin \alpha) - 1], \tag{125}$$

With this expression for r, the generalized forces (torques) associated with ϕ and θ are then found from (121) and (122b), in the form

$$Q_\phi = -m r^2 \omega^2 \sin \alpha \cos \alpha, \quad Q_\theta = 2m \omega r \dot{r} \sin^2 \alpha. \tag{126a,b}$$

These results may be interpreted as follows: The forces acting on m in the positive ϕ-direction are the gravity component $-m g \sin \alpha$ and the reaction component R_ϕ exerted by the wire. In a linear displacement $\delta s_\phi = r\, \delta\phi$, in which r and θ are considered to be held fixed, the work done by these forces would be $(R_\phi - m g \sin \alpha) r\, \delta\phi$. By equating this work to $Q_\phi\, \delta\phi$, we obtain the reaction R_ϕ in the form

$$R_\phi = m g \sin \alpha - m r \omega^2 \sin \alpha \cos \alpha.$$

Further, in a linear displacement $\delta s_\theta = r \sin \alpha\, \delta\theta$, in which r and ϕ are considered to be held fixed, the work done by the reaction component R_θ would be $R_\theta\, r \sin \alpha\, \delta\theta$. Since the force of gravity has no component in the θ-direction, there follows $R_\theta r \sin \alpha\, \delta\theta = Q_\theta\, \delta\theta$. Hence, the circumferential force exerted by the wire is given by

$$R_\theta = 2m \omega \dot{r} \sin \alpha.$$

In this example ϕ is fixed and, if we think of θ as prescribed, the system essentially has only one degree of freedom. *The reactions R_ϕ and R_θ were obtained by considering the system as a degenerate case of a system with three degrees of freedom.*

If the reactions were *not* required, we would write directly

$$x = r \cos \omega t \sin \alpha, \quad y = r \sin \omega t \sin \alpha, \quad z = r \cos \alpha,$$

and so obtain T immediately in terms of the one independent coordinate r, in the form

$$T = \tfrac{1}{2} m(\dot{r}^2 + r^2 \omega^2 \sin^2 \alpha).$$

Lagrange's equation for that coordinate would then be obtained as

$$\frac{d}{dt}(m\dot{r}) - m r \omega^2 \sin^2 \alpha = Q_r \equiv m g \cos \alpha,$$

in accordance with (124).

2.11. Constraints in dynamical systems. In some cases it is inconvenient or impossible to choose n independent generalized coordinates to specify the configuration of a system having n degrees of freedom. Instead, if n coordinates are used, we may have k auxiliary equations relating these n coordinates. The system is then usually said to have $n - k$ degrees of freedom, and the k restrictive equations are known as the *equations of constraint*. If these equations are of the form

$$\left.\begin{aligned} &\phi_1(q_1, q_2, \cdots, q_n) = 0, \\ &\phi_2(q_1, q_2, \cdots, q_n) = 0, \\ &\cdots\cdots\cdots\cdots \\ &\phi_k(q_1, q_2, \cdots, q_n) = 0 \end{aligned}\right\}, \tag{127}$$

then two possible procedures are available.

Clearly, if the k equations *can* be conveniently resolved to express k of the q's in terms of the $n - k$ remaining q's, then these latter q's are independent and the governing differential equations are determined by the methods of Section 2.9.

If this procedure is not convenient, the method of *Lagrange multipliers*, as described in Section 2.6, is useful. The method consists in first forming the variational conditions which follow from (127):

$$\left.\begin{aligned} &\frac{\partial \phi_1}{\partial q_1}\delta q_1 + \cdots + \frac{\partial \phi_1}{\partial q_n}\delta q_n = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &\frac{\partial \phi_k}{\partial q_1}\delta q_1 + \cdots + \frac{\partial \phi_k}{\partial q_n}\delta q_n = 0 \end{aligned}\right\}. \tag{128}$$

These equations are then multiplied respectively by functions $\lambda_1, \ldots, \lambda_k$, integrated over (t_1, t_2), and added to the equation of Hamilton's principle. Then, as was indicated in Section 2.6, we obtain n equations, each of the form

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} = Q_i + \lambda_1 \frac{\partial \phi_1}{\partial q_i} + \cdots + \lambda_k \frac{\partial \phi_k}{\partial q_i} \qquad (i = 1, 2, \cdots, n). \tag{129}$$

These n equations, together with the k equations (127), then comprise $n + k$ equations in the $n + k$ unknown quantities $q_1, \ldots, q_n$ and $\lambda_1, \ldots, \lambda_k$. If the λ's are eliminated, the resultant n equations serve to determine the n q's.

In equation (129), Q_i is, as before, determined by the fact that $Q_i\,\delta q_i$ is the work done by the external forces when q_i is varied by δq_i and the remaining q's are held fixed. However, here it is important to notice that *such a variation may violate the physical constraint conditions,* which may require that a displacement δq_i should actually be necessarily accompanied by changes in certain of the other q's. It is useful to notice that, for a *conservative* system, (129) is obtained by replacing V by $V - \Sigma\lambda_r\phi_r$ in (97b). This last quantity is sometimes called the *reduced* potential energy.

From (129) it is apparent that a term $\lambda_k\,\partial\phi_k/\partial q_i$ is of the nature of a generalized force, due to the kth constraint and associated with the ith coordinate. Each constraint may thus contribute an additional generalized force to each of the equations of motion.

However, we notice that the work done in *any* set of displacements by the force due to the kth constraint is given by

$$\lambda_k \frac{\partial \phi_k}{\partial q_1}\,\delta q_1 + \lambda_k \frac{\partial \phi_k}{\partial q_2}\,\delta q_2 + \cdots + \lambda_k \frac{\partial \phi_k}{\partial q_n}\,\delta q_n.$$

Hence, in virtue of equations (128), *the work done by the (fixed) constraint vanishes if the displacements satisfy the constraint conditions.* Displacements which are compatible with the constraint conditions are often called *virtual displacements.*

In certain cases, a constraint condition may *not* be expressible in the form (127), but may be of the form

$$C_1\,\delta q_1 + \cdots + C_n\,\delta q_n = 0, \tag{130}$$

where the left-hand member is not proportional to the variation of *any* function. In such a case the constraint is said to be *nonholonomic.* If k nonholonomic constraints are involved, it is not *possible* to eliminate certain of the q's by solving equations similar to (127), so that n coordinates are still needed to specify the configuration. Nevertheless, the system is usually said to possess only $n - k$ degrees of freedom.

In any case, it is clear that the method of Lagrange multipliers (which involves only the *variations* of the coordinates) is again directly applicable in that the functions $\partial\phi_1/\partial q_1, \ldots, \partial\phi_1/\partial q_n$ are merely replaced by the functions $C_1, \ldots, C_n$. The general problem of the rolling of a disk on a plane is found to be one involving nonholonomic constraints, and is solvable by these methods.*

The basic ideas of this section may be illustrated by two elementary examples. The simple pulley of Figure 2.7 possesses one degree of freedom, and q_1 is a suitable coordinate. The kinetic energy of the system, neglecting the weight of the cord, is

$$T = \tfrac{1}{2}(m_1 + m_2)\dot{q}_1^2.$$

If q_1 is increased by δq_1, the work done by gravity is given by

$$m_1 g\, \delta q_1 - m_2 g\, \delta q_1$$

and hence

$$Q_1 = (m_1 - m_2)g.$$

Thus the equation of motion is

$$(m_1 + m_2)\ddot{q}_1 = (m_1 - m_2)g, \tag{131}$$

as is also obvious from other considerations.

Suppose, however, *to illustrate the preceding developments in a simple case,* that the *two* coordinates q_1 and q_2 of Figure 2.8 are used. These two coordinates are clearly not independent since, if the total length of the cord (assumed to be inextensible) is L, the constraint

$$q_1 + q_2 = L \tag{132}$$

must be imposed. In terms of q_1 and q_2, the kinetic energy of the system is

$$T = \tfrac{1}{2}m_1\dot{q}_1^2 + \tfrac{1}{2}m_2\dot{q}_2^2. \tag{133}$$

* See, for example, Reference 4.

If q_1 is increased by δq_1 *and* q_2 *is held fixed* (violating the constraint condition), the work done by gravity in the displacement δq_1 is $m_1 g\, \delta q_1$. Thus, we must have

$$Q_1 = m_1 g$$

and, similarly,

$$Q_2 = m_2 g.$$

From (132) we have also

$$\delta q_1 + \delta q_2 = 0. \tag{134}$$

This condition clearly requires that the displacements satisfy the constraint condition, that is, that they be *virtual* displacements, so that the work done by the constraining tension vanishes.

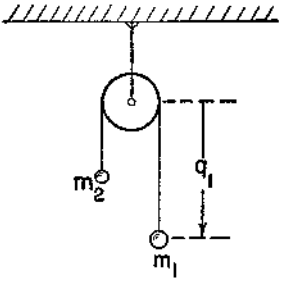


FIGURE 2.7

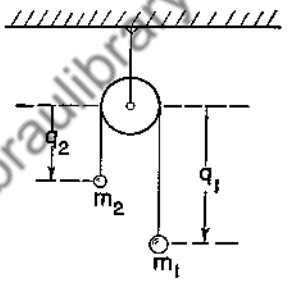


FIGURE 2.8

With the introduction of a Lagrange multiplier, the equations corresponding to (129) become

$$\left.\begin{aligned} m_1 \ddot{q}_1 &= m_1 g + \lambda, \\ m_2 \ddot{q}_2 &= m_2 g + \lambda \end{aligned}\right\}. \tag{135a,b}$$

Equations (135a,b) and (132) are the desired three equations in q_1, q_2, and λ. The elimination of λ between (135a,b) gives

$$m_1 \ddot{q}_1 - m_2 \ddot{q}_2 = (m_1 - m_2) g, \tag{136}$$

and the elimination of q_2 between (136) and (132) then leads to (131).

From (135) we see that λ is the force exerted by the tension in the cord on each of the masses. By eliminating q_1 between (131) and (135a), we find that

$$\lambda = -2\left(\frac{m_1 m_2}{m_1 + m_2}\right) g. \tag{137}$$

The negative sign corresponds to the fact that the tensile force acts in the negative direction relative to q_1 and q_2.

As a second example, involving two constraints, we consider the rolling of a right circular cylinder of mass m on another cylinder, assuming the axes of the cylinders to be parallel. We choose the angles θ_1 and θ_2 of Figure 2.9 and the distance r between the centers as coordinates, noticing in advance that *so long as the cylinders are in contact* θ_1 and θ_2 are not independent, and r is actually constrained

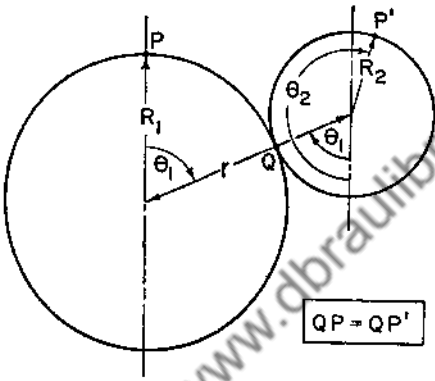


FIGURE 2.9

to remain constant. We notice that the rolling cylinder rotates through an angle $\theta_2 - \theta_1$ as the angle θ_1 is generated by the line of centers, and also that the kinetic energy of the rolling cylinder is composed of two parts: one of the form $\frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}_1^2)$ due to translation of the center of gravity, and one of the form $\frac{1}{2}(\frac{1}{2}m\,R_2^2)\dot{\theta}_2^2$ due to rotation *about* the center of gravity. Hence we have

$$T = \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}_1^2 + \tfrac{1}{2}R_2^2\dot{\theta}_2^2). \tag{138}$$

The potential energy can clearly be taken as

$$V = m\,g\,r\cos\theta_1 + \text{constant}. \tag{139}$$

The requirement of contact leads to the constraint equation

$$r - (R_1 + R_2) = 0, \tag{140}$$

while, if *pure rolling* is present, the condition $R_1\theta_1 = R_2(\theta_2 - \theta_1)$ leads to the frictional constraint equation

$$(R_1 + R_2)\theta_1 - R_2\theta_2 = 0. \tag{141}$$

The variational forms of (140) and (141) are then

$$\left.\begin{aligned} 1 \cdot \delta r + 0 \cdot \delta\theta_1 + 0 \cdot \delta\theta_2 &= 0, \\ 0 \cdot \delta r + (R_1 + R_2)\,\delta\theta_1 - R_2\,\delta\theta_2 &= 0 \end{aligned}\right\}. \tag{142a,b}$$

With the introduction of two Lagrange multipliers λ_1 and λ_2, the Lagrange equations (129) take the form

$$\left.\begin{aligned} m\ddot{r} &= -m\,g\cos\theta_1 + m\,r\,\dot{\theta}_1^2 + \lambda_1, \\ \frac{d}{dt}(m\,r^2\dot{\theta}_1) &= m\,g\,r\sin\theta_1 + \lambda_2(R_1 + R_2), \\ \tfrac{1}{2}m\,R_2^2\ddot{\theta}_2 &= -\lambda_2 R_2 \end{aligned}\right\}, \tag{143a,b,c}$$

the coefficients $\{1, 0, 0\}$ of λ_1 and $\{0, R_1 + R_2, -R_2\}$ of λ_2, in successive equations, being read from (142a,b).

If r and θ_2 are eliminated by use of (140) and (141), equations (143a,c) give

$$\lambda_1 = m\,g\cos\theta_1 - m(R_1 + R_2)\dot{\theta}_1^2, \tag{144a}$$

$$\lambda_2 = -\tfrac{1}{2}m\,R_2\ddot{\theta}_2 = -\tfrac{1}{2}m(R_1 + R_2)\ddot{\theta}_1, \tag{144b}$$

and the combination of (143b) and (144b) gives

$$m(R_1 + R_2)^2\ddot{\theta}_1 = m\,g(R_1 + R_2)\sin\theta_1 - \tfrac{1}{2}m(R_1 + R_2)^2\ddot{\theta}_1$$

or

$$\tfrac{3}{2}(R_1 + R_2)\ddot{\theta}_1 - g\sin\theta_1 = 0. \tag{145}$$

This is the required equation of motion, valid so long as contact persists.

From (143a) we see that λ_1 is the normal force (in the r-direction) exerted by the stationary cylinder on the rolling cylinder. Equation (144a) shows that this force is *positive* only when $g\cos\theta_1 > (R_1 + R_2)\dot{\theta}_1^2$, after which contact ceases and the constraints are removed. Equations (143a,b,c) then hold with $\lambda_1 = \lambda_2 = 0$.

From (143c) it follows that $-\lambda_2$ is the frictional force exerted on the rolling cylinder ($-\lambda_2 R_2$ is the corresponding torque about its center). By combining (144b) and (145), we obtain

$$-\lambda_2 = \tfrac{1}{3}m\,g\sin\theta_1. \tag{146}$$

If the constraints were of no interest, we would more economically introduce (140) and (141) directly into (138) and (139), to obtain

$$T = \tfrac{1}{2}m(R_1 + R_2)^2\dot{\theta}_1^2$$

and $$V = m\,g(R_1 + R_2)\cos\theta_1 + \text{constant}.$$

Since θ_1 is now an independent coordinate, equation (145) follows immediately as the relevant Lagrange equation. It should be noticed that no information as to the range of validity of (145) or as to the nature of the subsequent motion would then be obtained.

While the examples just considered, in which the constraint equations were of sufficiently simple form to permit direct elimination of superfluous coordinates, do not illustrate the *efficiency* of the method of Lagrange multipliers in more involved problems, they may serve to illustrate the *technique* involved. Further, they indicate the fact that such multipliers are very often capable of useful physical interpretation, and that their use in connection with superfluous coordinates may lead to the determination of unknown constraints when they are of interest.

At the same time, it may be appropriate to point out that the simplicity involved in the use of Lagrange's equations stems from the fact that unknown constraints can generally be omitted from consideration when they are *not* of interest.

2.12. Small vibrations about equilibrium. Normal coordinates. In many problems in dynamics we deal with a system for which there exists a stable *equilibrium* configuration in which the system can remain permanently at rest, and such that motions with small displacements and velocities can persist near the equilibrium state. If the system is specified by n generalized coordinates $q_1, q_2, \ldots, q_n$, we can choose these coordinates in such a way that they are all zero at an equilibrium position. For simplicity, we consider here only the case when $n = 2$, so that the system possesses two degrees of freedom and is completely specified by the coordinates q_1 and q_2. The results to be obtained are readily generalized.

In the case of a *conservative* system with two degrees of freedom, there exists a potential energy function V which depends only upon q_1 and q_2, and which can generally be expanded in a power series, near $q_1 = q_2 = 0$, of the form

$$V(q_1, q_2) = V_0 + \left(\frac{\partial V}{\partial q_1}\right)_0 q_1 + \left(\frac{\partial V}{\partial q_2}\right)_0 q_2 + \frac{1}{2}\left[\left(\frac{\partial^2 V}{\partial q_1^2}\right)_0 q_1^2 + 2\left(\frac{\partial^2 V}{\partial q_1\,\partial q_2}\right)_0 q_1 q_2 + \left(\frac{\partial^2 V}{\partial q_2^2}\right)_0 q_2^2\right] + \cdots, \tag{147}$$

where a zero subscript indicates evaluation at the equilibrium position $q_1 = q_2 = 0$. But, since V must be stationary at equilibrium, the linear terms must vanish. The constant V_0 is irrelevant, and it can be taken to be zero. Hence, if the terms of order greater than two in the expansion of V are neglected, we can write

$$V = \tfrac{1}{2}(a_{11}q_1^2 + 2a_{12}q_1q_2 + a_{22}q_2^2), \tag{148}$$

where the a's are constants defined by comparison with (147), so that (to a first approximation) V is a homogeneous quadratic function of q_1 and q_2, with constant coefficients.

The kinetic energy T is of the form

$$T = \tfrac{1}{2}(b_{11}\dot{q}_1^2 + 2b_{12}\dot{q}_1\dot{q}_2 + b_{22}\dot{q}_2^2), \tag{149}$$

where the b's may depend upon q_1 and q_2. For small departures from equilibrium, and small velocities, these coefficients may be replaced by their values when $q_1 = q_2 = 0$. Hence, in such cases, (149) expresses T as a homogeneous quadratic function of $\dot{q}_1$ and $\dot{q}_2$, with constant coefficients.

With these approximations, Lagrange's equations in the form (97b) become merely

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) + \frac{\partial V}{\partial q_i} = 0 \qquad (i = 1, 2) \tag{150}$$

or, with the notation of (148) and (149),

$$\left.\begin{aligned} b_{11}\ddot{q}_1 + b_{12}\ddot{q}_2 + a_{11}q_1 + a_{12}q_2 &= 0, \\ b_{12}\ddot{q}_1 + b_{22}\ddot{q}_2 + a_{12}q_1 + a_{22}q_2 &= 0 \end{aligned}\right\}. \tag{151a,b}$$

It is important to notice that these differential equations are *linear, with constant coefficients.*

If the equilibrium state is to be *stable*, V must possess a relative *minimum* at $q_1 = q_2 = 0$, so that the form (148) must be *positive* unless $q_1 = q_2 = 0$. Also, since the kinetic energy T cannot be negative, and cannot vanish unless all the *velocities* vanish, the

same must be true of the form (149). Quadratic forms having this property are said to be *positive definite* forms (see Section 1.17).

If we consider the coordinates q_1 and q_2 as the components of a *vector* $\mathbf{q}$, equations (151a,b) can be combined into the matrix equation

$$\mathbf{b}\,\ddot{\mathbf{q}} + \mathbf{a}\,\mathbf{q} = \mathbf{0}. \tag{152}$$

Following the usual procedure for solving such sets of equations, we seek solutions of the form

$$\mathbf{q} = \mathbf{x}\cos(\omega t + \gamma), \tag{153}$$

where the elements of $\mathbf{x}$ are the amplitudes of the required solutions, and are to be independent of t. Equation (152) then takes the form

$$\mathbf{a}\,\mathbf{x} = \omega^2 \mathbf{b}\,\mathbf{x}. \tag{154}$$

In this way we are led to a characteristic-value problem of the type considered in Section 1.25, with ω^2 identified with the parameter λ of that section. The characteristic values of ω^2 are thus the roots of the characteristic equation

$$|\,\mathbf{a} - \omega^2\mathbf{b}\,| \equiv \begin{vmatrix} a_{11} - \omega^2 b_{11} & a_{12} - \omega^2 b_{12} \\ a_{12} - \omega^2 b_{12} & a_{22} - \omega^2 b_{22} \end{vmatrix} = 0. \tag{155}$$

Since $\mathbf{a}$ and $\mathbf{b}$ are symmetric and positive definite, the results of Section 1.25 show that the roots of this quadratic equation in ω^2 are *real and positive*. However, they need not be *distinct*. Corresponding to each distinct root, the problem possesses a nontrivial vector solution. Furthermore, to a double root there correspond two linearly independent solutions which can be specified in infinitely many ways. Two solution vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ which correspond to distinct characteristic values of ω^2 are orthogonal with respect to both $\mathbf{a}$ and $\mathbf{b}$; that is, we have the relations

$$\mathbf{v}_1{}^T\,\mathbf{a}\,\mathbf{v}_2 = 0, \qquad \mathbf{v}_1{}^T\,\mathbf{b}\,\mathbf{v}_2 = 0 \tag{156}$$

in this case. Furthermore, the two linearly independent solution vectors corresponding to a repeated root of (155) can be so orthogonalized by the generalized Schmidt procedure, if this is desirable.

Thus if $\mathbf{v}_1 = \{v_{11}, v_{12}\}$ and $\mathbf{v}_2 = \{v_{21}, v_{22}\}$ are linearly independent characteristic vectors corresponding to $\omega_1{}^2$ and $\omega_2{}^2$, respec-

tively, where ω_1 and ω_2 need not be distinct, the solutions corresponding to the assumption (153) are then given by the expressions

$$\mathbf{q}^{(1)} = \mathbf{v}_1 \cos(\omega_1 t + \gamma_1), \qquad \mathbf{q}^{(2)} = \mathbf{v}_2 \cos(\omega_2 t + \gamma_2) \tag{157}$$

where γ_1 and γ_2 are arbitrary constants. The most general solution is then obtained by superposition, in the vector form

$$\mathbf{q} = c_1\mathbf{v}_1 \cos(\omega_1 t + \gamma_1) + c_2\mathbf{v}_2 \cos(\omega_2 t + \gamma_2), \tag{158}$$

or in the expanded form

$$\left.\begin{aligned} q_1 &= c_1 v_{11} \cos(\omega_1 t + \gamma_1) + c_2 v_{21} \cos(\omega_2 t + \gamma_2), \\ q_2 &= c_1 v_{12} \cos(\omega_1 t + \gamma_1) + c_2 v_{22} \cos(\omega_2 t + \gamma_2) \end{aligned}\right\}, \tag{159a,b}$$

where c_1 and c_2 are also arbitrary constants.

It follows that *the most general motion of the specified system is a superposition of two simple-harmonic motions*, the *natural frequencies* of which are $\omega_i/2\pi$. The fact that this statement is true even in the case when (155) possesses repeated roots is of particular importance.

The general solution (158) or (159a,b) can be written in the matrix form

$$\begin{Bmatrix} q_1 \\ q_2 \end{Bmatrix} = \begin{bmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{bmatrix} \begin{Bmatrix} c_1 \cos(\omega_1 t + \gamma_1) \\ c_2 \cos(\omega_2 t + \gamma_2) \end{Bmatrix} \tag{160a}$$

or

$$\mathbf{q} = \mathbf{M}\,\mathbf{C}, \tag{160b}$$

where $\mathbf{M}$ is a *modal matrix* having the components of successive characteristic vectors as the elements of its successive *columns*, and $\mathbf{C}$ denotes the vector multiplied by $\mathbf{M}$ in (160a). If the equal members of (160b) are premultiplied by $\mathbf{M}^{-1}$, there follows simply

$$\boldsymbol{\alpha} = \mathbf{C}, \tag{161}$$

where the vector $\boldsymbol{\alpha}$ is defined by the relation

$$\mathbf{q} = \mathbf{M}\,\boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} = \mathbf{M}^{-1}\mathbf{q}. \tag{162a,b}$$

Thus the new coordinates α_1 and α_2 so defined are such that the general solution of (151a,b) or (152) can be expressed in the simple form

$$\alpha_1 = c_1 \cos(\omega_1 t + \gamma_1), \quad \alpha_2 = c_2 \cos(\omega_2 t + \gamma_2), \tag{163}$$

the two modes of vibration then being *uncoupled*. These new coordinates are often called *normal coordinates* of the problem.

Since the characteristic vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are each determined only within a multiplicative arbitrary constant, the modal matrix $\mathbf{M}$ and the normal coordinates α_1, α_2 are not uniquely defined. For some purposes it is convenient to *normalize* the vectors relative to either a or b.* In particular, we may determine multiples of $\mathbf{v}_1$ and $\mathbf{v}_2$, say $\mathbf{e}_1$ and $\mathbf{e}_2$, which are normalized relative to b, so that

$$\mathbf{e}_i^T \mathbf{b}\, \mathbf{e}_j = \delta_{ij} \tag{164}$$

where δ_{ij} is the Kronecker delta. It then follows that the *normalized modal matrix* $\mathbf{M}$ made up of these *normalized* characteristic vectors has the property that it satisfies the equation

$$\mathbf{M}^T \mathbf{b}\, \mathbf{M} = \mathbf{I}, \tag{165}$$

where I is the *unit matrix*. By postmultiplying both members of (165) by $\mathbf{M}^{-1}$, there then follows $\mathbf{M}^{-1} = \mathbf{M}^T \mathbf{b}$, so that (162) takes the more convenient form

$$\mathbf{q} = \mathbf{M}\, \boldsymbol{\alpha}, \qquad \boldsymbol{\alpha} = \mathbf{M}^T \mathbf{b}\, \mathbf{q} \tag{166a,b}$$

in this case. Furthermore, it then follows (as in Section 1.25) that also

$$\mathbf{M}^T \mathbf{a}\, \mathbf{M} = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{bmatrix}. \tag{167}$$

Consequently, with the substitution (166), the potential and kinetic energies then become

$$\left.\begin{aligned} V &= \tfrac{1}{2}\mathbf{q}^T \mathbf{a}\, \mathbf{q} = \tfrac{1}{2}\boldsymbol{\alpha}^T(\mathbf{M}^T \mathbf{a}\, \mathbf{M})\boldsymbol{\alpha} = \tfrac{1}{2}(\omega_1^2\alpha_1^2 + \omega_2^2\alpha_2^2), \\ T &= \tfrac{1}{2}\dot{\mathbf{q}}^T \mathbf{b}\, \dot{\mathbf{q}} = \tfrac{1}{2}\dot{\boldsymbol{\alpha}}^T(\mathbf{M}^T \mathbf{b}\, \mathbf{M})\dot{\boldsymbol{\alpha}} = \tfrac{1}{2}(\dot{\alpha}_1^2 + \dot{\alpha}_2^2) \end{aligned}\right\}. \tag{168a,b}$$

The corresponding Lagrangian equations are then simply

$$\ddot{\alpha}_1 + \omega_1^2\alpha_1 = 0, \qquad \ddot{\alpha}_2 + \omega_2^2\alpha_2 = 0, \tag{169}$$

and equations (163) follow immediately.

More generally, if the columns of $\mathbf{M}$ are *not* necessarily normal-

* In some references, the coordinates α_i and α_j are called *natural* coordinates, and are said to be *normal* coordinates only when they are *normalized* in a certain way, as the terminology suggests.

ized relative to $\mathbf{b}$, it is easily shown that the substitution $\mathbf{q} = \mathbf{M}\,\boldsymbol{\alpha}$ reduces V and T to the forms

$$V = \tfrac{1}{2}(\omega_1^2 f_1^2 \alpha_1^2 + \omega_2^2 f_2^2 \alpha_2^2), \quad T = \tfrac{1}{2}(f_1^2 \dot{\alpha}_1^2 + f_2^2 \dot{\alpha}_2^2) \qquad (170a,b)$$

where

$$f_i^2 = \mathbf{v}_i^T\,\mathbf{b}\,\mathbf{v}_i. \qquad (171)$$

The corresponding Lagrangian equations,

$$f_1^2(\ddot{\alpha}_1 + \omega_1^2 \alpha_1) = 0, \quad f_2^2(\ddot{\alpha}_2 + \omega_2^2 \alpha_2) = 0$$

are seen to be of the same form as (169), in accordance with the previously established validity of (163). However, unless (165) is true the relation (166b) is not a consequence of (166a), and (162b) must be used instead, for the purpose of expressing α_1 and α_2 explicitly in terms of q_1 and q_2.

Unless it is desirable to actually reduce the expressions for the potential and kinetic energies to the standard forms (168a,b), it is clear that there is little to be gained by normalizing the characteristic vectors in the developments under consideration.

In the more general case, the impressed force system may consist of a conservative part, derivable from a potential energy V, and also of a *dissipative* (nonconservative) part. In particular, there may exist resistive forces R_1 and R_2, associated with q_1 and q_2, which are proportional to the velocities, and hence are expressible in the forms

$$R_1 = -(r_{11}\dot{q}_1 + r_{12}\dot{q}_2), \quad R_2 = -(r_{21}\dot{q}_1 + r_{22}\dot{q}_2). \qquad (172)$$

These terms would then be added to the right-hand members of (151a,b), thus introducing velocity terms into the linearized equations of motion. In the special case when $r_{21} = r_{12}$ (in particular, when these two coupling coefficients are zero), if we define the function

$$F = \tfrac{1}{2}(r_{11}\dot{q}_1^2 + 2r_{12}\dot{q}_1\dot{q}_2 + r_{22}\dot{q}_2^2), \qquad (173)$$

we see that

$$R_i = -\frac{\partial F}{\partial \dot{q}_i} \qquad (i = 1, 2). \qquad (174)$$

The function F is then known as *Rayleigh's dissipation function.*

If we denote any external forces associated with q_1 and q_2 which are dissipative but *not* derivable from a dissipation function by

Q_1' and Q_2', respectively, equation (150) must be modified to read

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) + \frac{\partial V}{\partial q_i} + \frac{\partial F}{\partial \dot{q}_i} = Q_i' \qquad (i = 1, 2), \tag{175}$$

to include both conservative and dissipative forces. With the notation of (148), (149), and (173), this set of equations can be combined in the matrix form

$$\mathbf{b}\,\ddot{\mathbf{q}} + \mathbf{r}\,\dot{\mathbf{q}} + \mathbf{a}\,\mathbf{q} = \mathbf{Q}' \tag{176}$$

where $\mathbf{b}$, $\mathbf{r}$, and $\mathbf{a}$ are symmetric square matrices and $\mathbf{Q}'$ is the column vector $\{Q_1', Q_2'\}$.

The coefficients b_{ij} in (149) are often known as the *inertia coefficients* associated with q_1 and q_2, and the coefficients a_{ij} in (148) as the *stiffness coefficients*. The coefficients r_{ij} in (173), which govern deviations from simple-harmonic motions when suitable dissipative forces are present, are often called the associated *resistance coefficients*.

2.13. Numerical example. To illustrate the results of the preceding section, we consider the determination of natural modes of small vibration of the compound pendulum of Section 2.9. We obtain the form (148),

$$V = \tfrac{1}{2}(m_1 + m_2)g\,L_1\theta_1^2 + \tfrac{1}{2}m_2 g\,L_2\theta_2^2, \tag{177}$$

by retaining leading terms in the expansion of (100b) and the form (149),

$$T = \tfrac{1}{2}(m_1 + m_2)L_1^2\dot{\theta}_1^2 + m_2L_1L_2\dot{\theta}_1\dot{\theta}_2 + \tfrac{1}{2}m_2L_2^2\dot{\theta}_2^2, \tag{178}$$

by setting $\theta_1 = \theta_2 = 0$ in (100a). That $\theta_1 = \theta_2 = 0$ actually specifies an equilibrium state (as is clear from physical considerations) follows mathematically from the fact that $\partial V/\partial\theta_1 = \partial V/\partial\theta_2 = 0$ when $\theta_1 = \theta_2 = 0$.

The resultant equations of small oscillations,

$$(m_1 + m_2)L_1^2\ddot{\theta}_1 + m_2L_1L_2\ddot{\theta}_2 + (m_1 + m_2)g\,L_1\theta_1 = 0 \tag{179a}$$

and

$$m_2L_1L_2\ddot{\theta}_1 + m_2L_2^2\ddot{\theta}_2 + m_2g\,L_2\theta_2 = 0, \tag{179b}$$

are equivalent to the results of linearizing (101a,b) in the displacements and velocities.

We now consider explicitly the special case in which

$$m_1 = m_2 \equiv m, \qquad L_1 = L_2 \equiv L, \tag{180}$$

so that, after removing a factor $m L^2$ from the equal members of the governing equations, there follows

$$\left.\begin{aligned} 2\ddot{\theta}_1 + \ddot{\theta}_2 + 2\frac{g}{L}\theta_1 = 0, \\ \ddot{\theta}_1 + \ddot{\theta}_2 + \frac{g}{L}\theta_2 = 0 \end{aligned}\right\}. \tag{181a,b}$$

Corresponding to the assumption

$$\boldsymbol{\theta} = \mathbf{x}\cos(\omega t + \gamma), \tag{182}$$

the equation corresponding to (154) is obtained in the form

$$\mathbf{a}\,\mathbf{x} = \bar{\omega}^2\mathbf{b}\,\mathbf{x}, \tag{183}$$

where

$$\mathbf{a} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{184a,b}$$

and where $\bar{\omega}$ is a *dimensionless* parameter defined by the relation

$$\bar{\omega}^2 = \frac{L}{g}\omega^2. \tag{185}$$

The matrices **a** and **b** so defined differ from those defined in the preceding section only in that their elements have been made dimensionless.

We have then to deal with the matrix

$$[\mathbf{a} - \bar{\omega}^2\mathbf{b}] = \begin{bmatrix} 2(1-\bar{\omega}^2) & -\bar{\omega}^2 \\ -\bar{\omega}^2 & 1-\bar{\omega}^2 \end{bmatrix}, \tag{186}$$

the vanishing of the determinant of which leads to the characteristic equation

$$\bar{\omega}^4 - 4\bar{\omega}^2 + 2 = 0. \tag{187}$$

Corresponding to the smaller root,

$$\bar{\omega}_1^2 = 0.586, \tag{188}$$

the matrix (186) takes the form

$$[\mathbf{a} - \bar{\omega}_1^2\mathbf{b}] = \begin{bmatrix} 0.828 & -0.586 \\ -0.586 & 0.414 \end{bmatrix}$$

and the elements of the modal column are proportional to the *cofactors* of the elements in either *row* of this matrix. By choosing the first row, we may hence take $\mathbf{v}_1$ in the form

$$\mathbf{v}_1 = \begin{Bmatrix} 0.414 \\ 0.586 \end{Bmatrix}. \tag{189}$$

Similarly, in correspondence with the other root of (187),

$$\bar{\omega}_2^2 = 3.414, \tag{190}$$

the modal column $\mathbf{v}_2$ may be taken in the form

$$\mathbf{v}_2 = \begin{Bmatrix} -2.414 \\ 3.414 \end{Bmatrix}. \tag{191}$$

The general solution of (181a,b) can then be expressed by the relations

$$\left.\begin{aligned} q_1 &= 0.414c_1 \cos(\omega_1 t + \gamma_1) - 2.414c_2 \cos(\omega_2 t + \gamma_2), \\ q_2 &= 0.586c_1 \cos(\omega_1 t + \gamma_1) + 3.414c_2 \cos(\omega_2 t + \gamma_2) \end{aligned}\right\}, \tag{192}$$

where

$$\omega_1 = 0.765\sqrt{\frac{g}{L}}, \qquad \omega_2 = 1.848\sqrt{\frac{g}{L}}, \tag{193}$$

and where c_1, c_2, γ_1, and γ_2 are arbitrary constants.

The modal matrix made up of the elements of (189) and (191) is then of the form

$$\mathbf{M} = \begin{bmatrix} 0.414 & -2.414 \\ 0.586 & 3.414 \end{bmatrix}, \tag{194}$$

and the corresponding normal coordinates are defined by the matrix equation

$$\boldsymbol{\alpha} = \mathbf{M}^{-1}\mathbf{q} = \begin{bmatrix} 1.207 & 0.854 \\ -0.207 & 0.146 \end{bmatrix} \mathbf{q}. \tag{195}$$

The modal columns (189) and (191) happen to be normalized in such a way that the elements in each column add to unity. By multiplying each column by a suitable constant, these columns can be normalized in other ways, and multiples of the present α_1 and α_2 (in terms of which the natural modes are *also* uncoupled) are then obtained.

2.14. Variational problems for deformable bodies. General variational principles have been established in connection with the theory of elasticity,* as well as in many other fields. In this section no attempts are made to establish such general theories. Instead, it is shown in what way the variational problem can be derived *from the differential equation and associated boundary conditions*, in certain illustrative cases. In later sections it is shown that such formulations are readily adapted to *approximate* analysis.

We start with the problem of determining *small deflections of a rotating string* of length L. The governing differential equation is then of the form

$$\frac{d}{dx}\left(F\frac{dy}{dx}\right) + \rho\,\omega^2 y + p = 0, \tag{196}$$

where $y(x)$ is the displacement of a point from the axis of rotation, $F(x)$ is the tension, $\rho(x)$ the linear mass density, ω the angular velocity of rotation, and $p(x)$ is the intensity of a distributed radial load. Suitable end conditions are also to be prescribed.

In order to formulate a corresponding variational problem, we first multiply both sides of (196) by a variation δy and integrate the result over $(0, L)$ to obtain

$$\int_0^L \frac{d}{dx}\left(F\frac{dy}{dx}\right)\delta y\,dx + \int_0^L \rho\,\omega^2 y\,\delta y\,dx + \int_0^L p\,\delta y\,dx = 0. \tag{197}$$

The second and third integrands are the variations of $\frac{1}{2}\rho\,\omega^2 y^2$ and $p\,y$, respectively. If the first integral is transformed by integration by parts, it takes the form

$$\left[F\frac{dy}{dx}\,\delta y\right]_0^L - \int_0^L F\frac{dy}{dx}\,\delta\frac{dy}{dx}\,dx,$$

and the integrand in this form is the variation of $\frac{1}{2}F\,(dy/dx)^2$. Thus the left-hand member of (197) can be transformed to the left-hand member of the equation

$$\delta\int_0^L \left[\frac{1}{2}\rho\,\omega^2 y^2 + p\,y - \frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right]dx + \left[F\frac{dy}{dx}\,\delta y\right]_0^L = 0. \tag{198}$$

* See Reference 5.

If we impose at each of the two ends the condition that either

$$y = y_0 \quad \text{or} \quad F\frac{dy}{dx} = 0 \qquad \text{(when } x = 0, L\text{)} \tag{199}$$

where y_0 is a prescribed constant, the integrated terms in (198) vanish and the condition becomes

$$\delta \int_0^L \left[\frac{1}{2}\rho\,\omega^2 y^2 + p\,y - \frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right] dx = 0. \tag{200}$$

Conversely, (196) is the Euler equation [(17a)] corresponding to (200). That is, if y renders the integral in (200) stationary it must satisfy (196), while if y satisfies (196) and end conditions of the type required in (199), then y renders the integral in (200) stationary.

The end conditions (199) or, equivalently,

$$\left[F\frac{dy}{dx}\,\delta y\right]_0^L = 0, \tag{201}$$

are the so-called *natural boundary conditions* of the variational problem (200). If we recall that (in the linearized theory) the product $F\,(dy/dx)$ is the component of the tensile force normal to the axis of rotation, we see that (201) *requires that the end tensions do no work*. This situation exists if no end motion is permitted ($\delta y = 0$), or if no end restraint (normal to the axis of rotation) is present [$F\,(dy/dx) = 0$].

Thus, *if the end tensions do no work*, we conclude that *of all functions $y(x)$ which satisfy the relevant end conditions, that one which also satisfies the relevant differential equation* (196) *renders the integral in* (200) *stationary*.

It is clear that the term $\frac{1}{2}\rho(\omega\,y)^2$ in (200) represents the *kinetic energy* of the string per unit length, since the speed of an element of the string is given by $\omega\,y$. Also, since $p\,\delta y\,dx$ is the element of work done by p on an element dx in a displacement δy, the term $-p\,y$ is *potential energy* per unit length due to the radial force distribution $p(x)$. To identify the remaining term, we notice that an element of original length dx stretches into an element of length

$$ds = \left[1 + \left(\frac{dy}{dx}\right)^2\right]^{1/2} dx.$$

The work per unit length done *against* the tensile force is then

$$F\frac{ds-dx}{dx} = F\left\{\left[1+\left(\frac{dy}{dx}\right)^2\right]^{\frac{1}{2}}-1\right\}$$
$$= F\left[1+\frac{1}{2}\left(\frac{dy}{dx}\right)^2+\cdots-1\right]$$
$$\approx \frac{1}{2}F\left(\frac{dy}{dx}\right)^2,$$

if higher powers of the slope dy/dx (assumed to be small) are neglected. Thus this term represents potential energy per unit length due to the tension in the string, to a first approximation.

Finally, (200) requires that the difference between the total kinetic and total potential energies be stationary, in analogy with Hamilton's principle.*

For the case of a *yielding support* at the end $x = 0$, the end condition at that point would be of the form

$$\left(F\frac{dy}{dx}\right)_{x=0} = k(y)_{x=0} \tag{202}$$

where k is the modulus of the support. There would then follow

$$\left(F\frac{dy}{dx}\,\delta y\right)_{x=0} = (k\,y\,\delta y)_{x=0} = \delta\left(\frac{1}{2}k\,y^2\right)_{x=0}.$$

Since this term would not vanish, (200) would be replaced by

$$\delta\left\{\int_0^L\left[\frac{1}{2}\rho\,\omega^2y^2+p\,y-\frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right]dx-\left(\frac{1}{2}k\,y^2\right)_{x=0}\right\}=0, \tag{203}$$

the additional term representing the potential energy stored in the support.

If the *slope* of the string at the end $x = 0$ were prescribed as $y'(0) = \alpha$, where α is small, the deflection $y(0)$ then being unknown, there would follow

$$\left(F\frac{dy}{dx}\,\delta y\right)_{x=0} = (F\,\alpha\,\delta y)_{x=0},$$

* For a complete analogy, we should require that the *time integral* of this difference over (t_1, t_2) be stationary. In the present case, however, the energy difference is *independent* of time.

and (200) would be replaced by

$$\delta\left\{\int_0^L\left[\frac{1}{2}\rho\,\omega^2y^2+p\,y-\frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right]dx-F(0)\,\alpha\,y(0)\right\}=0, \quad (204)$$

the additional term corresponding to work done by the component of the tension normal to the x-axis ($F\sin\alpha\approx F\,\alpha$) in the end displacement $y(0)$.

In both (203) and (204), admissible functions must satisfy the single end condition $y(L)=0$.

As a second example, we consider the case of *small deflections of a rotating shaft* of length L, subject to an axial end load P and to distributed transverse loading of intensity $p(x)$. The deflection $y(x)$ is then governed by the differential equation

$$\frac{d^2}{dx^2}\left(E\,I\,\frac{d^2y}{dx^2}\right)+P\,\frac{d^2y}{dx^2}-\rho\,\omega^2y-p=0, \quad (205)$$

where $E\,I$ is the bending stiffness of the shaft. We first form the equation

$$\int_0^L(E\,I\,y'')''\,\delta y\,dx+P\int_0^L y''\,\delta y\,dx-\int_0^L(\rho\,\omega^2y+p)\,\delta y\,dx=0. \quad (206)$$

If the first term is integrated twice by parts, it becomes

$$\Big[\,(E\,I\,y'')'\,\delta y-E\,I\,y''\,\delta y'\Big]_0^L+\int_0^L E\,I\,y''\,\delta y''\,dx,$$

and the new integrand is recognized as the variation of $\frac{1}{2}E\,I(y'')^2$. After one integration by parts, the second term in (206) becomes

$$\Big[P\,y'\,\delta y\Big]_0^L-P\int_0^L y'\,\delta y'\,dx=\Big[P\,y'\,\delta y\Big]_0^L-\delta\int_0^L\frac{1}{2}P(y')^2\,dx.$$

Thus (206) implies the equation

$$\delta\int_0^L\left[\frac{1}{2}E\,I\,(y'')^2-\frac{1}{2}P(y')^2-\frac{1}{2}\rho\,\omega^2y^2-p\,y\right]dx$$
$$+\Big[\,\{(E\,I\,y'')'+P\,y'\}\,\delta y-E\,I\,y''\,\delta y'\Big]_0^L=0. \quad (207)$$

The value of $[(E\,I\,y'')'+P\,y']\,\delta y$ at an end point can be interpreted as the work done by the total transverse shear at that point in a *displacement* δy, while the value of $E\,I\,y''\,\delta y'$ is the work

done by the end bending moment $E\,I\,y''$ in a *rotation* (change of *slope*) $\delta y'$. The total work done by end forces and moments will vanish in case of satisfaction of the *natural boundary conditions*

$$\left[\{(E\,I\,y'')' + P\,y'\}\,\delta y - E\,I\,y''\,\delta y'\right]_0^L = 0$$

or, explicitly,

$$\left.\begin{aligned} &y = y_0 \quad \text{or} \quad (E\,I\,y'')' + P\,y' = 0 \\ \textit{and} \quad &y' = y_0' \quad \text{or} \quad E\,I\,y'' = 0 \end{aligned}\right\} \quad (\text{when } x = 0,\ L). \tag{208}$$

For any case in which such conditions are to be satisfied, the variational problem (207) reduces to the form

$$\delta \int_0^L \left[\frac{1}{2} E\,I(y'')^2 - \frac{1}{2} P(y')^2 - \frac{1}{2}\rho\,\omega^2 y^2 - p\,y\right] dx = 0. \tag{209}$$

Here $\frac{1}{2}\rho\,\omega^2 y^2$ is the kinetic energy per unit length, and the remaining terms $\frac{1}{2}E\,I(y'')^2$, $-\frac{1}{2}P(y')^2$, and $-p\,y$ can be identified with potential energies per unit length due to bending and to the end thrust and lateral loading, respectively.

If end supports do work in bending the shaft, additional terms must be added to the integral in (209) in analogy with (203) and (204).

We notice that (209) involves the *second* as well as the first derivative of y. That (205) is truly the Euler equation of (209) can be established directly [see (43)], or by retracing the above steps leading from (205) to (209).

As a third example, we consider *small steady-state forced vibration of a rectangular membrane*. The basic equation is of the form

$$\frac{\partial}{\partial x}\left(F\,\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(F\,\frac{\partial u}{\partial y}\right) + P = \rho\,\frac{\partial^2 u}{\partial t^2}$$

where u is displacement, F tension, ρ is surface mass density, and $P(x, y, t)$ is the impressed periodic normal force. If P is of the form $P = p(x, y)\sin(\omega t + \alpha)$, we may write the steady-state displacement u in the form

$$u = w(x, y)\sin(\omega t + \alpha),$$

where w is the amplitude of the oscillation. The amplitude w must then satisfy the equation

$$(F\,w_x)_x + (F\,w_y)_y + \rho\,\omega^2 w + p = 0. \tag{210}$$

After multiplying by the variation $\delta w(x, y)$ and integrating the results over the membrane ($x_1 \leqq x \leqq x_2$, $y_1 \leqq y \leqq y_2$), there follows

$$\int_{x_1}^{x_2}\int_{y_1}^{y_2} (F\,w_x)_x\,\delta w\,dx\,dy + \int_{x_1}^{x_2}\int_{y_1}^{y_2} (F\,w_y)_y\,\delta w\,dx\,dy + \delta\int_{x_1}^{x_2}\int_{y_1}^{y_2}\left(\frac{1}{2}\rho\,\omega^2 w^2 + p\,w\right)dx\,dy = 0. \qquad (211)$$

After a partial integration with respect to x, the first term takes the form

$$\int_{y_1}^{y_2}\left\{\left[F\,w_x\,\delta w\right]_{x_1}^{x_2} - \int_{x_1}^{x_2} F\,w_x\,\delta w_x\,dx\right\}dy \equiv -\delta\int_{x_1}^{x_2}\int_{y_1}^{y_2}\frac{1}{2}F\,w_x^2\,dx\,dy + \int_{y_1}^{y_2}\left[F\,w_x\,\delta w\right]_{x_1}^{x_2}dy.$$

If the second integral is transformed in a similar way, (211) becomes

$$\delta\int_{x_1}^{x_2}\int_{y_1}^{y_2}\left[-\frac{1}{2}F(w_x^2 + w_y^2) + \frac{1}{2}\rho\,\omega^2 w^2 + p\,w\right]dx\,dy + \int_{y_1}^{y_2}\left[F\,w_x\,\delta w\right]_{x_1}^{x_2}dy + \int_{x_1}^{x_2}\left[F\,w_y\,\delta w\right]_{y_1}^{y_2}dx = 0. \qquad (212)$$

This result can be written more concisely in a form independent of the coordinate system as follows:

$$\delta\iint_A\left[\frac{1}{2}F\,(\nabla w)^2 - \frac{1}{2}\rho\,\omega^2 w^2 - p\,w\right]dS - \oint_C F\,\frac{\partial w}{\partial n}\,\delta w\,ds = 0. \qquad (212a)$$

Here A is the area of the undeformed membrane and C is its closed boundary. The term $\partial w/\partial n$ is the derivative of w in the direction of the outward *normal* to the boundary. By more general considerations (see Problem 20), it can be shown that (212a) is the proper variational form for a membrane of arbitrary contour.

The single integral in (212a) represents total work of the transverse component of the edge tensions, and vanishes if w is prescribed along C or, more generally, if the restraining transverse force $F\,\partial w/\partial n$ vanishes at all parts of the boundary which are not fixed. In all such cases the relevant variational problem becomes

$$\delta\iint_A\left[\frac{1}{2}F\,(\nabla w)^2 - \frac{1}{2}\rho\,\omega^2 w^2 - p\,w\right]dS = 0. \qquad (213)$$

The term $\frac{1}{2}\rho\,\omega^2 w^2$ is the kinetic energy (per unit area) corresponding to positions of maximum displacement, while the term $\frac{1}{2}F(w_x^2 + w_y^2)$ represents the potential energy stored in the membrane at such instants as a result of the stretching, and $-p\,w$ is the corresponding potential energy due to the loading.

If, instead of prescribing w along part of the boundary C, and requiring that $F\,\partial w/\partial n$ *vanish* along the remainder of the boundary, we require that $\partial w/\partial n = \psi(s)$ along the portion C' where w is not prescribed, equation (212a) shows that the term $-\delta \int_{C'} F\psi\, w\, ds$ must be added to the left-hand member of (213).

When $F = 1$ and $\omega = 0$, equation (210) is *Poisson's equation*. When also $p = 0$, this equation becomes *Laplace's equation*. The variational form of the *Dirichlet problem*, where $w = \phi(s)$ along C, then takes the form $\delta \iint_A \frac{1}{2}(\nabla w)^2\, dS = 0$, where the varied functions are to take on the prescribed values along C. The variational form of the *Neumann problem*, where $\partial w/\partial n = \psi(s)$ along C, becomes

$$\delta\left\{\iint_A \frac{1}{2}(\nabla w)^2\, dS - \oint_C \psi\, w\, ds\right\} = 0,$$

where the varied functions are *unrestricted* along C.

2.15. Useful transformations. Certain formulas of frequent use in transformations of the type considered in the preceding section are collected together in this section, for convenient reference.

The formula

$$\int_{x_1}^{x_2} (p f_x)_x\, \delta f\, dx = -\delta \int_{x_1}^{x_2} \left(\frac{1}{2} p f_x^2\right) dx + \Big[p f_x\, \delta f\Big]_{x_1}^{x_2},$$

established by integration by parts, implies the relation

$$(p f_x)_x\, \delta f = -\delta(\tfrac{1}{2} p f_x^2) + (p f_x\, \delta f)_x. \tag{214}$$

Here p is an explicit function of x, which is not to be varied. In a similar way, the following relations can be established:

$$(s f_{xx})_{xx}\, \delta f = \delta(\tfrac{1}{2} s f_{xx}^2) + [(s f_{xx})_x\, \delta f - s f_{xx}\, \delta f_x]_x. \tag{215}$$

$$[(p f_y)_x + (p f_x)_y]\, \delta f = -\delta(p f_x f_y) + (p f_y\, \delta f)_x + (p f_x\, \delta f)_y. \tag{216}$$

$$2 f_{xy}\, \delta f = -\delta(f_x f_y) + (f_y\, \delta f)_x + (f_x\, \delta f)_y. \tag{217}$$

$$2f_{xxyy}\,\delta f = \begin{cases} \delta(f_{xy}{}^2) + 2(f_{xyy}\,\delta f)_x + 2(f_{xxy}\,\delta f)_y - 2(f_{xy}\,\delta f)_{xy}, & \text{(218a)} \\ \delta(f_{xx}f_{yy}) + (f_{xyy}\,\delta f - f_{yy}\,\delta f_x)_x + (f_{xxy}\,\delta f - f_{xx}\,\delta f_y)_y. & \text{(218b)} \end{cases}$$

The differentiations in (214) and (215) may be total or partial. In each case, the truth of the relation can be verified directly by expanding both sides of the equation.

In each of the preceding formulas, the left-hand member is expressed as the sum of an *exact variation* and one or more *derivatives*. It is of interest to notice that $2f_{xxyy}\,\delta f$ can be expressed thus in two different ways, according to (218a,b). The two alternatives can be combined by expressing the left-hand member as $(1-\alpha)$ times (218a) plus α times (218b), where α is a completely arbitrary constant. Thus we may write

$$\begin{aligned} 2f_{xxyy}\,\delta f = {} & \delta[(1-\alpha)f_{xy}{}^2 + \alpha f_{xx}f_{yy}] \\ & + [(2-\alpha)f_{xyy}\,\delta f - \alpha f_{yy}\,\delta f_x]_x + [(2-\alpha)f_{xxy}\,\delta f - \alpha f_{xx}\,\delta f_y]_y \\ & - 2(1-\alpha)[f_{xy}\,\delta f]_{xy}, \end{aligned} \qquad \text{(219)}$$

where α is an arbitrary constant. This form reduces to (218a) when $\alpha = 0$, and to (218b) when $\alpha = 1$.

If we take $p = 1$ in (214), and add to this expression the result of replacing x by y, we obtain the further useful result

$$\nabla^2 f\,\delta f = -\delta[\tfrac{1}{2}(\nabla f)^2] + (f_x\,\delta f)_x + (f_y\,\delta f)_y. \qquad \text{(220)}$$

As a further example of the use of these formulas, we consider the product

$$\nabla^4 f\,\delta f \equiv (f_{xxxx} + 2f_{xxyy} + f_{yyyy})\,\delta f.$$

If use is made of (215) and (219), there follows

$$\begin{aligned} \nabla^4 f\,\delta f = {} & \delta(\tfrac{1}{2}f_{xx}{}^2) + [f_{xxx}\,\delta f - f_{xx}\,\delta f_x]_x \\ & + \delta[(1-\alpha)f_{xy}{}^2 + \alpha f_{xx}f_{yy}] + [(2-\alpha)f_{xyy}\,\delta f - \alpha f_{yy}\,\delta f_x]_x \\ & + [(2-\alpha)f_{xxy}\,\delta f - \alpha f_{xx}\,\delta f_y]_y - 2(1-\alpha)[f_{xy}\,\delta f]_{xy} \\ & + \delta(\tfrac{1}{2}f_{yy}{}^2) + [f_{yyy}\,\delta f - f_{yy}\,\delta f_y]_y, \end{aligned}$$

or, after collecting terms,

$$\nabla^4 f\,\delta f = \delta\left[\frac{1}{2}(f_{xx}{}^2 + f_{yy}{}^2) + (1-\alpha)f_{xy}{}^2 + \alpha f_{xx}f_{yy}\right]$$

$$+\frac{\partial}{\partial x}\left\{\left[\frac{\partial \nabla^2 f}{\partial x}+(1-\alpha)f_{xyy}\right]\delta f-(f_{xx}+\alpha f_{yy})\,\delta f_x\right\}$$

$$+\frac{\partial}{\partial y}\left\{\left[\frac{\partial \nabla^2 f}{\partial y}+(1-\alpha)f_{xxy}\right]\delta f-(f_{yy}+\alpha f_{xx})\,\delta f_y\right\}$$

$$-\frac{\partial^2}{\partial x\,\partial y}[2(1-\alpha)f_{xy}\,\delta f]. \qquad (221)$$

2.16. The variational problem for the elastic plate. As a final illustration of the preceding methods, we consider the problem of determining the amplitude w of small deflections of a thin, initially flat, elastic plate of constant thickness. If the amplitude of the periodic impressed force is denoted by $p(x, y)$ and the circular frequency by ω, it is known* that under certain simplifying assumptions the governing differential equation is of the form

$$D\,\nabla^4 w-\rho\,\omega^2 w-p=0. \qquad (222)$$

Here D is a constant known as the *bending stiffness* of the plate.

We consider here a rectangular plate ($0 \leqq x \leqq a$, $0 \leqq y \leqq b$). Then, by multiplying both sides of (222) by a variation δw and integrating the result over the area of the plate, there follows

$$\int_0^a\int_0^b D\,\nabla^4 w\,\delta w\,dx\,dy-\delta\int_0^a\int_0^b\left[\frac{1}{2}\rho\,\omega^2 w^2+p\,w\right]dx\,dy=0. \qquad (222a)$$

If use is made of equation (221), this condition is transformed immediately into the requirement that

$$\delta\int_0^a\int_0^b\left\{\frac{1}{2}D[w_{xx}^2+w_{yy}^2+2\alpha\,w_{xx}w_{yy}+2(1-\alpha)w_{xy}^2]-\frac{1}{2}\rho\,\omega^2 w^2-p\,w\right\}dx\,dy$$

$$+\int_0^b\left\{\left[D\frac{\partial \nabla^2 w}{\partial x}+(1-\alpha)D\,w_{xyy}\right]\delta w-D(w_{xx}+\alpha\,w_{yy})\,\delta w_x\right\}_{x=0}^{x=a}dy$$

$$+\int_0^a\left\{\left[D\frac{\partial \nabla^2 w}{\partial y}+(1-\alpha)D\,w_{xxy}\right]\delta w-D(w_{yy}+\alpha\,w_{xx})\,\delta w_y\right\}_{y=0}^{y=b}dx$$

$$-\left[\left[2D(1-\alpha)w_{xy}\,\delta w\right]_{x=0}^{x=a}\right]_{y=0}^{y=b}=0. \qquad (223)$$

* See Reference 6.

It is to be noticed that (223) and (222a) are equivalent for *any* constant α. In the physical problem under consideration, α is identifiable with the physical constant known as *Poisson's ratio*. Its value is between zero and one-half, and is dependent upon the plate material.

If the conditions of edge support are such that no deflection or rotation of the edges is permitted, the plate is said to be *clamped*. In this case w is prescribed as zero along the complete boundary, while $\partial w/\partial x$ must be zero on each boundary $x =$ constant and $\partial w/\partial y$ must vanish on the boundaries $y = 0$ and $y = b$. In view of the fact that the corresponding variations are to vanish when these quantities are prescribed, it follows that the partially integrated terms in (223) vanish, and the variational problem takes the form

$$\delta \int_0^a \int_0^b \left\{ \frac{1}{2} D[w_{xx}^2 + w_{yy}^2 + 2\alpha\, w_{xx} w_{yy} + 2(1-\alpha) w_{xy}^2] - p\, w - \frac{1}{2} \rho\, \omega^2 w^2 \right\} dx\, dy = 0. \quad (224)$$

That part of the integrand which involves D is known as the *strain energy* per unit area. The term $-p\,w$ again represents additional potential energy per unit area due to the transverse loading, and the term $\frac{1}{2}\rho\,\omega^2 w^2$ represents the kinetic energy per unit area, each of these quantities being evaluated at a position of maximum deflection.

The *natural boundary conditions* of the problem are obtained by equating to zero the integrands of the single (line) integrals, which are evaluated along the boundary. Thus, at the boundaries $x = 0$ and $x = a$, one must have either

$$w \text{ prescribed} \quad \text{or} \quad D \frac{\partial \nabla^2 w}{\partial x} + (1-\alpha) D\, w_{xyy} = 0 \quad (225a)$$

and

$$w_x \text{ prescribed} \quad \text{or} \quad D(w_{xx} + \alpha\, w_{yy}) = 0. \quad (225b)$$

The natural boundary conditions at the boundaries $y = 0$ and $y = b$ are obtained by interchanging x and y in this statement.

In the theory of elasticity it is shown that the quantity $-D\,\partial \nabla^2 w/\partial x$ is to be interpreted as the transverse *shearing force*

(Q_x) at a boundary x = constant, the quantity $(1 - \alpha)D\, w_{xy}$ as the corresponding *twisting moment* (M_{xy}), and the quantity $-D(w_{xx} + \alpha\, w_{yy})$ as the corresponding *bending moment* (M_{xx}). From (223) it follows that the effective transverse edge force associated with a deflection δw along an edge x = constant must be of the form $R_x = Q_x - \partial M_{xy}/\partial y$. The discovery of this fact, by *physical* reasoning, constituted a significant advance in the theory of small deflections of elastic plates. The presence of the last expression in (223), which involves values of M_{xy} at the four *corners* of the plate, corresponds to the possible presence of *concentrated* reactions at the corners.

The variations of appropriate line integrals, obtained by reference to (223), must be added to the left-hand member of (224) when edge deflections and/or rotations are not prescribed, but edge forces and/or moments are given.

It should be noticed that the case of *static* loading is contained in the above discussion when $\omega = 0$.

The present section is intended to illustrate two important facts. First, it has been shown (in a fairly complicated physical problem) that mere knowledge of the governing *differential equation* can lead to information as to which mathematical quantities should be *prescribed at the boundary*, and hence which *mathematical* quantities *must* be of principal *physical* interest.

Second, it is seen that, once the differential equation *and* the relevant boundary conditions are known, the corresponding variational problem (if one exists) can be obtained *without specialized knowledge of the physical details of the problem involved*. On the other hand, a sufficiently general knowledge of the *physical* background of the problem would permit one to *write down* the variational problem and, if it were desirable, *derive the relevant differential equation from it*.

In the remaining sections of this chapter, it is indicated that the variational formulation of a problem is often particularly well adapted to numerical procedures for obtaining an approximate solution.

2.17. The Ritz method. The so-called *Ritz method* is a procedure for obtaining approximate solutions of problems expressed in variational form. In the case when a function $y(x)$ is to be determined, the procedure consists essentially in assuming that the

desired extremal of a given problem can be approximated by a linear combination of n suitably chosen functions, in the form

$$y \approx c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_n\phi_n(x), \tag{226}$$

where the c's are constants to be determined. Usually the functions $\phi_k(x)$ are to be so chosen that this expression satisfies the specified boundary conditions for any choice of the c's. Thus, if y is to *vanish* at the ends of the interval under consideration, we require that *each* of the ϕ's satisfy the same condition. Otherwise, the choice of the functions ϕ_k is to a large extent arbitrary. In physical problems, the general nature of the desired solution is usually known, and a set of ϕ's is chosen in such a way that *some* linear combination of them may be expected to satisfactorily approximate the solution.

The quantity I to be made stationary is then expressed as a function of the c's, and the c's are so determined that the resultant expression is stationary. Thus in place of attempting to determine that function which renders I stationary with reference to *all* admissible slightly modified functions, we consider only the family of functions of type (226), and determine that member of the family for which I is stationary with reference to slightly modified functions belonging to the family. It is clear that the efficiency of the procedure depends upon the choice of appropriate approximating functions ϕ_k.

A more elaborate procedure consists in obtaining a *sequence* of approximations, in which the first assumption is merely $c_1\phi_1$, the second $c_1\phi_1 + c_2\phi_2$, and so forth, the nth assumption being of the form (226). The relevant c's are determined *at each stage* of the process by the method outlined above. By comparing successive approximations, an estimate of the degree of accuracy attained at any stage of the calculation can be obtained. In order that this process converge as $n \to \infty$, the functions $\phi_1, \phi_2, \ldots, \phi_n, \ldots$ should comprise an infinite set of functions such that the unknown function $y(x)$ can with certainty be approximated to any specified degree of accuracy by *some* linear combination of the ϕ's. Frequently it is convenient to choose as the nth approximation a *polynomial* of degree n, satisfying appropriate end conditions. In certain cases the use of special functions such as sine or cosine

harmonics, Legendre polynomials, and so forth, may afford computational advantages.

To illustrate this procedure in a simple case, we consider the problem of determining small static deflections of a string fixed at its ends ($x = 0, L$) and subject to a uniformly distributed load of intensity q. We assume that the tension F in the string can be considered as constant. With $\omega = 0$ and $p = -q$, the variational problem (198) becomes

$$\delta \int_0^L \left(\frac{1}{2} F y'^2 + q y\right) dx = 0, \tag{227}$$

the integrated terms vanishing in virtue of the end conditions

$$y(0) = y(L) = 0, \tag{228}$$

which require that δy must vanish at the end points. The functions

$$\phi_k(x) = \sin \frac{k\pi x}{L} \qquad (k = 1, 2, 3, \cdots),$$

which satisfy (228), are convenient admissible coordinate functions.

If, for simplicity, we assume a three-term expansion of the form

$$y \approx c_1 \sin \frac{\pi x}{L} + c_2 \sin \frac{2\pi x}{L} + c_3 \sin \frac{3\pi x}{L}, \tag{229}$$

the result of replacing y by its approximation in (227) is of the form

$$\delta \int_0^L \left[\frac{F}{2}\frac{\pi^2}{L^2}\left(c_1 \cos \frac{\pi x}{L} + 2c_2 \cos \frac{2\pi x}{L} + 3c_3 \cos \frac{3\pi x}{L}\right)^2 + q\left(c_1 \sin \frac{\pi x}{L} + c_2 \sin \frac{2\pi x}{L} + c_3 \sin \frac{3\pi x}{L}\right)\right] dx = 0. \tag{230}$$

The integrations can be carried out explicitly. Thus, making use of the *orthogonality* of the harmonics,* we obtain

$$\delta \left[\frac{F}{2}\frac{\pi^2}{L^2}\frac{L}{2}(c_1^2 + 4c_2^2 + 9c_3^2) + q\frac{L}{\pi}\left(2c_1 + 0c_2 + \frac{2}{3}c_3\right)\right] = 0. \tag{231}$$

* See Section 1.29.

Noticing that here the c's are the quantities to be varied, we next write (231) in the form

$$\frac{F}{2}\frac{\pi^2}{L}\left[\left(c_1+\frac{4qL^2}{\pi^3F}\right)\delta c_1+4c_2\,\delta c_2+\left(9c_3+\frac{4qL^2}{3\pi^3F}\right)\delta c_3\right]=0. \quad (232)$$

But since the δc's are arbitrary, their coefficients in (232) must vanish, giving the evaluations

$$c_1=-\frac{4qL^2}{\pi^3F},\quad c_2=0,\quad c_3=-\frac{4qL^2}{27\pi^3F}. \quad (233)$$

The "best" approximation of the form (229) to the required extremal is thus of the form

$$y\approx-\frac{4qL^2}{\pi^3F}\left(\sin\frac{\pi x}{L}+\frac{1}{27}\sin\frac{3\pi x}{L}\right). \quad (234)$$

The exact solution of this particular problem is readily found by elementary methods, in the form

$$y=-\frac{q}{2F}x(L-x), \quad (235)$$

and it can be verified that (234) comprises the leading terms of the Fourier sine-series expansion of (235) over the interval $(0, L)$, and that (234) does indeed afford a good approximation to (235) over that interval.

It is useful to notice not only that the Euler equation of (227) is the governing differential equation

$$F\,y''-q=0, \quad (236)$$

but also that if y satisfies the natural boundary conditions of (227) then (227) *is*, in virtue of the equivalence of (197) and (198), *equivalent to the equation*

$$\int_0^L(F\,y''-q)\,\delta y\,dx=0. \quad (237)$$

With the approximation of (229), this condition becomes

$$\int_0^L\left[F\frac{\pi^2}{L^2}\left(c_1\sin\frac{\pi x}{L}+4c_2\sin\frac{2\pi x}{L}+9c_3\sin\frac{3\pi x}{L}+q\right)\right]\cdot\left[\delta c_1\sin\frac{\pi x}{L}+\delta c_2\sin\frac{2\pi x}{L}+\delta c_3\sin\frac{3\pi x}{L}\right]dx=0. \quad (238)$$

If the integrations are carried out, equation (232) is obtained directly. This last procedure is equivalent to calculating the variation of (230) *before* carrying out the integration, and it frequently involves a reduced amount of calculation.

While the procedure of forming (237) directly from (236) is a convenient one, before it is employed in other cases one should make certain that the differential equation involved *is* indeed the Euler equation of *some* variational problem, $\delta I = 0$, whose *natural boundary conditions* include those which govern the problem at hand.

For example, the equation

$$(x^2y')' + x\,y = x \qquad (0 \leqq x \leqq L)$$

is readily transformed, by the methods of the preceding sections, to the variational problem

$$\delta \int_0^L \left[-\frac{1}{2}\,x^2(y')^2 + \frac{1}{2}\,x\,y^2 - x\,y \right] dx + \Big[x^2\,y'\,\delta y \Big]_0^L = 0.$$

If the specified boundary conditions are such that

$$\Big[x^2y'\,\delta y \Big]_0^L = 0,$$

the variational problem can thus be taken as

$$\frac{1}{2}\,\delta \int_0^L [x^2(y')^2 - x\,y^2 + 2x\,y]\,dx = 0$$

or, after calculating the variation,

$$\int_0^L [(x^2y')' + x\,y - x]\,\delta y\,dx = 0.$$

The basic equation, when expanded, becomes

$$x^2y'' + 2x\,y' + x\,y = x.$$

While this equation is equivalent to the equation

$$x\,y'' + 2y' + y = 1,$$

this last form cannot be transformed to a proper variational problem, $\delta I = 0$, merely by multiplication by δy and subsequent integration by parts. The correct multiplicative factor is seen to be

$x\,\delta y$. Thus, the form

$$\int_0^L (x\,y'' + 2y' + y - 1)\,\delta y\,dx = 0$$

is *not* the consequence of a proper variational problem and, as was seen above, the "weighting function" x should properly be introduced in the integrand, in order to ensure convergence of a *sequence* of approximations, and hence to increase the probability that a good approximation will be afforded by a given finite number of terms.

In the case of a differential equation of order greater than two, it may happen that no such weighting function *exists*. However, it is readily verified that the abbreviated procedure is valid (*when appropriate boundary conditions are prescribed*) if the governing equation is of the precise form

$$L\,y \equiv (p\,y')' + q\,y = f \qquad (x_1 \leqq x \leqq x_2) \tag{239a}$$

or

$$L\,y \equiv (s\,y'')'' + (p\,y')' + q\,y = f \qquad (x_1 \leqq x \leqq x_2), \tag{239b}$$

where p, q, and s are functions of x or constants. That is, such an equation *is* the Euler equation of a proper variational problem $\delta I = 0$, which is equivalent to the condition

$$\int_{x_1}^{x_2} (L\,y - f)\,\delta y\,dx = 0 \tag{240}$$

when $y(x)$ satisfies the appropriate natural boundary conditions. Any linear second-order equation can be written in the form (239a), by suitably defining $p(x)$ and $q(x)$. While *not all* equations of the fourth order can be reduced to (239b), the reduction *is* possible in most cases which arise in practice.

It may be noticed that, if the "unnatural" condition $y'(0) = \alpha$ and the natural condition $y(L) = 0$ were imposed on the solution of (236), the variational problem (198) would take the form

$$\delta\left[\int_0^L \left(\frac{1}{2}F\,y'^2 + q\,y\right)dx + F\,\alpha\,y(0)\right] = 0,$$

in place of (227). By integrating by parts, this condition can be transformed into the problem

$$\int_0^L (F\,y'' - q)\,\delta y\,dx + F[y'(0) - \alpha]\,\delta y(0) = 0,$$

in place of (237). Here the approximating series (226) must satisfy the condition $y(L) = 0$ for all values of the c's, but it need not satisfy the condition $y'(0) = \alpha$ identically. However, if it does not do so, care should be taken that the latter condition *can* be satisfied for *some* choice of the c's.

As a second illustration of the Ritz method, we consider the solution of the boundary-value problem for which $y(x)$ must satisfy the differential equation

$$\frac{d^2y}{dx^2} + x\,y = -x \tag{241a}$$

and the homogeneous conditions

$$y(0) = 0, \qquad y(1) = 0. \tag{241b}$$

Since (241a) is in the form of (239a), and the end conditions correspond to the requirement that δy vanish at the end points, the variational problem corresponding to (241a,b) can be expressed immediately in the reduced form

$$\int_0^1 (y'' + x\,y + x)\,\delta y\,dx = 0. \tag{242}$$

An appropriate assumption corresponding to (226) and satisfying (241b) is of the form

$$y = x(1 - x)(c_1 + c_2x + c_3x^2 + \cdots), \tag{243}$$

according to which (242) takes the form

$$\int_0^1 [(-2 + x^2 - x^3)c_1 + (2 - 6x + x^3 - x^4)c_2 + \cdots + x] \cdot [\delta c_1\,(x - x^2) + \delta c_2\,(x^2 - x^3) + \cdots]\,dx = 0.$$

The result of carrying out the indicated integrations is then of the form

$$(-\tfrac{8}{60}c_1 - \tfrac{11}{70}c_2 + \cdots + \tfrac{1}{12})\,\delta c_1 + (-\tfrac{11}{70}c_1 - \tfrac{107}{840}c_2 + \cdots - \tfrac{1}{20})\,\delta c_2 + \cdots = 0. \tag{244}$$

If only a one-term approximation is assumed,

$$y^{(1)} = c_1x(1 - x), \tag{245}$$

we have $c_2 = c_3 = \ldots = 0$, and hence also $\delta c_2 = \delta c_3 = \ldots = 0$, and (244) reduces to the condition

$$(-\tfrac{1}{6}\tfrac{8}{0}c_1 + \tfrac{1}{1}\tfrac{}{2})\,\delta c_1 = 0.$$

In virtue of the arbitrariness of δc_1, we must then have

$$c_1 = \tfrac{5}{1}\tfrac{}{9}$$

and hence the "best" solution of form (245) is given by

$$y^{(1)} = 0.263x(1 - x). \tag{246}$$

Similarly, for a two-term approximation of the form

$$y^{(2)} = c_1x(1 - x) + c_2x^2(1 - x), \tag{247}$$

the vanishing of the coefficients of the arbitrary variations δc_1 and δc_2 in (244) leads to the simultaneous equations

$$\left.\begin{aligned} 0.317c_1 + 0.157c_2 &= 0.0833, \\ 0.157c_1 + 0.127c_2 &= 0.0500 \end{aligned}\right\}, \tag{248}$$

if only three significant figures are retained in the calculations. From these equations we obtain the numerical results

$$c_1 = 0.177, \qquad c_2 = 0.173,$$

so that the "best" solution of form (247) is given by

$$y^{(2)} = (0.177x + 0.173x^2)(1 - x). \tag{249}$$

In dealing similarly with the characteristic-value problem consisting of the equation

$$\frac{d^2y}{dx^2} + \lambda\, x\, y = 0, \tag{250}$$

in place of (241a), and the boundary conditions of (241b), the assumption of (243) is found to lead to the equation

$$\left[\left(-\frac{1}{3} + \frac{\lambda}{60}\right)c_1 + \left(-\frac{1}{6} + \frac{\lambda}{105}\right)c_2 + \cdots\right]\delta c_1$$
$$+\left[\left(-\frac{1}{6} + \frac{\lambda}{105}\right)c_1 + \left(-\frac{2}{15} + \frac{\lambda}{168}\right)c_2 + \cdots\right]\delta c_2 + \cdots = 0.$$

Corresponding to a one-term approximation (245), we obtain the condition

$$\left(-\frac{1}{3}+\frac{\lambda}{60}\right)c_1 = 0,$$

since δc_1 is arbitrary, and hence obtain a nontrivial approximate solution only if

$$\lambda = \lambda_1^{(1)} = 20. \tag{251}$$

The coefficient c_1 is then arbitrary.

Corresponding to a two-term approximation (247), we obtain the two conditions

$$\left.\begin{aligned}(0.333 - 0.0167\lambda)c_1 + (0.167 - 0.00955\lambda)c_2 &= 0,\\ (0.167 - 0.00955\lambda)c_1 + (0.133 - 0.00595\lambda)c_2 &= 0\end{aligned}\right\}. \tag{252}$$

Hence a nontrivial approximate solution can be obtained only if the determinant of the coefficients of the c's vanishes. The expansion of this determinantal equation takes the form

$$3\lambda^2 - 364\lambda + 5880 = 0, \tag{253}$$

with the two roots

$$\lambda_1^{(2)} \doteq 19.2, \qquad \lambda_2^{(1)} \doteq 102. \tag{254}$$

Thus we obtain a *second* approximation to the smallest characteristic number λ_1, and a *first* approximation to the second characteristic number λ_2. For each such value of λ, the two equations of (252) become equivalent and either can be used to express c_2 as a multiple of c_1 (with c_1 arbitrary), thus determining approximations to the corresponding characteristic functions. In more involved cases, the iterative matrix methods of Sections 1.23 to 25 are useful.

The true characteristic functions of the problem are arbitrary multiples of the functions

$$f_n(x) = x^{\frac{1}{2}}J_{\frac{1}{5}}(\tfrac{2}{5}\lambda_n{}^{\frac{1}{2}}x^{\frac{5}{4}})$$

where λ_n is the nth solution of the equation

$$J_{\frac{1}{5}}(\tfrac{2}{5}\lambda^{\frac{1}{2}}) = 0,$$

from which it is found that $\lambda_1 \doteq 18.9$ and $\lambda_2 \doteq 81.8$. As is indicated by this example, the accurate calculation of the *higher* characteristic numbers by iterative methods may involve considerable labor.

As a third example, we consider the calculation of the smallest critical frequency of a vibrating membrane in the form of an isosceles right triangle. We choose dimensionless rectangular coordinates in such a way that the vertices are at the origin and at the points (1, 0) and (0, 1) (Figure 2.10). If it is assumed that the tension in the membrane is (approximately) uniform, (210) gives the differential equation satisfied by the amplitude function w in the form

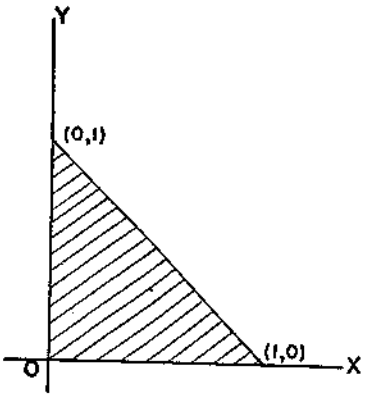


FIGURE 2.10

$$\nabla^2 w + \lambda w = 0, \tag{255}$$

where, if the legs of the triangle are of length a,

$$\lambda = \frac{\rho \omega^2 a^2}{F}. \tag{256}$$

If we require that the membrane be fixed along its boundary, the boundary condition is

$$w = 0 \text{ on the boundary.} \tag{257}$$

From the results of Section 2.14, the associated variational problem can be expressed in the form

$$\delta \iint_A \frac{1}{2} [(\nabla w)^2 - \lambda w^2] \, dx \, dy = 0 \tag{258a}$$

or, equivalently, in the reduced form

$$\iint_A (\nabla^2 w + \lambda w) \, \delta w \, dx \, dy = 0. \tag{258b}$$

Since the equation of the boundary can be written in the form

$$x y(x + y - 1) = 0,$$

appropriate approximating functions satisfying (257) are of the form

$$w = x y(x + y - 1)(c_1 + c_2 x + c_3 y + c_4 x^2 + \cdots). \tag{259}$$

For simplicity we here consider only a one-term approximation:

$$w^{(1)} = c_1 x\, y(x + y - 1). \tag{260}$$

When w is replaced by $w^{(1)}$ in (258b), there follows

$$c_1\, \delta c_1 \int_0^1 \int_0^{1-y} [2(x + y) + \lambda(x^2 y + x\, y^2 - x\, y)] \cdot (x^2 y + x\, y^2 - x\, y)\, dx\, dy = 0.$$

Since δc_1 is arbitrary, and $c_1 = 0$ leads to a trivial solution, the double integral must vanish. The integrations are readily carried out, if use is made of the known formula

$$\int_0^1 y^m (1 - y)^n\, dy = \frac{m!\, n!}{(m + n + 1)!}, \tag{261}$$

and there follows

$$-\frac{8}{6!} + \frac{\lambda}{7!} = 0 \quad \text{or} \quad \lambda = 56. \tag{262}$$

Thus a first approximation to the smallest characteristic value of λ is obtained. From (256), it follows that the smallest critical frequency is $\omega_1/2\pi$, where

$$\omega_1 \approx 7.48 \left(\frac{F}{\rho\, a^2}\right)^{1/2}. \tag{263}$$

The true value of the numerical factor in (263) is known to be $\pi\sqrt{5} \doteq 7.03$.

2.18. A semidirect method. The procedures described in the preceding section are often known as the *direct methods* in the calculus of variations. The approximating functions are completely specified at the start, and only the *constants of combination* are determined by variational methods. In the present section a modified procedure of frequent usefulness is outlined.

To fix ideas, suppose that the function w to be determined in a variational problem depends upon two independent variables x and y, and that the region of determination is the rectangle $(-a \leqq x \leqq a,\ -b \leqq y \leqq b)$. In the Ritz method we would assume an approximation in the form

$$w \approx c_1 \phi_1(x, y) + \cdots + c_n \phi_n(x, y) \tag{264}$$

where the ϕ's are completely specified functions, satisfying appropriate boundary conditions, and the c's are to be determined by variational methods. However, in many physical problems, while the general nature of the behavior of w in, say, the x-direction may be known, it may happen that the behavior in the y-direction is less predictable. In such cases, it is convenient to only *partially* specify the approximating functions by writing instead

$$w \approx \phi_1(x)f_1(y) + \cdots + \phi_n(x)f_n(y), \tag{265}$$

where the ϕ's are suitably chosen functions of x alone, satisfying appropriate conditions on the boundaries $x =$ constant, and the f's are unspecified functions of y alone, to be determined by a variational method. A procedure of this type, which may be called a *semidirect* method, then leads to a set of *ordinary differential equations* involving the unknown f's, together with a proper number of corresponding end conditions.

To illustrate the procedure, we consider small deflections of a square membrane ($|x| \leqq a$, $|y| \leqq a$), fixed along the edges $x =$ constant and along the edge $y = -a$, but unrestrained along the edge $y = a$, and subject to a distribution of static loading given by

$$p = -q(a^2 - x^2) \tag{266}$$

where q is a constant. If the tension F in the membrane is again assumed to be constant, the governing differential equation (211) is of the form

$$F\,\nabla^2 w = q(a^2 - x^2), \tag{267}$$

and the associated variational problem (212) becomes

$$-\delta \int_{-a}^{a}\int_{-a}^{a} \left[\frac{1}{2}F\,(\nabla w)^2 + q(a^2 - x^2)w\right] dx\,dy$$
$$+ \int_{-a}^{a} \Big[F\,w_x\,\delta w\Big]_{x=-a}^{x=a} dy + \int_{-a}^{a} \Big[F\,w_y\,\delta w\Big]_{y=-a}^{y=a} dx = 0. \tag{268}$$

The vanishing of the first line integral is assured if w satisfies the prescribed conditions

$$w(-a, y) = 0, \qquad w(a, y) = 0. \tag{269}$$

Since w is *not* prescribed along the edge $y = a$, the requirement that the second line integral vanish shows that in order that no

work be done by constraints along these edges we must have

$$w(x, -a) = 0, \qquad w_y(x, a) = 0, \tag{270}$$

in accordance with the approximations upon which the present formulation is based. Having thus formulated the boundary conditions appropriate to the variational problem, we are at liberty to use either (268) or the more convenient reduced form

$$\int_{-a}^{a}\int_{-a}^{a} [F \nabla^2 w - q(a^2 - x^2)] \, \delta w \, dx \, dy = 0. \tag{271}$$

In view of the nature of the loading (266), it may be suspected that the deflection $w(x, y)$ will be approximately parabolic in the x-direction, with the maximum deflection (along the x-axis) varying as a function of y. Thus a simple approximation of the form

$$w = (a^2 - x^2)f(y) \tag{272}$$

may be assumed. This approximation satisfies (269), regardless of the form of $f(y)$; in order that it satisfy (270) for all values of x, $f(y)$ must satisfy the end conditions

$$f(-a) = 0, \qquad f'(a) = 0. \tag{273}$$

If w is replaced by its approximation, there follows

$$\nabla^2 w = (a^2 - x^2)f''(y) - 2f(y)$$

and

$$\delta w = (a^2 - x^2) \, \delta f(y),$$

and hence (271) takes the form

$$\int_{-a}^{a} \left\{ \int_{-a}^{a} [F(a^2 - x^2)f''(y) - 2F f(y) - q(a^2 - x^2)] \cdot (a^2 - x^2) \, dx \right\} \delta f(y) \, dy = 0. \tag{274}$$

The integrations *with respect to* x can now be carried out explicitly, to express (274) in the form

$$\int_{-a}^{a} \left\{ F \left[\frac{16}{15} a^5 f''(y) - \frac{8}{3} a^3 f(y) \right] - \frac{16}{15} a^5 q \right\} \delta f(y) \, dy = 0. \tag{275}$$

From the arbitrariness of $\delta f(y)$ inside the interval $(-a, a)$, it follows that the quantity in braces must vanish, so that $f(y)$ must satisfy the differential equation

$$f''(y) - \frac{5}{2a^2} f(y) = \frac{q}{F}. \tag{276}$$

The solution of (276), satisfying (273), is found to be

$$f(y) = -\frac{2a^2q}{5F}\left[1 - \frac{\cosh \frac{1}{2}\sqrt{10}(1 - y/a)}{\cosh \sqrt{10}}\right] \tag{277}$$

and the introduction of (277) into (272) leads to the determination of the desired one-term approximation to the deflection.

REFERENCES

1. Bliss, G. A.: *Calculus of Variations*, University of Chicago Press, Chicago, 1925.
2. Bolza, O.: *Lectures on the Calculus of Variations*, G. E. Stechert and Company, New York, 1931.
3. Temple, G., and W. G. Bickley: *Rayleigh's Principle*, Oxford University Press, New York, 1933.
4. Webster, A. G.: *Dynamics*, B. G. Teubner, Leipzig, 1925.
5. Sokolnikoff, I. S.: *Mathematical Theory of Elasticity*, McGraw-Hill Book Company, Inc., New York, 1946 (Chapter 5).
6. Timoshenko, S.: *Theory of Plates and Shells*, McGraw-Hill Book Company, Inc., New York, 1940.

PROBLEMS

Section 2.1.

1. Suppose that $f(x, y)$ is stationary when $x = a$ and $y = b$ [so that $\partial f/\partial x = \partial f/\partial y = 0$ at (a, b)], and that $f(x, y)$ can be expanded in power series near (a, b).

(a) Show that the relevant power series then takes the form

$$f(x, y) - f(a, b) = \tfrac{1}{2}[(x - a)^2 f_{xx}(a, b) + 2(x - a)(y - b)f_{xy}(a, b) + (y - b)^2 f_{yy}(a, b)] + \cdots,$$

where-omitted terms are of degree greater than two in $(x - a)$ and $(y - b)$.

(b) Deduce that if $\partial f/\partial x = \partial f/\partial y = 0$ at (a, b), then $f(x, y)$ possesses a relative *minimum* at (a, b) if the matrix

$$\mathbf{M} \equiv \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}$$

is positive definite when $x = a$ and $y = b$, and that it possesses a relative *maximum* if $\mathbf{M}$ is negative definite at that point. [See Section 1.17.]

(c) Use the results of Section 1.18 to show that the stationary value is a maximum if $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b), and is a minimum if $f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b).

(d) Generalize the criterion of part (b) in the case of a function of n independent variables $x_1, x_2, \ldots, x_n$.

2. Of all rectangular parallelepipeds which have sides parallel to the coordinate planes, and which are inscribed in the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

determine the dimensions of that one which has the largest possible volume.

3. Determine the lengths of the principal semiaxes of the ellipse $Ax^2 + 2Bxy + Cy^2 = 1$, where $AC > B^2$, and deduce also that the area of the ellipse is given by $\pi/\sqrt{AC - B^2}$.

4. Of all parabolas which pass through the points (0, 0) and (1, 1), determine that one which, when rotated about the x-axis, generates a solid of revolution with least possible volume between $x = 0$ and $x = 1$. [Notice that the equation may be taken in the form $y = x + c\,x(1 - x)$, where c is to be determined.]

5. (a) If $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is a real vector, and $\mathbf{a}$ is a real symmetric square matrix of order n, show that the requirement that

$$F \equiv \mathbf{x}^T\,\mathbf{a}\,\mathbf{x} - \lambda\,\mathbf{x}^T\,\mathbf{x}$$

be stationary, for a prescribed $\mathbf{a}$, takes the form

$$\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}.$$

Deduce that the requirement that the quadratic form $A \equiv \mathbf{x}^T\,\mathbf{a}\,\mathbf{x}$ be stationary, subject to the constraint $B \equiv \mathbf{x}^T\,\mathbf{x} = \text{constant}$, leads to the requirement $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}$, where λ is a constant to be determined. [Notice that the same is true of the requirement that B be stationary, subject to the constraint $A = \text{constant}$, with a suitable redefinition of λ. (See also page 123 of the text.)]

(b) Show that, if we write

$$\lambda = \frac{\mathbf{x}^T\,\mathbf{a}\,\mathbf{x}}{\mathbf{x}^T\,\mathbf{x}} \equiv \frac{A}{B},$$

the requirement that λ be stationary leads again to the matrix equation $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}$. [Notice that the requirement $d\lambda = 0$ can be written in the form $(B\,dA - A\,dB)/B^2 = 0$ or $(dA - \lambda\,dB)/B = 0$.] Deduce that stationary values of the ratio $(\mathbf{x}^T\,\mathbf{a}\,\mathbf{x})/(\mathbf{x}^T\,\mathbf{x})$ are characteristic numbers of the symmetric matrix $\mathbf{a}$. (See also Problem 77 of Chapter 1.)

Section 2.2.

6. Establish the equivalence of equations (17b) and (17c).

7. It is required to determine the continuously differentiable function $y(x)$ which minimizes the integral $I = \int_0^1 (1 + y'^2)\, dx$, and satisfies the end conditions $y(0) = 0$, $y(1) = 1$.

(a) Obtain the relevant Euler equation, and show that the extremal is $y = x$.

(b) With $y(x) = x$, and the special choice $\eta(x) = x(1 - x)$, and with the notation of equation (12), calculate $I(\epsilon)$ and verify directly that $dI(\epsilon)/d\epsilon = 0$ when $\epsilon = 0$.

(c) By writing $y(x) = x + u(x)$, show that the problem becomes

$$I \equiv 2 + \int_0^1 u'^2\, dx = \text{minimum},$$

where $u(0) = u(1) = 0$, and deduce that $y(x) = x$ is indeed the required minimizing function.

8. Obtain the Euler equation and the associated natural boundary conditions, relevant to the determination of extremals of the integral $\int_0^1 F(x, y, y')\, dx$, in the following cases:

(a) $F = y'^2 + y\,y' + y^2$, (b) $F = x\,y'^2 - y\,y' + y$,

(c) $F = y'^2 + k^2 \cos y$, (d) $F = a(x)y'^2 - b(x)y^2$.

Section 2.3.

A *geodesic* on a given surface is a curve, lying on that surface, along which distance between two points is as small as possible. On a plane, a geodesic is a straight line. Determine equations of geodesics on the following surfaces:

9. Right circular cylinder. [Take $ds^2 = a^2\, d\theta^2 + dz^2$ and minimize $\int \sqrt{a^2 + (dz/d\theta)^2}\, d\theta$ or $\int \sqrt{a^2 (d\theta/dz)^2 + 1}\, dz$.]

10. Right circular cone. [Use spherical coordinates (Figure 2.5) with $ds^2 = dr^2 + r^2 \sin^2 \alpha\, d\theta^2$.]

11. Sphere. [Use spherical coordinates (Figure 2.5) with $ds^2 = a^2 \sin^2 \phi\, d\theta^2 + a^2\, d\phi^2$.]

12. Surface of revolution. [Write $x = r \cos \theta$, $y = r \sin \theta$, $z = f(r)$. Express the desired relation between r and θ in terms of an integral.]

Section 2.4.

13. If $I(y) = \int_0^1 \sqrt{1 + y'^2}\, dx$, calculate $I(x)$ and $I(\cosh x)$.

14. If $F = 1 + x + y + y'^2$, calculate the following quantities for $x = 0$:

(a) dF where $y = \sin x$ and $dx = \epsilon$.

(b) δF where $y = \sin x$ and $\delta y = \epsilon(x + 1)$.

15. If $I = \int_0^1 (x^2 - y^2 + y'^2)\, dx$, calculate both ΔI and δI when $y = x$ and $\delta y = \epsilon x^2$.

16. Let $y = 1 + x^2$, where x and y are functions of an independent variable t. Calculate $\delta \dfrac{dy}{dx}$ and $\dfrac{d}{dx}\delta y$ when $x = \dfrac{1}{t}$ and $\delta x = \epsilon t^2$, and verify the validity of equation (32a) in this case.

Section 2.5.

17. Derive the Euler equation of the problem

$$\delta \int_{x_1}^{x_2} F(x, y, y', y'')\, dx = 0$$

in the form

$$\frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y''}\right) - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) + \frac{\partial F}{\partial y} = 0,$$

and show that the associated natural boundary conditions are

$$\left[\left(\frac{d}{dx}\frac{\partial F}{\partial y''} - \frac{\partial F}{\partial y'}\right)\delta y\right]_{x_1}^{x_2} = 0 \quad \text{and} \quad \left[\frac{\partial F}{\partial y''}\,\delta y'\right]_{x_1}^{x_2} = 0.$$

18. Specialize the results of Problem 17 in the case of the problem

$$\delta \int_{x_1}^{x_2} [a(x)y''^2 - b(x)y'^2 + c(x)y^2]\, dx = 0.$$

19. Derive the Euler equation of the problem

$$\delta \iint_R F(x, y, u, u_x, u_y)\, dx\, dy = 0$$

in the form

$$\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) - \frac{\partial F}{\partial u} = 0,$$

subject to the requirement that $u(x, y)$ is prescribed along the closed boundary C of the region R.

20. Obtain the natural boundary condition relevant to Problem 19, in the form

$$\oint_C \left[\frac{\partial F}{\partial u_x}\cos \nu + \frac{\partial F}{\partial u_y}\sin \nu\right]\delta u\, ds = 0,$$

where s is arc length along C in the positive (counterclockwise) direction, and ν is the angle from the positive x-axis to the outward normal at a point of C. [Notice that $\iint_A \frac{\partial \phi}{\partial x}\,dx\,dy = \oint_C \phi \cos \nu\,ds$ and $\iint_A \frac{\partial \phi}{\partial y}\,dx\,dy = \oint_C \phi \sin \nu\,ds$.]

21. Specialize the results of Problems 19 and 20 in the case of the problem

$$\delta \iint_R [a(x, y)u_x^2 + b(x, y)u_y^2 - c(x, y)u^2]\,dx\,dy = 0.$$

In particular, show that if $b(x, y) = a(x, y)$ the natural boundary condition takes the form

$$\oint_C \left(a \frac{\partial u}{\partial n}\,\delta u\right) ds = 0,$$

where $\partial u/\partial n$ is the normal derivative of u on C.

22. Derive the Euler equation of the problem

$$\delta \int_{x_1}^{x_2} \int_{y_1}^{y_2} F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy})\,dx\,dy = 0,$$

where x_1, x_2, y_1, and y_2 are constants, in the form

$$\frac{\partial^2}{\partial x^2}\left(\frac{\partial F}{\partial u_{xx}}\right) + \frac{\partial^2}{\partial x\,\partial y}\left(\frac{\partial F}{\partial u_{xy}}\right) + \frac{\partial^2}{\partial y^2}\left(\frac{\partial F}{\partial u_{yy}}\right) - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) + \frac{\partial F}{\partial u} = 0,$$

and show that the associated natural boundary conditions are then

$$\left[\left(\frac{\partial}{\partial x}\frac{\partial F}{\partial u_{xx}} + \frac{\partial}{\partial y}\frac{\partial F}{\partial u_{xy}} - \frac{\partial F}{\partial u_x}\right)\delta u\right]_{x_1}^{x_2} = 0, \quad \left[\frac{\partial F}{\partial u_{xx}}\,\delta u_x\right]_{x_1}^{x_2} = 0,$$

and

$$\left[\left(\frac{\partial}{\partial y}\frac{\partial F}{\partial u_{yy}} + \frac{\partial}{\partial x}\frac{\partial F}{\partial u_{xy}} - \frac{\partial F}{\partial u_y}\right)\delta u\right]_{y_1}^{y_2} = 0, \quad \left[\frac{\partial F}{\partial u_{yy}}\,\delta u_y\right]_{y_1}^{y_2} = 0.$$

23. Specialize the results of Problem 22 in the case of the problem

$$\delta \int_{x_1}^{x_2} \int_{y_1}^{y_2} \left[\frac{1}{2}u_{xx}^2 + \frac{1}{2}u_{yy}^2 + \alpha\,u_{xx}u_{yy} + (1 - \alpha)u_{xy}^2\right] dx\,dy = 0,$$

where α is a constant. [Show that the Euler equation is of the form $\nabla^4 u = 0$, regardless of the value of α, whereas the natural boundary conditions are dependent upon α.]

Section 2.6.

24. A particle moves on the surface $\phi(x, y, z) = 0$ from the point (x_1, y_1, z_1) to the point (x_2, y_2, z_2) in the time T. Show that if it moves in such a way that the integral of its kinetic energy over that time is a minimum, its coordinates must also satisfy the equations $\ddot{x}/\phi_x = \ddot{y}/\phi_y = \ddot{z}/\phi_z$.

[Minimize $\displaystyle\int_0^T \frac{1}{2}(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)\,dt$, subject to the constraint $\phi = 0$.]

25. Specialize Problem 24 in the case when the particle moves on the unit sphere $x^2 + y^2 + z^2 - 1 = 0$, from (0, 0, 1) to (0, 0, −1), in time T. [Show first that the motion must be described by the equations $r \equiv \sqrt{x^2 + y^2} = \sin\dfrac{n\pi t}{T}$, $z = \cos\dfrac{n\pi t}{T}$, $\theta \equiv \tan^{-1}\dfrac{y}{x} = \text{const.}$, where n is an odd integer, so that motion is along a great circle of the sphere. Then show that the integrated kinetic energy is least when $n = 1$, and is then given by $\pi^2/(2T)$.]

26. Determine the equation of the shortest arc which passes through the points (0, 0) and (1, 0) and encloses a prescribed area A with the x-axis. [Reduce the problem of determining the arbitrary constants to the solution of a transcendental equation.]

27. (a) Show that the extremals of the problem

$$\delta\int_{x_1}^{x_2} [p(x)y'^2 - q(x)y^2]\,dx = 0, \qquad \int_{x_1}^{x_2} r(x)y^2\,dx = 1,$$

where $y(x_1)$ and $y(x_2)$ are prescribed, are solutions of the equation

$$\frac{d}{dx}\left(p\frac{dy}{dx}\right) + (q + \lambda r)y = 0,$$

where λ is a constant.

(b) Show that the natural boundary conditions are of the form

$$\left[p\frac{dy}{dx}\,\delta y\right]_{x_1}^{x_2} = 0,$$

so that the same result follows if $p\,y'$ is required to vanish at an end point where y is not prescribed.

28. Specialize Problem 27 in the following special case:

$$\delta\int_0^\pi y'^2\,dx = 0, \qquad \int_0^\pi y^2\,dx = 1;$$

$$y(0) = 0, \qquad y(\pi) = 0.$$

[Show that the extremals are of the form $y = \sqrt{2/\pi}\sin nx$, where n is an integer other than zero.]

29. Show that, if the constraint $\int_0^\pi y^2\,dx = 1$ is omitted in Problem 28, the only extremal is the trivial one $y \equiv 0$.

30. (a) Show that the extremals of the problem

$$\delta \int_{x_1}^{x_2} [s(x)y''^2 - p(x)y'^2 + q(x)y^2]\,dx = 0,$$

$$\int_{x_1}^{x_2} r(x)y^2\,dx = 1,$$

where $y(x_1)$, $y'(x_1)$, $y(x_2)$, and $y'(x_2)$ are prescribed, are solutions of the equation

$$\frac{d^2}{dx^2}\left(s\frac{d^2y}{dx^2}\right) + \frac{d}{dx}\left(p\frac{dy}{dx}\right) + (q - \lambda r)y = 0,$$

where λ is a constant.

(b) By considering the relevant natural boundary conditions, show that the same result follows if $(s\,y'')' + p\,y'$ is required to vanish at an end point where y is not prescribed, and $s\,y''$ is required to vanish at an end point where y' is not prescribed.

31. Specialize Problem 30 in the following special case:

$$\delta \int_0^\pi y''^2\,dx = 0, \qquad \int_0^\pi y^2\,dx = 1;$$

$$y(0) = y''(0) = 0, \qquad y(\pi) = y''(\pi) = 0.$$

[Show that the end conditions are appropriate, and that the extremals are of the form $y = \sqrt{2/\pi}\sin nx$, where n is an integer other than zero.]

32. Show that, if the constraint $\int_0^\pi y^2\,dx = 1$ is omitted in Problem 31, the only extremal is the trivial one $y \equiv 0$.

Section 2.7.

33. Verify that the Euler equation relevant to the problem $\delta\lambda = 0$, where

$$\lambda = \frac{\int_{x_1}^{x_2} (s\,y''^2 - p\,y'^2 + q\,y^2)\,dx}{\int_{x_1}^{x_2} r\,y^2\,dx}$$

is of the form

$$\frac{d^2}{dx^2}\left(s\frac{d^2y}{dx^2}\right) + \frac{d}{dx}\left(p\frac{dy}{dx}\right) + (q - \lambda r)y = 0,$$

and that the relevant natural boundary conditions at $x = x_1$ and $x = x_2$ are the following:

$(s\,y'')' + p\,y' = 0$ or y prescribed *and* $s\,y'' = 0$ or y' prescribed.

[Compare Problem 30.] Deduce that, when *homogeneous* natural boundary conditions are prescribed, stationary values of the ratio λ are characteristic values of the associated boundary-value problem.

34. The deflection y of a beam executing small free vibrations of frequency ω satisfies the differential equation

$$\frac{d^2}{dx^2}\left(E\,I\,\frac{d^2y}{dx^2}\right) - \rho\,\omega^2 y = 0,$$

where $E\,I$ is the flexural rigidity and ρ the linear mass density. Deduce from Problem 33 that the deflection modes are extremals of the problem

$$\delta\omega^2 \equiv \delta\left[\frac{\int_0^L E\,I\,y''^2\,dx}{\int_0^L \rho\,y^2\,dx}\right] = 0,$$

when appropriate homogeneous end conditions are prescribed, and where L is the length of the beam, and that stationary values of the ratio are squares of the natural frequencies. [The *bending moment* M is given (approximately) by $M = E\,I\,y''$, and the transverse *shearing force* S by $S = M' = (E\,I\,y'')'$. Notice that the natural boundary conditions are satisfied if either $S = 0$ or y is prescribed *and* either $M = 0$ or y' is prescribed at each end of the beam. It can be shown that the *smallest* stationary value of ω^2 is truly the *minimum* value of the ratio.]

35. Suppose that the tension F and linear density ρ of a freely vibrating string of length L are *nearly uniform*, and that the string is fixed at the ends $x = 0$ and $x = L$. Recalling that the natural vibration modes for a *uniform* string are multiples of the functions

$$y_n(x) = \sin\left(\frac{n\pi x}{L}\right) \qquad (n = 1, 2, \cdots),$$

motivate the approximate formula

$$\omega_n \approx \frac{n\,\pi}{L}\sqrt{\frac{\int_0^L F\cos^2(n\pi x/L)\,dx}{\int_0^L \rho\sin^2(n\pi x/L)\,dx}} \qquad (n = 1, 2, \cdots),$$

for the nth natural frequency, in the case under consideration. If the small deviations in F and ρ from uniformity are assumed to be *linear*, show also that the approximate values of the natural frequencies take the form

$$\omega_n \approx \frac{n\,\pi}{L}\sqrt{\frac{\bar{F}}{\bar{\rho}}} \qquad (n = 1, 2, \cdots),$$

where $\bar{F}$ and $\bar{\rho}$ are the *mean values* of F and ρ.

36. Obtain formulas analogous to those of Problem 35, in the case of the freely vibrating beam of Problem 34, both ends of which are *hinged* in such a way that both y and M vanish.

37. Let ω_i^2 represent the ith characteristic value of ω^2 for the problem consisting of the equation $(F\,y')' + \rho\,\omega^2 y = 0$ and of specific end conditions which require that at each end of the interval $(0, L)$ either y or $F\,y'$ vanishes, and denote the corresponding characteristic function by $\phi_i(x)$. Suppose also that the ϕ's are *normalized* in such a way that

$$\int_0^L \rho\,\phi_i\phi_j\,dx = \delta_{ij}$$

(see Section 1.29), and that the ω^2's are arranged in increasing order of magnitude.

(a) Show that

$$\int_0^L F\,\phi_i'^2\,dx = -\int_0^L (F\,\phi_i')'\,\phi_i\,dx = \omega_i^2.$$

(b) By making use of the fact that any continuously differentiable function $y(x)$ which satisfies the prescribed end conditions can be expressed in the form

$$y(x) = \sum_{k=1}^{\infty} c_k\phi_k(x) \qquad (0 \leqq x \leqq L),$$

where the series converges uniformly, and by taking into account the orthogonality of the ϕ's relative to ρ, show that the relation

$$\omega^2 = \frac{\int_0^L F\,y'^2\,dx}{\int_0^L \rho\,y^2\,dx}$$

takes the form

$$\omega^2 = \frac{\sum_{k=1}^{\infty} c_k^2\omega_k^2}{\sum_{k=1}^{\infty} c_k^2}.$$

(c) Show that $\omega^2 - \omega_1^2 \geqq 0$, and that $\omega^2 = \omega_1^2$ when $y(x) = \phi_1(x)$. Hence deduce that *the smallest characteristic value of* ω^2 *is the minimum value of the ratio in* (79) *for all admissible functions.*

(d) Show that, if $c_1 = c_2 = \cdots = c_{r-1} = 0$, there follows $\omega^2 - \omega_r^2 \geqq 0$, and that $\omega^2 = \omega_r^2$ when $y(x) = \phi_r(x)$. Hence deduce that *the* r*th characteristic value of* ω^2 *is the minimum value of the ratio in* (79) *for all admissible functions which are orthogonal to the first* $r - 1$ *characteristic functions.* (Compare Problem 77 of Chapter 1.)

Section 2.8.

38. A particle of mass m is falling vertically, under the action of gravity. If x is distance measured downward, and no resistive forces are present, show that the Lagrangian function is

$$L = T - V = m(\tfrac{1}{2}\dot{x}^2 + g\,x) + \text{constant},$$

and verify that the Euler equation of the problem $\delta \int_{t_1}^{t_2} L\,dt = 0$ is the proper equation of motion of the particle.

39. A particle of mass m is moving vertically, under the action of gravity and a resistive force numerically equal to k times the displacement x from an equilibrium position. Show that the equation of Hamilton's principle is of the form

$$\delta \int_{t_1}^{t_2} \left(\frac{1}{2}\,m\,\dot{x}^2 + m\,g\,x - \frac{1}{2}\,k\,x^2\right) dt = 0,$$

and obtain the Euler equation.

40. A particle of mass m is falling vertically under the action of gravity, and its motion is resisted by a force numerically equal to a constant c times its velocity $\dot{x}$. Show that the equation of Hamilton's principle takes the form

$$\delta \int_{t_1}^{t_2} \left(\frac{1}{2}\,m\,\dot{x}^2 + m\,g\,x\right) dt - \int_{t_1}^{t_2} c\,\dot{x}\,\delta x\,dt = 0.$$

41. Three masses are connected in series to a fixed support, by linear springs. Assuming that only the spring forces are present, and using the

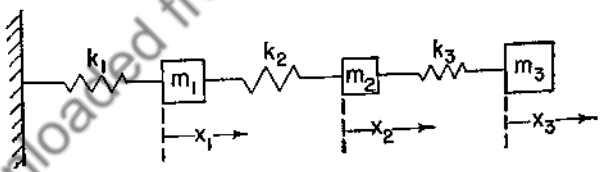


FIGURE 2.11

notation of Figure 2.11, show that the Lagrangian function of the system is

$$L = \tfrac{1}{2}[m_1\dot{x}_1^2 + m_2\dot{x}_2^2 + m_3\dot{x}_3^2 - k_1x_1^2 - k_2(x_2 - x_1)^2 - k_3(x_3 - x_2)^2] + \text{const.},$$

where the x_i represent displacements from equilibrium. [Notice that if the x_i are given increments δx_i the total work done *by* the springs is given by

$$\delta\Phi = -\delta V = [k_2(x_2 - x_1) - k_1x_1]\,\delta x_1 + [k_3(x_3 - x_2) - k_2(x_2 - x_1)]\,\delta x_2 + [-k_3(x_3 - x_2)]\,\delta x_3.]$$

Section 2.9.

42. Obtain the Lagrange equations relevant to the mechanical system of Problem 41.

43. A mass $4m$ is attached to a string which passes over a smooth pulley. The other end of the string is attached to a smooth pulley of mass m, over which passes a second string attached to masses m and $2m$. If the system starts from rest, determine the motion of the mass $4m$, using the coordinates q_1 and q_2 indicated in Figure 2.12.

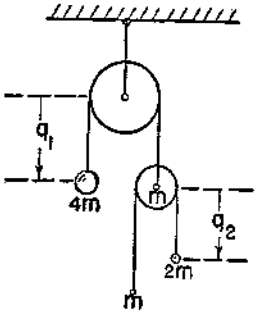


FIGURE 2.12

44. Obtain the Lagrangian equations for a triple pendulum consisting of three weights of equal mass m, connected in series to a fixed support by inextensible strings of equal length a, taking as the coordinates the angles θ_1, θ_2, and θ_3 made with the vertical by the three strings. Show also that, for small deviations from equilibrium, and small velocities, the Lagrangian function takes the approximate form

$$L = \frac{m a^2}{2}(3\dot{\theta}_1^2 + 2\dot{\theta}_2^2 + \dot{\theta}_3^2 + 4\dot{\theta}_1\dot{\theta}_2 + 2\dot{\theta}_2\dot{\theta}_3 + 2\dot{\theta}_1\dot{\theta}_3 - \frac{m g a}{2}(3\theta_1^2 + 2\theta_2^2 + \theta_3^2) + \text{const.}$$

45. Two particles of equal mass m are connected by an inextensible string which passes through a hole in a smooth horizontal table, the first particle resting on the table, and the second particle being suspended vertically. Initially, the first particle is caused to describe a circular path about the hole, with an angular velocity $\omega = \sqrt{g/a}$, where a is the radius of the path, so that the suspended mass is held at equilibrium. At the instant $t = 0$, the suspended mass is pulled downward a short distance and is released, while the first mass continues to rotate.

(a) If x represents the distance of the second mass below its equilibrium position at time t, and θ represents angular position of the first particle at time t, show that the Lagrangian function is given by

$$L = m[\dot{x}^2 + \tfrac{1}{2}(a - x)^2\dot{\theta}^2 + g\,x] + \text{const.},$$

and obtain the equations of motion.

(b) Show that the first integral of the θ-equation is of the form $(a - x)^2\dot{\theta} = a\sqrt{a\,g}$, and that the result of eliminating $\dot{\theta}$ between this equation and the x-equation becomes

$$2\ddot{x} + \left[\frac{1}{(1 - x/a)^3} - 1\right] g = 0.$$

(c) In the case when the displacement of the suspended mass from equilibrium is small, show that the suspended mass performs small vertical oscillations of period $2\pi\sqrt{2a/3g}$.

Section 2.10.

46. (a) In terms of Lagrange's function $L(q_1, \ldots, q_n; \dot{q}_1, \ldots, \dot{q}_n)$, such that $L = T - V$, show that the equations of motion become

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i} \qquad (i = 1, 2, \cdots, n).$$

(b) Show that the generalized momentum p_i corresponding to the ith coordinate q_i is given by

$$p_i \equiv \frac{\partial T}{\partial \dot{q}_i} = \frac{\partial L}{\partial \dot{q}_i}.$$

47. By noticing that T is a homogeneous quadratic form in the n $\dot{q}$'s, establish the identity

$$\sum_{k=1}^{n} p_k \dot{q}_k = 2T.$$

[Compare Problem 32 of Chapter 1.]

48. The *Hamiltonian function* H, of a conservative system, is defined as the *sum* of the kinetic and potential energies: $H = T + V$.

(a) By making use of the result of Problem 47, show that one may write

$$H = \sum_{k=1}^{n} p_k \dot{q}_k - L.$$

(b) Let the $3n$ variables $p_1, \ldots, p_n; q_1, \ldots, q_n; \dot{q}_1, \ldots, \dot{q}_n$ be considered independent. By using the result of Problem 46(b), show that

$$\frac{\partial H}{\partial \dot{q}_i} = 0,$$

so that H *is a function only of the* p's *and* q's, and is independent of the $\dot{q}$'s.

49. (a) By noticing that L is a function only of the q's and $\dot{q}$'s, use the result of Problem 48(a) to show that

$$\frac{\partial H}{\partial p_i} = \dot{q}_i \qquad (i = 1, 2, \cdots, n).$$

(b) By combining the results of Problem 48(a) and 46(a), show that the equations of motion can be written in the form

$$\frac{\partial H}{\partial q_i} = -\dot{p}_i \qquad (i = 1, 2, \cdots, n).$$

[The two sets of equations obtained in parts (a) and (b) are known as *Hamilton's canonical equations*.]

(c) By multiplying the ith equation of part (a) by $\dot{p}_i$, the ith equation of part (b) by $\dot{q}_i$, adding, and summing the results over i, deduce the equation of *conservation of total energy*, $dH/dt = 0$.

50. (a) For the simple pendulum of Figure 2.3, show that the Hamiltonian function (expressed in terms of coordinates and momenta) is of the form

$$H = \frac{p^2}{2m L^2} - m g L \cos \theta + \text{const.},$$

where p is the generalized momentum associated with the generalized coordinate $q = \theta$.

(b) Obtain Hamilton's canonical equations in the form

$$\frac{p}{m L^2} = \dot{\theta}, \qquad m g L \sin \theta = -\dot{p},$$

and show that they imply equation (99).

51. For a harmonic oscillator with one degree of freedom, show that the Hamiltonian function is of the form

$$H = \frac{p^2}{2m} + \frac{k q^2}{2} + \text{const.},$$

where k is the stiffness constant of the system. Show also that the canonical equations take the form $\dot{p} = -k q$ and $\dot{q} = p/m$.

52. A mass m moves in the xy-plane, under the action of a central force directed along the radius from the origin. If the position of the mass is specified by the polar coordinates r and θ, and the potential energy function is denoted by $V(r)$, express the Hamiltonian function in terms of r, θ, p_r, and p_θ, and obtain the four relevant canonical equations.

Section 2.11.

53. Solve the problem of the simple pendulum by taking as *two* coordinates of the mass m the distance r from the support to the mass and the angle θ of Figure 2.3, subject to the constraint $r = L$, and making use of a Lagrange multiplier. [Show that the equations corresponding to (129) become

$$m(\ddot{r} - r \dot{\theta}^2 - g \cos \theta) = \lambda,$$

$$m(r^2 \ddot{\theta} + 2r \dot{r} \dot{\theta} + g r \sin \theta) = 0.$$

By introducing the relation $r = L$, obtain equation (99) and deduce further that the tension S in the string is given by $S = -\lambda = m g \cos \theta + m L \dot{\theta}^2$.]

54. Suppose that the oscillations of a simple pendulum are not restricted to a plane. By appropriately introducing the spherical coordinates of

Figure 2.5, obtain the equations of motion in the form

$$\ddot{\phi} - \dot{\theta}^2 \sin \phi \cos \phi + \frac{g}{L} \sin \phi = 0, \qquad \dot{\theta} \sin^2 \phi = C,$$

and show also that the tension in the string is given by

$$S = m g \cos \phi + m L(\dot{\phi}^2 + \dot{\theta}^2 \sin^2 \phi).$$

Sections 2.12, 2.13.

55. *Potential energy of a linear spring.* Suppose that the force exerted by a spring is directed along the spring, and is proportional to its stretch e beyond its "natural length" L_0.

(a) Prove that the potential energy V_s stored in the spring is given by

$$V_s = \frac{k}{2} e^2 + \text{const.},$$

where k is the "spring constant" of proportionality. [Calculate the work done in stretching the spring from the length L_0 to the length $L_0 + e$.]

(b) Suppose that an unstretched spring of length L_0 coincides with the vector $a\,\mathbf{i} + b\,\mathbf{j} + c\,\mathbf{k}$, where $a^2 + b^2 + c^2 = {L_0}^2$, and that the subsequent displacement of one end relative to the other is defined by the vector $u\,\mathbf{i} + v\,\mathbf{j} + w\,\mathbf{k}$. Show that the potential energy is of the form

$$V_s = \frac{k}{2} [\sqrt{(a+u)^2 + (b+v)^2 + (c+w)^2} - L_0]^2 + \text{const.}$$

(c) Under the assumption of small displacements, obtain the expansion

$$V_s = \frac{k}{2} \left(\frac{a\,u + b\,v + c\,w}{L_0} \right)^2 + \cdots + \text{const.}$$

$$= \frac{k}{2} (l\,u + m\,v + n\,w)^2 + \cdots + \text{const.},$$

where omitted terms involve powers of u, v, and w greater than two, and l, m, and n are the direction cosines of the line of action of the spring in its natural position, so that *the contents of the parentheses comprise the component of the relative displacement vector in the direction of the natural position of the spring.*

56. A mass m is elastically restrained in space by a number of springs with spring constants k_i, which are attached to the points P_i. The mass is at equilibrium at the origin O, the springs then being of natural length (Figure 2.13). If the direction cosines of the radii OP_i are denoted by

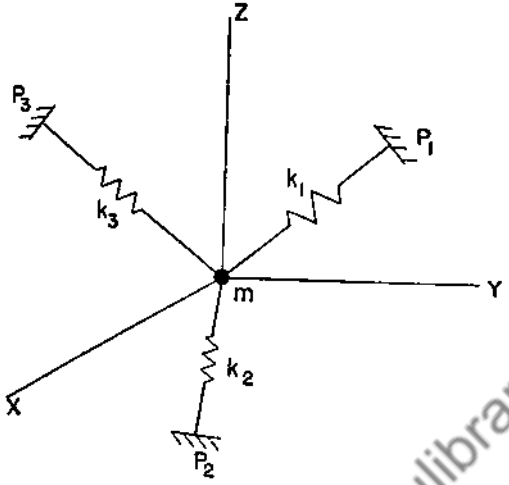


FIGURE 2.13

(l_i, m_i, n_i), and small displacements are assumed, obtain the potential energy stored in the springs when the mass is displaced to the position (x, y, z) in the form

$$V_s = \sum_i \frac{k_i}{2} (l_i x + m_i y + n_i z)^2.$$

[Use the result of Problem 55(c).] Also, obtain the equations of motion in the absence of external forces.

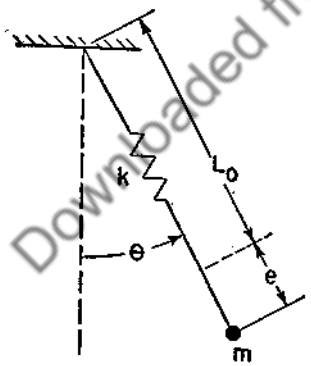


FIGURE 2.14

57. Suppose that a pendulum, vibrating in a plane, consists of a mass m attached to a fixed support by a linear spring with spring constant k (Figure 2.14).

(a) Show that the potential energy is given by

$$V = \frac{k}{2} e^2 - m\,g(L_0 + e) \cos\theta + \text{const.},$$

where L_0 is the natural length of the spring and e is its stretch.

(b) Show that the position of equilibrium is specified by $e = m\,g/k$, $\theta = 0$. With the introduction of the new coordinate $s = e - m\,g/k$, such that s is the stretch beyond the length $L = L_0 + m\,g/k$ assumed by the loaded spring in equilibrium under the action of gravity, obtain the relevant energy

functions in the form

$$T = \frac{m}{2}[\dot{s}^2 + (L+s)^2\dot{\theta}^2],$$

$$V = \frac{k}{2}\left(\frac{m\,g}{k} + s\right)^2 - m\,g(L+s)\cos\theta + \text{const.},$$

and deduce the equations of motion in the form

$$m\,\ddot{s} + k\,s - m(L+s)\dot{\theta}^2 + m\,g(1-\cos\theta) = 0,$$

$$m\frac{d}{dt}[(L+s)^2\dot{\theta}] + m\,g(L+s)\sin\theta = 0.$$

(c) Assuming small stretch and deflection, obtain the approximations

$$T = \frac{m}{2}(\dot{s}^2 + L^2\dot{\theta}^2), \qquad V = \frac{k}{2}s^2 + \frac{1}{2}m\,g\,L\,\theta^2 + \text{const.},$$

and deduce that in the linear theory the extensional and deflectional vibration modes are uncoupled, with frequencies $\sqrt{k/m}$ and $\sqrt{g/L}$, respectively, so that s and θ are normal coordinates.

58. In Problem 57, use as coordinates the components x and y of the displacement of the mass m from equilibrium position, in the horizontal and vertical directions, respectively. Show that there follows

$$T = \frac{m}{2}(\dot{x}^2 + \dot{y}^2),$$

$$V = \frac{k}{2}[\sqrt{x^2 + (y-L)^2} - L_0]^2 + m\,g\,y + \text{const.},$$

where $L_0 = L - \dfrac{m\,g}{k}$; obtain the expansion $V = \dfrac{k}{2}y^2 + \dfrac{m\,g}{2L}x^2 + \ldots$ + const., relevant to small oscillations; and compare the corresponding linearized equations of motion with the results of Problem 57(c). [Notice that the results of Problem 55(c) are not applicable here, since x and y are not measured from a position corresponding to zero stretch.]

59. The point of suspension of a simple pendulum is completely restrained from vertical motion, and is partially restrained from horizontal motion by a spring system which exerts a restoring force equal to $-k\,x$ when the horizontal displacement of that point is x (Figure 2.15). Obtain the equations of motion of the suspended mass m, assuming the string to be inextensible and of length L. Show that for small displacements there follows approximately $x = \dfrac{m\,g}{k}\theta$ and $\left(L + \dfrac{m\,g}{k}\right)\ddot{\theta} + g\,\theta = 0$, so that the

system is then equivalent to a simple pendulum of length $L + \frac{mg}{k}$ with a fixed support.

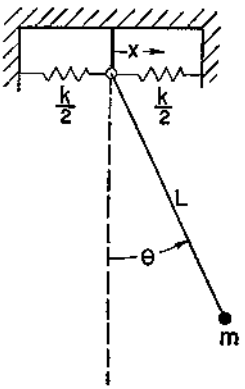


FIGURE 2.15

60. A mass m is attached to three symmetrically placed supports by linear springs. With the notation of Figure 2.16, the mass is at equilibrium at the origin, equidistant from the three supports (and in their plane), the springs then being unstretched.

(a) Assuming small oscillations, show that the potential energy stored in the springs, corresponding to a displacement (x, y) in the plane of the supports, is of the form

$$V = \tfrac{1}{8}[(4k_1 + k_2 + k_3)x^2 + 2\sqrt{3}\,(k_2 - k_3)xy + 3(k_2 + k_3)y^2],$$

and obtain the corresponding equations of motion of the mass.

(b) In the special case when $k_1 = 2k$, $k_2 = (2 + \sqrt{3})k$, and $k_3 = (2 - \sqrt{3})k$, determine the natural frequencies and the natural modes

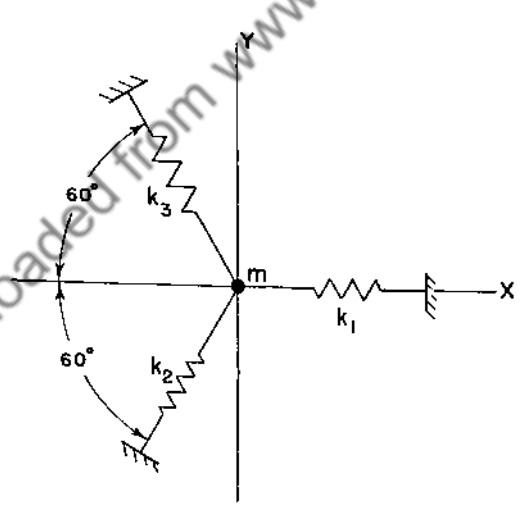


FIGURE 2.16

of small oscillations. Show also that the coordinates $\alpha_1 = (x + y)/\sqrt{2}$, $\alpha_2 = (x - y)/\sqrt{2}$ are then normal coordinates, and express the kinetic and potential energies of the system in terms of them.

61. A mass m under the action of gravity executes small oscillations near the origin on a frictionless paraboloid

$$z = \tfrac{1}{2}(Ax^2 + 2Bxy + Cy^2),$$

where $B^2 < AC$, $A > 0$, and where the z-axis is directed upward. Obtain the characteristic equation determining the natural frequencies. [Use x and y as the Lagrangian coordinates.]

62. A mass m under the action of gravity executes small oscillations near the origin on a frictionless ellipsoid

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} + \frac{(z - C)^2}{C^2} = 1,$$

where the z-axis is directed upward. Show that the coordinates x and y are normal coordinates, and that the natural frequencies are $\sqrt{g\,C}/A$ and $\frac{\sqrt{g\,C}}{B}$. [Show that, near the origin, there follows $z = \frac{C}{2}\left(\frac{x^2}{A^2} + \frac{y^2}{B^2}\right)$ + terms of higher order in x and y.]

63. From the analogy between coupled mechanical systems and coupled electric networks, in which linear displacement x corresponds to charge $Q = \int_{t_0}^{t} I\,dt$, where I is current, and where mass m corresponds to inductance L, spring constant k to reciprocal capacity $1/C$, damping coefficient r to resistance R, and impressed force F to impressed voltage E, deduce that to the potential energy V there must correspond the "electromagnetic energy"

$$V = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{C_i}Q_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{1}{C_{ij}}(Q_i - Q_j)^2 - \sum_{i=1}^{n}E_iQ_i,$$

where $\dot{Q}_i = I_i$ is the current flowing in the ith circuit, C_i is the capacitance of that circuit which is not in common with other circuits, $C_{ij} = C_{ji}$ is mutual capacitance in common with the ith and jth circuits, and E_i is the impressed voltage (positive in the positive direction of I_i) in the ith circuit. Show also that to the kinetic energy T there must correspond the "magnetic energy"

$$T' = \frac{1}{2}\sum_{i=1}^{n}L_i\dot{Q}_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}L_{ij}(\dot{Q}_i - \dot{Q}_j)^2,$$

where L_i and $L_{ij} = L_{ji}$ are coefficients of self-inductance and mutual inductance, respectively. Finally, show that to the Rayleigh dissipation function there must correspond the "heat dissipation function"

$$F = \frac{1}{2}\sum_{i=1}^{n}R_i\dot{Q}_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}R_{ij}(\dot{Q}_i - \dot{Q}_j)^2,$$

where R_i and $R_{ij} = R_{ji}$ are the resistances, after which the circuit equations are obtained in the Lagrangian form

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{Q}_i}\right) + \frac{\partial V}{\partial Q_i} + \frac{\partial F}{\partial \dot{Q}_i} = 0 \qquad (i = 1, 2, \cdots, n).$$

64. Derive the circuit equations relevant to the two networks of Figure 2.17 by the Lagrangian method of Problem 63, and verify the results by use of Kirchhoff's laws. Also, investigate the natural frequencies of

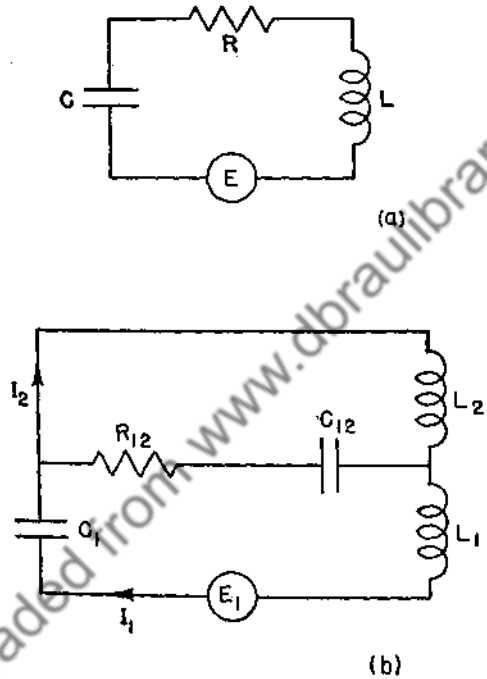


FIGURE 2.17

small oscillating currents in each of the two networks when the resistances are neglected. [In the second case, merely express the characteristic equation in terms of the vanishing of a determinant.]

Section 2.14.

65. Suppose that $y(x)$ satisfies the differential equation

$$(s\,y'')'' + (p\,y')' + q\,y = f$$

everywhere in the interval (x_1, x_2) except at an interior point ξ, and that one or more of the functions s, p, q, and f may be defined by different analytic expressions over the two subintervals (x_1, ξ) and (ξ, x_2).

(a) If y and y' are required to be continuous at $x = \xi$, obtain the relation

$$\delta \int_{x_1}^{x_2} \left(\frac{1}{2} s\, y''^2 - \frac{1}{2} p\, y'^2 + \frac{1}{2} q\, y^2 - f\, y \right) dx + \Big[\{(s\, y'')' + p\, y'\}\, \delta y - (s\, y'')\, \delta y' \Big]_{x_1}^{x_2} - \Big[(s\, y'')' + p\, y' \Big]_{\xi-}^{\xi+} \delta y(\xi) + \Big[s\, y'' \Big]_{\xi-}^{\xi+} \delta y'(\xi) = 0.$$

(b) Deduce the *natural transition conditions*

$$y(\xi+) = y(\xi-), \qquad y'(\xi+) = y'(\xi-),$$

$$[(s\, y'')]_{\xi+} = [s\, y'']_{\xi-}, \qquad [(s\, y'')' + p\, y']_{\xi+} = [(s\, y'')' + p\, y']_{\xi-}.$$

(c) Suppose that the conditions $y = y_1$ and $y' = y_1'$ are prescribed at $x = x_1$, and that the conditions $(s\, y'')' + p\, y' = S_2$ and $s\, y'' = M_2$ are prescribed at $x = x_2$. Further, suppose that it is required that $(s\, y'')' + p\, y'$ possess a *jump* of A, and $s\, y''$ a jump of B as the point $x = \xi$ is crossed in the positive direction. Show that the variational problem takes the form

$$\delta \left[\int_{x_1}^{x_2} \left(\frac{1}{2} s\, y''^2 - \frac{1}{2} p\, y'^2 + \frac{1}{2} q\, y^2 - f\, y \right) dx + S_2 y(x_2) - M_2 y'(x_2) - A\, y(\xi) + B\, y'(\xi) \right] = 0,$$

where admissible functions are to satisfy the conditions $y(x_1) = y_1$ and $y'(x_1) = y_1'$, and are to be continuously differentiable in (x_1, x_2).

66. Specialize the results of Problem 65 in the case of the equation

$$\frac{d^2}{dx^2} \left(E\, I\, \frac{d^2 y}{dx^2} \right) - \rho\, \omega^2 y = p,$$

which governs the steady-state amplitude of small forced vibration of a beam, and interpret the conditions in physical terms. [See the note to Problem 34.]

67. (a) Modify the treatments of Problem 65 in the case of the second-order equation

$$(p\, y')' + q\, y = f,$$

omitting the requirement that y' be continuous at $x = \xi$.

(b) Specialize the results of part (a) in the case of the equation

$$\frac{d}{dx} \left(F\, \frac{dy}{dx} \right) + \rho\, \omega^2 y = p,$$

which governs the steady-state amplitude of small forced vibration of a string, and interpret the conditions in physical terms.

68. Two unknown functions $y_1(x)$ and $y_2(x)$ are governed, over an interval (a, b), by the simultaneous equations

$$(p_{11}y_1')' + (p_{12}y_2')' + r_{11}y_1 + r_{12}y_2 = f_1,$$
$$(p_{12}y_1')' + (p_{22}y_2')' + r_{12}y_1 + r_{22}y_2 = f_2,$$

where p_{ij}, r_{ij}, and f_i are prescribed functions of x. By multiplying the first equation by a variation δy_1, the second by δy_2, adding, integrating over (a, b), and simplifying the result, show that the corresponding variational problem is of the form

$$\delta \int_a^b \left[\frac{1}{2} \left(p_{11}y_1'^2 + 2p_{12}y_1'y_2' + p_{22}y_2'^2 - r_{11}y_1^2 - 2r_{12}y_1y_2 - r_{22}y_2^2 \right) + f_1y_1 + f_2y_2 \right] dx = 0,$$

if the prescribed boundary conditions are compatible with the following ones:

$$\left[(p_{11}y_1' + p_{12}y_2')\, \delta y_1 \right]_a^b = 0, \quad \left[(p_{12}y_1' + p_{22}y_2')\, \delta y_2 \right]_a^b = 0.$$

69. Show that the equations

$$x\, y_1'' + 2y_1' + x\, y_1 - x^2y_2 = \phi_1,$$
$$x\, y_2'' + 4y_2' - y_1 + x\, y_2 = \phi_2$$

are reducible to the standard form of Problem 68, and obtain the relevant variational problem.

70. By starting with the known differential equation, or otherwise, deduce the following variational problems in the cases noted. In each case, u represents deflection at time t, and f represents the corresponding impressed force intensity.

(a) Transverse deformation of a string:

$$\delta \int_{t_1}^{t_2} \int_0^L \left[\frac{1}{2} \rho \left(\frac{\partial u}{\partial t} \right)^2 - \frac{1}{2} F \left(\frac{\partial u}{\partial x} \right)^2 + f\, u \right] dx\, dt = 0.$$

(b) Transverse deformation of a beam:

$$\delta \int_{t_1}^{t_2} \int_0^L \left[\frac{1}{2} \rho \left(\frac{\partial u}{\partial t} \right)^2 - \frac{1}{2} E\, I \left(\frac{\partial^2 u}{\partial x^2} \right)^2 + f\, u \right] dx\, dt = 0.$$

(c) Transverse deformation of a membrane:

$$\delta \int_{t_1}^{t_2} \iint_A \left[\frac{1}{2} \rho \left(\frac{\partial u}{\partial t} \right)^2 - \frac{1}{2} F\, (\nabla u)^2 + f\, u \right] dS\, dt = 0.$$

(d) Longitudinal deformation of a rod:

$$\delta \int_{t_1}^{t_2} \int_0^L \left[\frac{1}{2} \rho \left(\frac{\partial u}{\partial t} \right)^2 - \frac{1}{2} E A \left(\frac{\partial u}{\partial x} \right)^2 + f u \right] dx\, dt = 0.$$

[E is Young's modulus, A the cross-sectional area.]

Section 2.15.

71. Derive equations (215), (216), and (217) by considering the integral of each left-hand member over an appropriate interval or region, and transforming the integral by integration by parts.

72. Verify equation (219), by expanding the right-hand member or otherwise.

73. A linear partial differential equation of second order, of the form

$$F[w] \equiv a\, w_{xx} + 2b\, w_{xy} + c\, w_{yy} + d\, w_x + e\, w_y + f\, w + g = 0,$$

where the coefficients may be functions of x and y, is derivable from a variational problem $\delta \iint_R G\, dx\, dy = 0$ if and only if the left-hand member can be reduced to the left-hand member of a so-called "self-adjoint" form

$$S[w] \equiv (p\, w_x)_x + (q\, w_y)_x + (q\, w_x)_y + (r\, w_y)_y + s\, w + t = 0,$$

by multiplication by a function $A(x, y)$, which may be termed a "reducing factor."

(a) By requiring that $S[w]$ be identical with $A\, F[w]$, show that there must follow

$$p = A\, a, \quad q = A\, b, \quad r = A\, c, \quad s = A\, f, \quad t = A\, g,$$

and that the reducing factor A must then satisfy the simultaneous first-order partial differential equations

$$a \frac{\partial A}{\partial x} + b \frac{\partial A}{\partial y} = (d - a_x - b_y)A,$$

$$b \frac{\partial A}{\partial x} + c \frac{\partial A}{\partial y} = (e - b_x - c_y)A,$$

Unless these equations possess a common solution, the equation $F[w] = 0$ is not derivable from a variational problem.

(b) Suppose that a reducing factor A exists. By multiplying the equation $F[w] = 0$ by $A\, \delta w\, dx\, dy$, integrating the result over the relevant region R, and making use of equations (214) and (216), show that the variational problem is of the form

$$\delta \iint_R \left[\frac{1}{2}(a\, w_x^2 + 2b\, w_x w_y + c\, w_y^2 - f\, w^2) - g\, w \right] A\, dx\, dy = 0,$$

if appropriate boundary conditions are prescribed.

(c) Apply the preceding technique to the differential equation

$$x^2w_{xx} + 2x\,w_{xy} + x^2w_{yy} + 3x\,w_x + 2w_y + x\,w + g = 0.$$

[Show that the reducing factor must be a constant multiple of $A = x$.]

Section 2.16.

74. When the plate considered in Section 2.16 is also subjected to compressive forces N_1 and N_2, parallel to its surface and in the x- and y-directions, respectively, and to a shearing force S parallel to its surface, the approximate governing differential equation differs from equation (222) in that the zero right-hand member is replaced by $-(N_1w_{xx} + 2S\,w_{xy} + N_2w_{yy})$. Show that the integrand of (224) is then to be modified by the addition of the expression

$$-\tfrac{1}{2}(N_1w_x^2 + 2S\,w_xw_y + N_2w_y^2).$$

75. Suppose that a rectangular plate of uniform thickness is acted on only by a uniform compressive force N in the x-direction.

(a) Show that the variational problem derived in Problem 74 takes the form

$$\frac{D}{2}\,\delta \int_0^a \int_0^b [w_{xx}^2 + w_{yy}^2 + 2\alpha\,w_{xx}w_{yy} + 2(1-\alpha)w_{xy}^2]\,dx\,dy - \frac{N}{2}\,\delta \int_0^a \int_0^b w_x^2\,dx\,dy = 0.$$

(b) Deduce that the critical buckling loads (for which the problem possesses a nontrivial solution) are stationary values of the ratio

$$N = \frac{D\int_0^a \int_0^b [w_{xx}^2 + w_{yy}^2 + 2\alpha\,w_{xx}w_{yy} + 2(1-\alpha)w_{xy}^2]\,dx\,dy}{\int_0^a \int_0^b w_x^2\,dx\,dy},$$

where w satisfies the appropriate support conditions along the boundary. [Compare equation (69) and Problem 33. It can be shown that the *smallest* stationary value of N is the *minimum* value of the ratio.]

Section 2.17.

76. Use the Ritz method to obtain an approximate solution of the problem

$$\frac{d}{dx}\left(x\frac{dy}{dx}\right) + y = x, \qquad y(0) = 0, \quad y(1) = 1,$$

in the form $y \approx x + x(1-x)(c_1 + c_2x)$.

77. Use the Ritz method to find two successive approximations to the smallest characteristic value of λ in the problem

$$\frac{d}{dx}\left[(1+x)\frac{dy}{dx}\right] + \lambda y = 0, \qquad y(0) = 0, \quad y(1) = 0,$$

assuming first $y(x) \approx c_1 x(1-x)$, and second $y(x) \approx (c_1 + c_2 x)x(1-x)$.

78. Suppose that a small mass M is attached at the point $x = a$ to a vibrating string of linear mass density ρ and length L, and that $M \ll \rho L$. If the string is fixed at the ends $x = 0$ and $x = L$, show that the variational problem for small vibrations of frequency ω is of the form

$$\delta \int_0^L \left[\frac{1}{2}\rho\,\omega^2 y^2 - \frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right] dx + \delta\left[\frac{1}{2}M\,\omega^2[y(a)]^2\right] = 0$$

or, equivalently,

$$\int_0^L \left[F\frac{d^2y}{dx^2} + \rho\,\omega^2 y\right]\delta y\,dx + M\,\omega^2 y(a)\,\delta y(a) = 0,$$

where F is the tension in the string. Assuming that F and ρ are constant, and that the deflection modes differ slightly from those in which M is absent, show that the nth natural frequency is approximately given by

$$\omega_n \approx \frac{n\pi}{L}\sqrt{\frac{F}{\rho}}\left(1 - \frac{M}{\rho L}\sin^2\frac{n\pi a}{L}\right).$$

[Compare the use of the procedure of Problem 35.]

79. A uniform square plate of length a is subject to a uniformly distributed compressive load N in the x-direction, in the plane of the plate. The plate is clamped along its complete boundary ($x = 0$, $x = a$, $y = 0$, $y = a$). Show that the approximation

$$w \approx C\left(1 - \cos\frac{2\pi x}{a}\right)\left(1 - \cos\frac{2\pi y}{a}\right),$$

for the fundamental buckling mode, satisfies the relevant boundary conditions, and determine a corresponding approximation to the critical buckling load N_{cr}. [Use the result of Problem 75, or, equivalently, use the relation

$$D\int_0^a\int_0^a \nabla^4 w\,\delta w\,dx\,dy + N_{cr}\int_0^a\int_0^a w_{xx}\,\delta w\,dx\,dy = 0.$$

The required approximation is given by $N_{cr} \approx 32\pi^2 D/3a^2 \doteq 105D/a^2$, whereas the true value is known to be $103.5D/a^2$.]

30. (a) Establish the relations

$$\int_{x_1}^{x_2} p\,y'^2\,dx = -\int_{x_1}^{x_2} (p\,y')'y\,dx + \Big[p\,y'y\Big]_{x_1}^{x_2}$$

and

$$\int_{x_1}^{x_2} s\,y''^2\,dx = \int_{x_1}^{x_2} (s\,y'')''y\,dx + \Big[s\,y''y' - (s\,y'')'y\Big]_{x_1}^{x_2}.$$

(b) Use the results of part (a) to show that the expression

$$\delta\int_{x_1}^{x_2}\left(\frac{1}{2}s\,y''^2 - \frac{1}{2}p\,y'^2 + \frac{1}{2}q\,y^2 - f\,y\right)dx$$

can be written, not only in the form

$$\int_{x_1}^{x_2}[(s\,y'')'' + (p\,y')' + q\,y - f]\,\delta y\,dx + \Big[(s\,y'')\,\delta y' - \{(s\,y'')' + p\,y'\}\,\delta y\Big]_{x_1}^{x_2},$$

but also in the form

$$\frac{1}{2}\delta\left\{\int_{x_1}^{x_2}[(s\,y'')'' + (p\,y')' + q\,y - 2f]y\,dx + \Big[(s\,y'')y' - \{(s\,y'')' + p\,y'\}y\Big]_{x_1}^{x_2}\right\}.$$

31. Let R denote a region of the xy-plane, with boundary C made up of one or more closed curves, and suppose that w is to satisfy Laplace's equation in R, that w is prescribed as $\phi(s)$ along the portion C' of C, and that $\partial w/\partial n$ is prescribed as $\psi(s)$ along the remainder of the boundary C'', where s represents distance along C. By calculating the variation, verify that the problem

$$\delta\left[\frac{1}{2}\iint_R (\nabla w)^2\,dx\,dy - \int_{C'}(w-\phi)\frac{\partial w}{\partial n}\,ds - \int_{C''}\psi\,w\,ds\right] = 0$$

is equivalent to the problem

$$\iint_R \nabla^2 w\,\delta w\,dx\,dy + \int_{C'}(w-\phi)\,\delta\frac{\partial w}{\partial n}\,ds - \int_{C''}\left(\frac{\partial w}{\partial n} - \psi\right)\delta w\,ds = 0,$$

and hence deduce that the desired solution is an extremal of either formulation of this variational problem, where the admissible functions are *unrestricted* along C. [Either C' or C'' may, of course, be identified with the whole of C. Notice that, if use is made of the Ritz method, the linear combination of approximating functions need not identically satisfy the prescribed conditions along either C' or C''. (In two-dimensional problems, it is often inconvenient to choose approximating functions which have this property.) In particular, it is possible to choose, as approximating

functions, special solutions of Laplace's equation, so that the *double integral* in the *second* form vanishes identically, and then to determine the constants of combination in such a way that the sum of the line integrals vanishes.]

82. Let the symbol Y represent y_1 when $x = x_1$ and y_2 when $x = x_2$. With this notation, verify that the problem

$$\delta \left\{ \int_{x_1}^{x_2} \left(\frac{1}{2} p\, y'^2 - \frac{1}{2} q\, y^2 - f\, y \right) dx - \Big[p(y - Y)y' \Big]_{x_1}^{x_2} \right\} = 0$$

is equivalent to the problem

$$\int_{x_1}^{x_2} [(p\, y')' + q\, y - f]\, \delta y\, dx + \Big[p(y - Y)\, \delta y' \Big]_{x_1}^{x_2} = 0.$$

Hence deduce that the extremal of this problem, when the admissible functions are *unrestricted* at $x = x_1$ and $x = x_2$, is the solution of the equation $(p\, y')' + q\, y = f$ for which $y(x_1) = y_1$ and $y(x_2) = y_2$.

Section 2.18.

83. The edges $x = 0$, $x = a$, and $y = 0$ of a vibrating square membrane are fixed. Whereas the edge $y = a$ is unrestrained, the thickness of the membrane is abruptly increased at that edge. Suppose that the additional material may be considered as concentrated along the edge, with linear mass density $A\, \rho/h$, where A is the effective cross-sectional area and h is the uniform membrane thickness.

(a) Show that the relevant variational problem is of the form

$$\delta \int_0^a \int_0^a \left[\frac{1}{2} \rho\, \omega^2 w^2 - \frac{1}{2} F(w_x^2 + w_y^2) \right] dx\, dy + \delta \int_0^a \left[\frac{1}{2} \frac{A\, \rho}{h}\, \omega^2 w^2 \right]_{y=a} dx = 0,$$

and that this requirement is equivalent to the condition

$$\int_0^a \int_0^a [F \nabla^2 w + \rho\, \omega^2 w]\, \delta w\, dx\, dy - \int_0^a \left[\left(F \frac{\partial w}{\partial y} - \frac{A\, \rho}{h}\, \omega^2 w \right) \delta w \right]_{y=a} dx = 0,$$

where w is to vanish along the three fixed edges.

(b) Suppose that ρ, F, and h are considered to be constant, whereas A may vary moderately along the edge $y = a$. By assuming approximate deflection modes in the form

$$w = f_m(y) \sin \frac{m\pi x}{a} \qquad (m = 1, 2, \cdots),$$

where $f_m(0) = 0$, show that $f_m(y)$ must satisfy the differential equation

$$f_m''(y) + \left(\frac{\rho\, \omega^2}{F} - \frac{m^2\pi^2}{a^2} \right) f_m(y) = 0,$$

and the homogeneous end conditions

$$f_m(0) = 0, \qquad a f_m'(a) - \alpha_m \omega^2 f_m(a) = 0,$$

where

$$\alpha_m = \frac{2\rho}{h F} \int_0^a A(x) \sin^2 \frac{m\pi x}{a}\, dx.$$

(c) Deduce that corresponding critical frequencies are of the approximate form

$$\omega_{mn} \simeq \sqrt{\frac{F}{\rho a^2}}\, \sqrt{m^2\pi^2 + k_{mn}^2},$$

where k_{mn} is the nth solution of the equation

$$\alpha_m \tan k = \frac{\rho a^2}{F} \frac{k}{k^2 + m^2\pi^2}.$$

[The approximation can be shown to be *exact* when A is constant.]

(d) Specialize this result in the two limiting cases in which the edge $y = a$ is unstiffened ($\alpha_m = 0$) and in which it is fixed ($\alpha_m = \infty$).

84. A uniform square plate is clamped along the edges $x = 0$, $x = a$, and $y = 0$, and completely unrestrained along the edge $y = a$, and is subject to a uniform loading $p = -p_0$ normal to its surface. If an approximate deflection

$$w = x^2(a - x)^2 f(y) \equiv \phi(x) f(y)$$

is assumed, where $f(y)$ satisfies the conditions $f(0) = f'(0) = 0$ along the edge $y = 0$, use equation (223) to show that the relevant natural boundary conditions along the edge $y = a$ take the form

$$k_1 f'''(a) + (2 - \alpha) k_2 f'(a) = 0,$$

$$k_1 f''(a) + \alpha k_2 f(a) = 0,$$

where

$$k_1 = \int_0^a \phi^2\, dx = \frac{a^9}{630}, \qquad k_2 = \int_0^a \phi\, \phi''\, dx = -\frac{2a^7}{105},$$

and where α is Poisson's ratio for the plate material. Show also that the differential equation governing $f(y)$ is obtained, from the condition

$$\int_0^a \left[\int_0^a [D \nabla^4(\phi f) + p_0]\phi\, dx \right] \delta f\, dy = 0$$

for arbitrary δf, in the form

$$k_1 f^{iv}(y) + 2k_2 f''(y) + k_3 f(y) = -k_4 \frac{p_0}{D},$$

where k_1 and k_2 are as defined above, and

$$k_3 = \int_0^a \phi\, \phi^{iv}\, dx = \frac{4a^5}{5}, \qquad k_4 = \int_0^a \phi\, dx = \frac{a^5}{30}.$$

CHAPTER THREE

Difference Equations

3.1. Introduction. For a given function $f(x)$, we may in general calculate the change in the function when x is increased by a positive amount h. This change in f is called the first *forward difference* of f, relative to the increment h, and is denoted by $\Delta f(x)$:

$$\Delta f(x) = f(x + h) - f(x). \tag{1}$$

The corresponding differences of higher order are then defined by iteration, according to the formulas

$$\Delta^2 f(x) = \Delta[\Delta f(x)] = f(x + 2h) - 2f(x + h) + f(x), \tag{2a}$$

$$\Delta^3 f(x) = \Delta[\Delta^2 f(x)] = f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x), \tag{2b}$$

and so forth.

An equation which may be considered as relating differences of an unknown function is known as a *difference equation.* Thus, for example, it is readily verified that the relations

$$y(x + 2h) + A\, y(x + h) + B\, y(x) = \phi(x) \tag{3a}$$

and

$$\Delta^2 y(x) + (A + 2)\Delta y(x) + (A + B + 1)y(x) = \phi(x), \tag{3b}$$

where a certain increment h is implied in (3b), are equivalent forms of the *same* linear difference equation.

The *order* of the equation is defined to be the difference between the largest and smallest arguments involved, in units of the increment or *spacing* h, *when the equation is written in a form similar to* (3a). Thus, if $B \neq 0$, equation (3a) or (3b) is of the second order. It may be verified that the equation $\Delta^2 y(x) + 2\Delta y(x) + y(x) = \phi(x)$

is equivalent to the equation $y(x + 2h) = \phi(x)$, and hence is actually of order *zero*. If $\phi(x)$ is prescribed and $y(x)$ is to be determined, the solution of this equation is obviously $y(x) = \phi(x - 2h)$.

A *linear* difference equation is one which involves no products or nonlinear functions of the unknown function and its differences.

Suppose, for simplicity, that the coefficients A and B in (3a) are constants, and that $B \neq 0$. If equation (3a) is satisfied for a particular value of x, say $x = x_0$, we may write

$$y(x_0 + 2h) = \phi(x_0) - A\, y(x_0 + h) - B\, y(x_0).$$

Similarly, by setting $x = x_0 + h$, we obtain the result

$$y(x_0 + 3h) = \phi(x_0 + h) - A\, y(x_0 + 2h) - B\, y(x_0 + h)$$

or, making use of the preceding relation,

$$y(x_0 + 3h) = \phi(x_0 + h) - A\, \phi(x_0) - (B - A^2)y(x_0 + h) + AB\, y(x_0).$$

By repetitions of this process, it is then clear that the value of $y(x_0 + kh)$ can be obtained for any positive integer $k \geqq 2$ in terms of prescribed values of ϕ, *and in terms of the two arbitrarily assigned values* $y(x_0)$ *and* $y(x_0 + h)$.

If we write

$$x = x_0 + kh \quad \text{or} \quad k = \frac{x - x_0}{h}, \tag{4}$$

so that k is dimensionless distance from a reference point x_0, in units of the spacing h, it follows that k then takes on the integral values 0, 1, 2, . . . at the points x_0, $x_0 + h$, $x_0 + 2h$, ...

In many applications it is found that the independent variable x takes on *only* integral values, say $x = 0, 1, 2, \ldots$, in the sense that the function $y(x)$ to be determined is *defined* only for integral values of the argument. In such cases it is often convenient to replace x by the symbol k to indicate more explicitly the fact that the variable takes on only discrete values and is not a "continuous variable." In other cases, $y(x)$ may be defined for all values of x in some continuous range, but it may be that the *difference equation* governs only those values of x for which $x = x_0$, $x = x_0 + h$, . . . , $x_0 + kh$, ... By using the notation of (4), we then conveniently place these values of x into correspondence with the

integers 0, 1, . . . , k, . . . , so that again the difference equation deals only with integral arguments.

Accordingly, we introduce the abbreviations

$$f_0 \equiv f(x_0), \quad f_1 \equiv f(x_0 + h), \quad \cdots, \quad f_k \equiv f(x_0 + kh), \quad \cdots, \tag{5a}$$

and write also

$$\triangle f_k \equiv f_{k+1} - f_k, \quad \triangle^2 f_k \equiv f_{k+2} - 2f_{k+1} + f_k, \quad \cdots, \tag{5b}$$

in place of (1) and (2). A constant spacing h is assumed in each case.

With this notation, the linear difference equation (3) can be written in the form

$$y_{k+2} + A_k y_{k+1} + B_k y_k = \phi_k \tag{6a,b}$$

or

$$\triangle^2 y_k + a_k \triangle y_k + b_k y_k = \phi_k.$$

If the values ϕ_0, ϕ_1, ϕ_2, and so forth, are *known*, and if the *two initial values* y_0 and y_1 of the unknown function y are *prescribed*, it is seen that the successive values y_2, y_3, and so forth, can be determined step by step from (6a), as long as the coefficients A_k and B_k are defined (and finite) for $k = 0, 1, 2, \ldots$. More generally, in the case of a linear difference equation of order n, of the general form

$$C_k^{(n)} y_{k+n} + C_k^{(n-1)} y_{k+n-1} + \cdots + C_k^{(1)} y_{k+1} + C_k^{(0)} y_k = \phi_k, \tag{7}$$

it is clear that we may arbitrarily prescribe n initial values y_0, y_1, . . . , y_{n-1} and determine the values y_n, y_{n+1}, . . . in terms of them, if the ratios of each of the remaining coefficients in (7) to the *leading* coefficient $C_k^{(n)}$ are defined (and finite) for $k = 0, 1, 2, \ldots$. If these ratios are finite and $C_k^{(0)} \neq 0$ for $0 \leqq k \leqq K$, the general solution of the difference equation over that range involves *exactly n arbitrary constants.* The specified conditions, which serve to determine these constants, need not prescribe the n initial values directly, but may be expressed in various other ways, as will be seen.

To illustrate the solution of a difference equation, we may consider the problem of determining the solution of the equation

$$y_{k+1} - r\, y_k = 1 \qquad (k \geqq 0), \tag{8a}$$

where r is a constant, with the initial condition

$$y_0 = 1. \tag{8b}$$

By setting k successively equal to 0, 1, and 2, there follows easily $y_1 = 1 + r$, $y_2 = 1 + r + r^2$, and $y_3 = 1 + r + r^2 + r^3$. Inductive reasoning then leads, in this case, to an explicit form for y_k,

$$y_k = 1 + r + r^2 + \cdots + r^k. \tag{9}$$

While this form is *explicit*, it is not *closed*, since omitted terms are necessarily indicated by dots unless k is specified. However, since the terms in (9) form a "geometric progression," a closed form is obtained by recalling a result of elementary algebra. This result may be rederived here by noticing that (9) also satisfies the relation $y_{k+1} - y_k = r^{k+1}$, and by eliminating y_{k+1} between this relation and equation (8a) to give

$$y_k = \frac{1 - r^{k+1}}{1 - r} \qquad (r \neq 1). \tag{10}$$

When $r = 1$, it is obvious that $y_k = k + 1$.

Whether or not an explicit form, or a closed form, can be obtained for y_k in other cases, it is always possible to determine successive values of y step by step (when the coefficients of the governing equation are sufficiently well behaved). In Sections 3.4 to 3.11, we consider certain important cases in which explicit solutions *can* be obtained, whereas in the remainder of the chapter we indicate in what ways the possibility of *step-by-step* solution of *difference* equations leads to techniques for obtaining *approximate* solutions to certain problems governed by ordinary or partial *differential* equations.

Before proceeding to these matters, however, it is desirable to introduce certain operational notations and indicate other related applications, and (Section 3.3) to illustrate typical problems which are conveniently formulated in terms of difference equation.

3.2. Difference operators. In addition to the *forward difference* operator Δ, defined by

$$\Delta f_k = f_{k+1} - f_k, \tag{11a}$$

there are conventionally defined also the *backward difference* operator ∇, defined by

$$\nabla f_k = f_k - f_{k-1}, \tag{11b}$$

and the *central difference* operator δ, defined by

$$\delta f_k = f_{k+\frac{1}{2}} - f_{k-\frac{1}{2}}. \tag{11c}$$

To this list we may add the *shifting operator* E, defined by

$$E f_k = f_{k+1}, \tag{12}$$

and the *differential operator* D, defined by

$$D f_k = \left(\frac{df}{dx}\right)_{x=x_k}. \tag{13}$$

In all cases except (13), the spacing h is implied.

It is readily verified that all these operators satisfy the commutative and distributive laws of ordinary real numbers. We say that two operators are *equal* when both give the same result when applied to any function for which both operations are defined. With this understanding, it follows immediately that

$$\Delta = E - 1, \quad \nabla = 1 - E^{-1} = \frac{E-1}{E}, \quad \delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}, \tag{14a,b,c}$$

where, for example, $E^{-1} f_k = f_{k-1}$ and $E^{\frac{1}{2}} f_k = f_{k+\frac{1}{2}}$. Other useful relations are of the form

$$\Delta = E^{\frac{1}{2}}\delta = E\nabla, \quad \delta = E^{\frac{1}{2}}\nabla = E^{-\frac{1}{2}}\Delta, \quad \nabla = E^{-\frac{1}{2}}\delta = E^{-1}\Delta \tag{15,a,b,c}$$

and

$$\nabla\Delta = \Delta\nabla = \delta^2. \tag{16}$$

For any function $f(x)$ which is regular at $x = x_k$ (in particular, for any *polynomial*), the Taylor series expansion

$$f(x_k + h) = f(x_k) + \frac{h}{1!} f'(x_k) + \cdots + \frac{h^n}{n!} f^{(n)}(x_k) + \cdots$$

can then be written in the symbolic form

$$E f_k = \left[1 + \frac{h D}{1!} + \frac{(h D)^2}{2!} + \cdots + \frac{(h D)^n}{n!} + \cdots\right] f_k \tag{17}$$

or, more briefly,

$$E f_k = e^{hD} f_k.$$

Hence, we are led to the curious and useful operational relation

$$E = e^{hD}, \tag{18}$$

which is merely a symbolic way of writing (17). By making use of (14a,b,c), we may then express the operators $\triangle$, ∇, and δ in terms of the differential operator D.

In particular, since (18) implies that $h\,D = \log E$, reference to (14a) gives the formal operational relation

$$h\,D = \log(1 + \triangle) = \triangle - \tfrac{1}{2}\triangle^2 + \tfrac{1}{3}\triangle^3 - \cdots, \tag{19}$$

which is equivalent to the relation

$$\left(\frac{df}{dx}\right)_{x=x_k} = \frac{1}{h}\left(\triangle f_k - \frac{1}{2}\triangle^2 f_k + \frac{1}{3}\triangle^3 f_k - \cdots\right). \tag{20}$$

Thus we obtain a formula for the first derivative of a function at a point, in terms of its forward differences at that point, assuming appropriate convergence. Similar formulas are obtainable in terms of backward differences and central differences. Further, it is possible to generate (by analogous methods) formulas for numerical interpolation, extrapolation, and integration, by the use of differences.

In the present work, we will not be concerned with these last topics, which were mentioned here only to indicate the scope of an important phase of the *calculus of finite differences.** In so far as the operator D is concerned, it is sufficient for present purposes to notice that, from the definition of the derivative, the *quantities* $\triangle f_k/h$, $\nabla f_k/h$, *and* $\delta f_k/h$ *each tend to* $D f_k$ *as* h *tends to zero* and are, in general, available as approximations to the derivative of $f(x)$ at x_k when h is sufficiently small. Similarly, *the quantities* $\triangle^n f_k/h^n$, $\nabla^n f_k/h^n$, *and* $\delta^n f_k/h^n$ *each approximate* $d^n f/dx^n$ *at* $x = x_k$ *for small values of* h.

The operators $\triangle$, δ^2, and E are to be used principally in the remainder of this chapter, although reference will occasionally be made to the operator ∇. The relation (14a) is particularly useful, for example, in transforming one of the forms (6a,b) to the other. Thus, (6a) may be written in the form

$$(E^2 + A_k E + B_k)y_k = \phi_k,$$

* See References 1 and 2.

and may be transformed to (6b) by replacing E by $\Delta + 1$,

$$[(\Delta + 1)^2 + A_k(\Delta + 1) + B_k]y_k = \phi_k,$$

and performing the indicated algebraic operations.

3.3. Formulation of difference equations. To illustrate the occurrence of difference equations in practice, we consider first the problem of determining small deflections of a tightly stretched

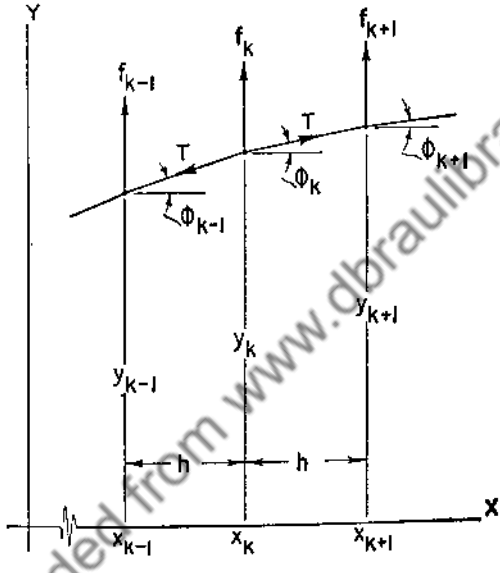


FIGURE 3.1

string, due to a number of concentrated forces f_k applied at equally spaced points x_k along the string. We assume that the string is under a large uniform tension T, and that the slope of each segment of the deflected string is small. The weight of the string is neglected. With the notation of Figure 3.1, the deflections y_k and y_{k+1} at successive points of load application differ by

$$y_{k+1} - y_k = h \tan \phi_k, \tag{21}$$

where h is the horizontal spacing. Also, for force equilibrium at (x_k, y_k) we must have the relation

$$T(\sin \phi_k - \sin \phi_{k-1}) + f_k = 0. \tag{22}$$

For small slope angles, we may identify the tangent and sine to a first approximation. Hence the introduction of (21) into (22) leads to the equation

$$\frac{T}{h}[(y_{k+1} - y_k) - (y_k - y_{k-1})] + f_k = 0$$

or

$$y_{k+1} - 2y_k + y_{k-1} = -\frac{h}{T} f_k. \tag{23}$$

This relation is a linear difference equation of second order, with constant coefficients, for the determination of the deflections at the points of force application. We may consider k as representing dimensionless distance along the horizontal axis, in units of the physical spacing h. If there are N masses, so that the length of the string is $L = (N + 1)h$, and if the two ends are fixed to the x-axis, with one end at the origin, then the *loaded points* are denoted by $k = 1, 2, \ldots, N$. Thus (23) is valid only for those values of k. The *end conditions*

$$y_0 = 0, \qquad y_{N+1} = 0 \tag{24}$$

complete the formulation of the problem.

Once the deflections of the loaded points are determined, the deflections of intermediate points are determined by the *linearity* of the deflection curve between loaded points.

While (23) is perhaps the most convenient form of the governing equation, we notice that the forms

$$\delta^2 y_k = -\frac{h}{T} f_k, \quad \Delta^2 y_k = -\frac{h}{T} f_{k+1}, \quad \nabla^2 y_k = -\frac{h}{T} f_{k-1} \tag{25a,b,c}$$

are equivalent statements of the basic condition. Equation (25a) can be written also in the form

$$T \frac{\delta^2 y(x_k)}{h^2} = -\frac{f(x_k)}{h}. \tag{26}$$

If we let the spacing h tend to zero, so that the discrete loading tends to become continuous, the ratio $f(x_k)/h$ tends toward the linear intensity of *a distributed loading*, say $p(x_k)$, at the point x_k. Hence, if we replace x_k by the coordinate x of an arbitrary interior point and proceed to the limit as $h \to 0$, equation (23) tends toward the

well-known *differential* equation

$$T\frac{d^2y}{dx^2} = -p(x),$$

which governs the case of a distributed load.

In other problems the additional conditions supplementing the difference equation may be in the form of *initial* conditions (as will be seen) in place of *boundary* (end) conditions. Further, the coefficients in the difference equation may depend upon k, as would be the case in the preceding analysis if the assumption of equal tensions in the several segments were abandoned. Finally, *nonlinear* equations may be obtained, as in the case when large deflections of the string are considered, so that the sines and tangents of the slope angles cannot be equated.

As an example of a formulation of a different nature, we consider the evaluation of the integral

$$I_k(\phi) = \int_0^\pi \frac{\cos k\theta - \cos k\phi}{\cos\theta - \cos\phi}\,d\theta \tag{27}$$

where k is zero or a positive integer.* For $k = 0$ and 1, the integration is readily carried out. We next attempt to determine a linear combination of, say, I_{k+1}, I_k, and I_{k-1} which can be integrated in a simple way. Corresponding to the combination

$$A\,I_{k+1} + B\,I_k + C\,I_{k-1},$$

we find that

$$A\cos(k+1)\theta + B\cos k\theta + C\cos(k-1)\theta$$
$$= [(A + C)\cos\theta + B)]\cos k\theta - [(A - C)\sin\theta]\sin k\theta.$$

This combination will contain $\cos\theta - \cos\phi$, the denominator in (27), as a factor if we take

$$A = C, \qquad B = -2A\cos\phi. \tag{28}$$

Thus, if we set $A = 1$, for convenience, and define the operator

$$L = E - 2\cos\phi + E^{-1}, \tag{29}$$

* This integral is of importance, for example, in the Prandtl "lifting line" theory of aerodynamics.

there follows

$$L \cos k\theta = 2(\cos \theta - \cos \phi) \cos k\theta,$$

and also we are fortunate in that

$$L \cos k\phi = 2(\cos \phi - \cos \phi) \cos k\phi = 0.$$

Hence, if we apply the operator L to (27), there follows

$$L I_k(\phi) = 2 \int_0^\pi \frac{(\cos \theta - \cos \phi) \cos k\theta - 0}{\cos \theta - \cos \phi} d\theta$$

$$= 2 \int_0^\pi \cos k\theta \, d\theta = 0 \qquad (k = 1, 2, 3, \cdots).$$

Thus the integral (27) satisfies the linear difference equation

$$I_{k+1} - 2I_k \cos \phi + I_{k-1} = 0 \qquad (k \geqq 1). \tag{30}$$

This equation can be treated as a *recurrence formula*, to deduce that

$$I_2 = 2I_1 \cos \phi - I_0,$$

$$I_3 = 2I_2 \cos \phi - I_1 = (4 \cos^2 \phi - 1)I_1 - 2I_0 \cos \phi,$$

and so forth, by setting k successively equal to 1, 2, ...

But the *initial values* I_0 and I_1 are easily obtained, in the form

$$I_0 = 0, \qquad I_1 = \pi. \tag{31}$$

Hence we see that the difference equation (30), and the initial conditions (31), serve to determine the value of (27) for any positive integral value of k, by step-by-step calculation. However, it is clearly desirable to solve this problem explicitly; that is, to obtain a general expression for I_k which is valid for all positive integral values of k.

In the following sections we consider the problem of finding such explicit solutions in those cases when the equation to be solved is linear, with constant coefficients. It will be shown that for the *homogeneous* equation (with right-hand member zero), an explicit solution can always be obtained in *closed form*, in terms of elementary functions. The same is true of the *nonhomogeneous* equation if the right-hand member is of one of several frequently occurring general types. In other cases, the explicit solution can be obtained in terms of *finite sums* which may or may not be expressible

in closed form, in terms of elementary functions. In particular, solutions will be obtained for the problems formulated in the present section.

3.4. Homogeneous linear difference equations with constant coefficients. We consider first the general *homogeneous* equation of order n, and write it, for convenience, in the form

$$y_{k+n} + A_1 y_{k+n-1} + \cdots + A_{n-1} y_{k+1} + A_n y_k = 0, \tag{32}$$

where the A's are assumed to be constants, and $A_n \neq 0$. With the use of the operator E,

$$E\, y_k = y_{k+1},$$

defined in Section 3.2, equation (32) becomes

$$E^n y_k + A_1 E^{n-1} y_k + \cdots + A_{n-1} E\, y_k + A_n y_k = 0. \tag{32a}$$

If we further define the *linear difference operator*

$$L \equiv E^n + A_1 E^{n-1} + \cdots + A_{n-1} E + A_n, \tag{33}$$

equation (32) or (32a) can be written in the abbreviated form

$$L\, y_k = 0. \tag{34}$$

As in the case of the analogous *differential* equations, it may be expected that (32) will possess solutions of the exponential form e^{rk}, where r is a suitably chosen constant. However, here it is usually more convenient to write $\beta = e^r$, and to attempt to determine solutions of the form

$$y_k = \beta^k. \tag{35}$$

From the results

$$\begin{aligned} E\,\beta^k &= \beta^{k+1} = \beta(\beta^k), \\ E^2\beta^k &= \beta^2(\beta^k), \quad \cdots, \quad E^m\beta^k = \beta^m(\beta^k), \end{aligned} \tag{36}$$

it follows that the result of the operation $L\,\beta^k$ will be merely a linear combination of constant multiples of β^k itself, the coefficients being independent of k. In fact, with the notation of (33), direct substitution shows that

$$L\,\beta^k = (\beta^n + A_1\beta^{n-1} + \cdots + A_{n-1}\beta + A_n)\beta^k. \tag{37}$$

Hence $y_k = \beta^k$ will satisfy (32) if β is a root of the determinantal equation

$$\beta^n + A_1\beta^{n-1} + \cdots + A_{n-1}\beta + A_n = 0. \tag{38}$$

Suppose first that the n roots of (38), say $\beta = \beta_1, \beta_2, \ldots, \beta_n$, are all real and distinct. *Zero* roots are excluded by the condition $A_n \neq 0$, which insures that (32) be indeed of order n. Then, from the *homogeneity* and *linearity* of (32), it follows that any linear combination of the solutions $\beta_1^k, \beta_2^k, \ldots, \beta_n^k$ will also be a solution. That is, any expression of the form

$$y = c_1\beta_1^k + c_2\beta_2^k + \cdots + c_n\beta_n^k \tag{39}$$

then satisfies (32), for arbitrary values of the n constants of combination. As in the case of analogous *differential* equations, it can be shown that since (32) is of order n, and linear, and since (39) involves n *independent* arbitrary constants, (39) represents the *most general* solution of (32).*

In some special cases solutions of the form e^{rk} are more convenient. By writing

$$\beta_m = e^{r_m}, \qquad r_m = \log \beta_m, \tag{40}$$

the general solution (39) then takes the form

$$y_k = c_1e^{r_1k} + c_2e^{r_2k} + \cdots c_ne^{r_nk}, \tag{41}$$

in those cases when the n roots of (38) are distinct.

We remark that the difference equation need not be written in the specific form (32) before the substitution (35) is made. In any case, direct substitution will lead to the equation which plays the role of (38) in determining permissible values of β. The possible presence of zero roots, in such cases, indicates merely that the order of the difference equation is then smaller than the degree of the determinantal equation so obtained. Thus, while the equation $y_{k+2} - y_{k+1} = 0$ is actually of the *first* order, the assumption (35) leads to the extraneous root $\beta = 0$, in addition to the relevant root $\beta = 1$ corresponding to the obvious general solution $y_k = \text{constant}$.

EXAMPLE 1. For the equation

$$y_{k+2} - 5y_{k+1} + 6y_k = 0,$$

the assumption $y_k = \beta^k$ leads to the requirement

$$(\beta^2 - 5\beta + 6)\beta^k = 0,$$

* More specifically, the constants in (39) can be determined in such a way that (39) is identified with any solution of (32) for all relevant *integral values* of k. The case when k is a *continuous* variable is treated in Problem 16.

from which there follows $\beta = 2, 3$. Hence, the general solution is of the form

$$y_k = c_1 2^k + c_2 3^k.$$

EXAMPLE 2. For the equation

$$y_{k+1} - 2y_k \cosh \alpha + y_{k-1} = 0,$$

where α is a constant, the assumption $y_k = e^{rk}$ is found to be more convenient. We thus obtain the requirement

$$(e^r - 2 \cosh \alpha + e^{-r})e^{rk} = 0,$$

from which there follows $\cosh r = \cosh \alpha$, and hence $r = \pm\alpha$. The general solution can then be written in the form

$$y_k = c_1 e^{\alpha k} + c_2 e^{-\alpha k} \quad \text{or} \quad y_k = C_1 \cosh \alpha k + C_2 \sinh \alpha k.$$

Suppose that $\beta = \beta_1$ is a real *double* root of (38). There then follows

$$L\,\beta^k = (\beta - \beta_1)^2(\beta - \beta_3) \cdots (\beta - \beta_n)\beta^k.$$

Hence, in this case, we have not only $(L\,\beta^k)_{\beta=\beta_1} = L\,\beta_1{}^k = 0$, but also

$$\left[\frac{\partial}{\partial\beta}(L\,\beta^k)\right]_{\beta=\beta_1} = L\left[\frac{\partial}{\partial\beta}(\beta^k)\right]_{\beta=\beta_1} = L(k\beta^{k-1})_{\beta=\beta_1} = 0.$$

Thus a second solution, complementing the known solution $\beta_1{}^k$, can be taken as $k\,\beta_1{}^{k-1}$ or, since β_1 is a constant, as $k\,\beta_1{}^k$. More generally, it is readily shown that *the part of the solution corresponding to an m-fold root β_1 is given by*

$$y_k = \beta_1{}^k(c_1 + c_2 k + \cdots + c_m k^{m-1}). \tag{42}$$

EXAMPLE 3. The difference equation

$$\Delta^3 y_k - 3\Delta y_k + 2y_k = 0$$

is reduced by use of (14a) to the form

$$y_{k+3} - 3y_{k+2} + 4y_k = 0.$$

The determinantal equation for β is found to be

$$(\beta + 1)(\beta - 2)^2 = 0,$$

from which there follows $\beta = -1, 2, 2$. Hence the general solution can be taken in the form

$$y_k = c_1(-1)^k + 2^k(c_2 + c_3 k).$$

Finally, if the part of the solution corresponding to a pair of *conjugate complex* roots

$$\beta_1 = a + i\,b, \qquad \beta_2 = a - i\,b \tag{43}$$

is written in the form

$$\begin{aligned} y_k &= A(a + i\,b)^k + B(a - i\,b)^k \\ &= A(\rho\, e^{i\phi})^k + B(\rho\, e^{-i\phi})^k, \end{aligned}$$

where (ρ, ϕ) comprise the polar coordinates of the point (a, b):

$$\rho = \sqrt{a^2 + b^2}, \qquad \phi = \tan^{-1}\frac{b}{a}, \tag{44}$$

there follows also

$$y_k = \rho^k(A\, e^{i\phi k} + B\, e^{-i\phi k}).$$

Hence, by writing $c_1 = A + B$ and $c_2 = i(A - B)$, *the part of the solution corresponding to the roots* (43) *can be expressed in the real form*

$$y_k = \rho^k(c_1 \cos \phi k + c_2 \sin \phi k), \tag{45}$$

where ρ *and* ϕ *are defined by* (44). It is seen that the angle ϕ can be taken as any angle for which $\cos \phi = a/\rho$. Multiple complex roots are treated in an obvious way.

EXAMPLE 4. For the equation

$$y_{k+1} - 2y_k + 2y_{k-1} = 0,$$

the assumption $y_k = \beta^k$ leads to the condition

$$\beta - 2 + \frac{2}{\beta} = 0 \quad \text{or} \quad \beta^2 - 2\beta + 2 = 0,$$

from which there follows $\beta = 1 \pm i$. From (44) we obtain $\rho = \sqrt{2}$, $\phi = \pi/4$, and hence (45) gives the solution

$$y_k = 2^{k/2}\left(c_1 \cos \frac{k\pi}{4} + c_2 \sin \frac{k\pi}{4}\right).$$

EXAMPLE 5. For the equation

$$y_{k+1} - 2y_k \cos \alpha + y_{k-1} = 0,$$

the assumption $y_k = \beta^k$ leads to the requirement

$$\beta = \cos \alpha \pm i \sin \alpha,$$

from which there follows $\rho = 1$ and $\phi = \alpha$. Thus (45) gives the general solution

$$y_k = c_1 \cos \alpha k + c_2 \sin \alpha k.$$

From Examples 2 and 5, we deduce that the solution of the difference equation

$$y_{k+1} - 2A\, y_k + y_{k-1} = 0, \tag{46}$$

where A is a real constant, is of the form

$$y_k = c_1 e^{\alpha k} + c_2 e^{-\alpha k} \quad \text{or} \quad y_k = C_1 \cosh \alpha k + C_2 \sinh \alpha k \tag{47a}$$

where $$A = \cosh \alpha \qquad (A > 1),$$

and also

$$y_k = c_1 \cos \alpha k + c_2 \sin \alpha k \tag{47b}$$

where $$A = \cos \alpha \qquad (|\, A \,| < 1).$$

In a similar way, when $A < -1$, it is found that the solution is of the form

$$y_k = (-1)^k [c_1 e^{\alpha k} + c_2 e^{-\alpha k}]$$

or $$y_k = (-1)^k [C_1 \cosh \alpha k + C_2 \sinh \alpha k] \tag{47c}$$

where $$A = -\cosh \alpha \qquad (A < -1).$$

In the intermediate cases, we find the solutions

$$y_k = c_1 + c_2 k \qquad (A = 1) \tag{47d}$$

and $$y_k = (-1)^k (c_1 + c_2 k) \quad (A = -1). \tag{47e}$$

Equations of the form (46) occur rather frequently in practice. We note that (46) can be written in the equivalent forms

$$\left.\begin{aligned} y_{k+1} - 2y_k + y_{k-1} &= 2(A - 1)y_k, \\ \delta^2 y_k + 2(1 - A)y_k &= 0 \end{aligned}\right\}, \tag{46'}$$

and hence is analogous to a *differential* equation of the form

$$\frac{d^2y}{dx^2} + \lambda\, y = 0,$$

where the constant λ corresponds to $2(1 - A)/h^2$.

In illustration, we have seen in the preceding section that the integral

$$I_k(\phi) = \int_0^\pi \frac{\cos k\theta - \cos k\phi}{\cos \theta - \cos \phi}\, d\theta \tag{48}$$

satisfies the difference equation

$$I_{k+1} - 2I_k \cos \phi + I_{k-1} = 0 \qquad (k \geqq 1), \tag{49}$$

with the initial conditions

$$I_0 = 0, \qquad I_1 = \pi. \tag{50}$$

From (46) and (47b), it follows that I_k must be of the general form

$$I_k(\phi) = c_1 \cos k\phi + c_2 \sin k\phi.$$

The initial conditions give

$$c_1 = 0, \qquad c_2 \sin \phi = \pi,$$

and hence there follows

$$\int_0^\pi \frac{\cos k\theta - \cos k\phi}{\cos \theta - \cos \phi}\, d\theta = \pi \frac{\sin k\phi}{\sin \phi} \tag{51}$$

for any nonnegative integral value of k.

3.5. Particular solutions of nonhomogeneous linear equations. The general solution of a nonhomogeneous linear equation of the form

$$L\, y_k \equiv y_{k+n} + A_1 y_{k+n-1} + \cdots + A_{n-1} y_{k+1} + A_n y_k = \phi_k \tag{52}$$

can be expressed as the sum

$$y_k = y_k^{(H)} + y_k^{(P)}, \tag{53}$$

where $y_k^{(H)}$ is the general solution of the homogeneous equation

$$L\, y_k^{(H)} = 0, \tag{54}$$

and $y_k^{(P)}$ is any particular solution of (52). The coefficients may, of course, be functions of the argument k.

In those cases when (52) has *constant coefficients*, a method of "undetermined coefficients" similar to that applied in the solution of analogous *differential* equations can be used when the right-hand member ϕ_k is a linear combination of terms each having one of

the forms

$$a^k \text{ or } e^{bk}, \quad \sin ck, \quad \cos ck, \quad k^p \ (p = 0, 1, 2, \cdots), \tag{55}$$

or of products of such forms.

Terms of the form a^k or e^{bk}, where a and b may have any constant values, have the property that the operator E^m merely multiplies each such function of k by a constant which is independent of k. That is, we have the relations

$$E^m a^k = (a^m)a^k, \qquad E^m e^{bk} = (e^{bm})e^{bk}. \tag{56}$$

For terms of the form $\cos ck$ and $\sin ck$, there follows

$$\left.\begin{aligned} E^m \cos ck &= \cos c(k + m) \\ &= (\cos cm) \cos ck - (\sin cm) \sin ck, \\ E^m \sin ck &= \sin c(k + m) \\ &= (\sin cm) \cos ck + (\cos cm) \sin ck \end{aligned}\right\}. \tag{57}$$

Hence, the result of operating on $\cos ck$ or $\sin ck$ by any power of E can be expressed as a linear combination of $\cos ck$ *and* $\sin ck$, the constants of combination being independent of k. Finally, for a term k^p, where p is a positive integer (or zero), there follows

$$E^m k^p = (k + m)^p = k^p + p\, m\, k^{p-1} + \cdots + p\, m^{p-1} k + m^p. \tag{58}$$

Hence, if p is a nonnegative integer, the result of operating on k^p by any power of E can be expressed as a linear combination of the terms $k^p, k^{p-1}, \ldots, k, 1$.

Thus, we may speak of the "families" $\{a^k\}$, $\{e^{bk}\}$, $\{\sin ck, \cos ck\}$, and $\{k^p, k^{p-1}, \ldots, k, 1\}$ where p is a nonnegative integer, in the sense that the *family* of a term f_k is defined to be the set of all functions of which f_k and all operations $E^m f_k$ are linear combinations. Only the functions listed in (55), and products or linear combinations of such functions or products, have *finite* families. It is easily shown that the family of the product $f_k\, g_k$ consists of all possible products, in each of which one and only one member of each of the families of f_k and g_k appears.

The *method of undetermined coefficients*, for obtaining a particular solution of a linear difference equation with *constant coefficients*, corresponding to a right-hand term f_k with a *finite family*, is to be

applied after the general solution of the associated homogeneous equation has been obtained. The procedure may be outlined as follows:

1. Construct the family of that term.

2. If that family has no representatives in the *homogeneous* solution, assume $y_k^{(P)}$ as a linear combination of the members of that family, and determine the constants of combination in such a way that the difference equation is *identically* satisfied.

3. If that family has a representative in the homogeneous solution, multiply *each* member of the family by the smallest integral power of k for which all such representatives are removed, and assume as a particular solution a linear combination of the members of the *modified* family.

Proof that this procedure always succeeds in the cases described is lengthy, and is omitted. However, it can be shown that if the family of f_k possesses n members, the requirement that the assumed form reduce the left-hand member identically to f_k leads always to a set of n linear algebraic equations in the n unknown constants of combination, and that this set of equations always possesses a solution.

As an example, we consider the equation

$$y_{k+1} - 3y_k + 2y_{k-1} = 1 + a^k, \tag{59}$$

where a is a constant. The homogeneous solution is found to be

$$y_k^{(H)} = c_1 + c_2 2^k. \tag{60a}$$

Since $y_k = 1$ is a homogeneous solution, we assume a particular solution in the form

$$y_k^{(P)} = A\,k + B\,a^k. \tag{60b}$$

Substitution into (59) leads to the requirement

$$-A + \left(a - 3 + \frac{2}{a}\right) B\,a^k \equiv 1 + a^k,$$

from which there follows

$$A = -1, \qquad B = \frac{a}{(a-1)(a-2)},$$

if $a \neq 1$ or 2. Thus, with this stipulation, the required solution $y_k = y_k^{(H)} + y_k^{(P)}$ takes the form

$$y_k = c_1 + c_2 2^k - k + \frac{a^{k+1}}{(a-1)(a-2)}. \tag{61}$$

The two exceptional cases, in which (60b) involves a member of (60a), must be treated separately.

Certain other procedures, applicable to the determination of particular solutions of linear *differential* equations, can also be translated to analogous procedures for dealing with *difference* equations. In such cases, the process of *integration* translates into a process of *summation*, the sums involving only a finite number of terms. In certain cases, these sums can then be expressed in closed form.

As the simplest example, we consider the analogy to the differential equation $dy/dx = f(x)$,

$$\triangle y_k = f_k \quad \text{or} \quad y_{k+1} - y_k = f_k. \tag{62}$$

If we write

$$y_0 = c,$$

equation (62) gives successively

$$y_1 = c + f_0, \quad y_2 = c + f_0 + f_1, \quad y_3 = c + f_0 + f_1 + f_2,$$

and hence, by induction,

$$y_k = c + \sum_{n=0}^{k-1} f_n \equiv c + \sum_{n=1}^{k} f_{n-1}. \tag{63}$$

The arbitrary constant c is seen to be the general *homogeneous* solution of (62), corresponding to $f_k = 0$.

If we notice that a change in the lower limit of the summation can be compensated by a change in c, it follows that *a particular solution of the equation*

$$\triangle y_k = f_k \tag{64}$$

is given by

$$y_k = \Sigma^k f_{n-1}, \tag{65}$$

where we adopt the convention that *the symbol Σ^k indicates summation with respect to the relevant dummy variable* [denoted by n in (65)] *from an arbitrarily fixed integral lower limit to the variable upper limit k.*

We next indicate the application of the method of "variation of parameters" to the solution of linear difference equations, following a procedure analogous to that used in dealing with differential equations.

In the case of the general linear difference equation of the *first* order,

$$L\,y_k \equiv y_{k+1} + A\,y_k = \phi_k, \tag{66}$$

where A may be a function of k, we suppose that the general homogeneous solution

$$y_k^{(H)} = c\,u_k \tag{67}$$

has been obtained, so that

$$L\,u_k = 0 \tag{68}$$

and c is an arbitrary constant. We then attempt to find a solution of (66) in the form

$$y_k = C_k u_k, \tag{69}$$

where C_k is now an unknown function of k, to be determined.

For the determination of C_k, we first notice that from (69) there follows

$$y_{k+1} = C_{k+1}u_{k+1} \equiv C_k u_{k+1} + (C_{k+1} - C_k)u_{k+1}$$

and hence, with the usual abbreviation

$$\triangle C_k = C_{k+1} - C_k, \tag{70}$$

we may write also

$$y_{k+1} = C_k u_{k+1} + u_{k+1}\,\triangle C_k. \tag{71}$$

This artifice leads to a simple method of determining C_k. For the introduction of (69) into (66) then leads to the condition

$$C_k(u_{k+1} + A\,u_k) + u_{k+1}\,\triangle C_k = \phi_k \tag{72}$$

in which the coefficient of C_k vanishes in virtue of (68). Thus, (69) will satisfy (66) if C_k is determined in such a way that

$$\triangle C_k = \frac{\phi_k}{u_{k+1}}, \tag{73}$$

or, in accordance with (64) and (65),

$$C_k = \Sigma^k \frac{\phi_{n-1}}{u_n}. \tag{74}$$

For any conveniently fixed lower limit of summation, the introduction of (74) into (69) then gives a particular solution of (66),

$$y_k^{(P)} = u_k \sum^k \frac{\phi_{n-1}}{u_n}, \tag{75}$$

which may be added to (67) to give the complete solution.

EXAMPLE 6. For the equation

$$y_{k+1} - k\, y_k = k! \qquad (k \geqq 1),$$

a *homogeneous* solution is found (by induction or by inspection) in the form

$$y_k^{(H)} = c(k-1)! \equiv c\, u_k.$$

Hence (75) gives a particular solution

$$y_k^{(P)} = (k-1)! \sum_{n=1}^{k} \frac{(n-1)!}{(n-1)!} = k(k-1)! = k!,$$

and the general solution can be taken in the form

$$y_k = k! + c(k-1)! \qquad (k \geqq 1),$$

where c is an arbitrary constant.

In the case of the general linear difference equation of *second* order,

$$L\, y_k \equiv y_{k+2} + A\, y_{k+1} + B\, y_k = \phi_k, \tag{76}$$

we suppose again that the general *homogeneous* solution has been obtained in the form

$$y_k^{(H)} = c_1 u_k + c_2 v_k, \tag{77}$$

where c_1 and c_2 are arbitrary constants, and assume a solution to (76) in the form

$$y_k = C_k^{(1)} u_k + C_k^{(2)} v_k, \tag{78}$$

where $C_k^{(1)}$ and $C_k^{(2)}$ are functions of k to be determined. By proceeding as in the transition from (69) to (71), we find that

$$y_{k+1} = C_k^{(1)} u_{k+1} + C_k^{(2)} v_{k+1} + [u_{k+1}\, \Delta C_k^{(1)} + v_{k+1}\, \Delta C_k^{(2)}]. \tag{79}$$

As the first of two conditions needed to determine both $C_k^{(1)}$ and $C_k^{(2)}$, we require that the bracketed expression in (79) vanish, and hence arbitrarily impose the condition

$$u_{k+1}\, \Delta C_k^{(1)} + v_{k+1}\, \Delta C_k^{(2)} = 0. \tag{80}$$

There then follows

$$y_{k+1} = C_k^{(1)} u_{k+1} + C_k^{(2)} v_{k+1}, \tag{81}$$

and hence also

$$y_{k+2} = C_k^{(1)} u_{k+2} + C_k^{(2)} v_{k+2} + u_{k+2} \Delta C_k^{(1)} + v_{k+2} \Delta C_k^{(2)}. \tag{82}$$

If (78), (81), and (82) are introduced into (76), and use is made of the fact that u_k and v_k satisfy the homogeneous equation associated with (76), the second condition complementing (80) is readily obtained in the form

$$u_{k+2} \Delta C_k^{(1)} + v_{k+2} \Delta C_k^{(2)} = \phi_k. \tag{83}$$

Equations (80) and (83) permit the determination of $\Delta C_k^{(1)}$ and $\Delta C_k^{(2)}$ in terms of known quantities, after which $C_k^{(1)}$ and $C_k^{(2)}$ are determined by summation. The general solution of (76) is then obtained by adding (78) to (77).

EXAMPLE 7. We consider the equation

$$y_{k+1} - 2y_k \cos \alpha + y_{k-1} = f_k \qquad (k \geq 1),$$

noticing that, when this equation is written in the "standard form" (76), *there follows* $\phi_k = f_{k+1}$. The general homogeneous solution is $y_k^{(H)} = c_1 u_k + c_2 v_k$, where

$$u_k = \cos k\alpha, \qquad v_k = \sin k\alpha,$$

if $|\cos \alpha| < 1$, in accordance with (47b). If we assume a particular solution in the form

$$y_k^{(P)} = C_k^{(1)} u_k + C_k^{(2)} v_k,$$

equations (80) and (83) take the form

$$[\cos (k+1)\alpha] \Delta C_k^{(1)} + [\sin (k+1)\alpha] \Delta C_k^{(2)} = 0,$$
$$[\cos (k+2)\alpha] \Delta C_k^{(1)} + [\sin (k+2)\alpha] \Delta C_k^{(2)} = f_{k+1}.$$

The determinant of the coefficients reduces to $\sin \alpha$, and the solution by determinants gives

$$\Delta C_k^{(1)} = -\frac{f_{k+1} \sin (k+1)\alpha}{\sin \alpha}, \qquad \Delta C_k^{(2)} = +\frac{f_{k+1} \cos (k+1)\alpha}{\sin \alpha}.$$

There then follows, by summation,

$$C_k^{(1)} = -\sum_{n=1}^{k} \frac{f_n \sin n\alpha}{\sin \alpha}, \qquad C_k^{(2)} = +\sum_{n=1}^{k} \frac{f_n \cos n\alpha}{\sin \alpha},$$

and the general solution of the given equation takes the form

$$y_k = -\cos k\alpha \sum_{n=1}^{k} \frac{f_n \sin n\alpha}{\sin \alpha} + \sin k\alpha \sum_{n=1}^{k} \frac{f_n \cos n\alpha}{\sin \alpha} + c_1 \cos k\alpha + c_2 \sin k\alpha.$$

After an obvious reduction, this solution becomes

$$y_k = \frac{1}{\sin \alpha} \sum_{n=1}^{k} f_n \sin (k-n)\alpha + c_1 \cos k\alpha + c_2 \sin k\alpha.$$

We notice that this solution is valid only when $\sin \alpha \neq 0$, that is, when the coefficient $\cos \alpha$ in the given difference equation is numerically *smaller* than unity, as is required by equation (47b).

The extension of the above procedure to the treatment of linear equations of higher order is completely analogous to the corresponding extension in the case of *differential* equations.

3.6. The loaded string. To illustrate the basic types of problems involving linear difference equations, we take as our model the tightly stretched string (Figure 3.2) loaded at N equally

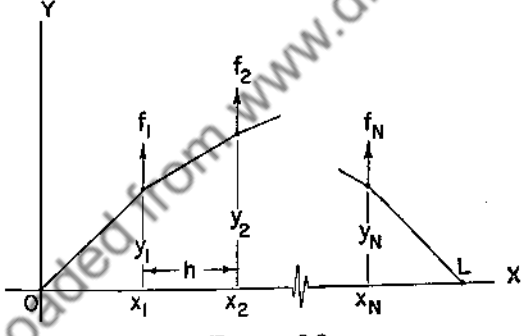


FIGURE 3.2

spaced points $x_k = kh$ $(k = 1, 2, \ldots, N)$ by forces f_k, and attached to the x-axis at the points $x_0 = 0$ and $x_{N+1} = (N+1)h = L$. In Section 3.3, it was shown that for *small* deflections y_k the problem is governed by the difference equation

$$\delta^2 y_k \equiv y_{k+1} - 2y_k + y_{k-1} = -\frac{h}{T} f_k \qquad (1 \leqq k \leqq N), \quad (84)$$

where h is the horizontal spacing and T the tension (assumed to be constant), and by the end conditions

$$y_0 = 0, \qquad y_{N+1} = 0. \quad (85a,b)$$

As a *first example*, we suppose that N beads of equal mass M are attached to the string, at the points x_k, and that the string hangs under the action of gravity. Assuming always that the mass of the string itself is relatively negligible, we then have

$$f_k = -Mg, \tag{86}$$

so that (84) becomes

$$y_{k+1} - 2y_k + y_{k-1} = \frac{Mgh}{T}. \tag{87}$$

The general homogeneous solution of (87) is of the form

$$y_k^{(H)} = c_1 + c_2 k. \tag{88}$$

Accordingly, the method of undetermined coefficients leads to the assumption

$$y_k^{(P)} = A\,k^2, \tag{89}$$

and substitution of this expression into (87) gives the condition

$$A = \frac{Mgh}{2T}. \tag{90}$$

Hence the general solution of (87) is of the form

$$y_k = \frac{Mgh}{2T}\,k^2 + c_1 + c_2 k. \tag{91}$$

The end conditions (85a,b) then give

$$c_1 = 0, \qquad c_2 = -\frac{Mgh}{2T}\,(N+1), \tag{92}$$

and (91) becomes

$$y_k = \frac{Mgh}{2T}\,k^2 - \frac{Mgh}{2T}\,(N+1)k$$

or, finally,

$$y_k = -\frac{Mgh}{2T}\,k(N+1-k) \qquad (0 \leqq k \leqq N+1). \tag{93}$$

If we notice that $x_k = kh$ and $L = (N+1)h$, this result can be written also in the form

$$y(x_k) = -\frac{Mg}{2Th}\,x_k(L - x_k). \tag{94}$$

Hence the segments of the deflected string are chords of the parabola

$$y = -\frac{Mg}{2Th}\,x(L - x). \tag{95}$$

In the limiting case as $h \to 0$, we may replace Mg/h by the linear intensity p of a load uniformly distributed in the horizontal direction, and the well-known parabolic form

$$y = -\frac{p}{2T}\,x(L - x)$$

is obtained.

As a *second example*, we suppose that the string, with beads attached, is rotating about the x-axis with uniform angular velocity ω, the mass of the string itself again being neglected. Then f_k is to be replaced in (84) by the inertia force

$$f_k = M\omega^2 y_k, \tag{96}$$

so that (84) takes the form

$$y_{k+1} - 2\left(1 - \frac{M\omega^2 h}{2T}\right) y_k + y_{k-1} = 0. \tag{97}$$

We stipulate first that the speed of rotation is sufficiently small that

$$1 - \frac{M\omega^2 h}{2T} > -1 \qquad \text{or} \qquad \frac{M\omega^2 h}{4T} < 1. \tag{98}$$

The coefficient of $2y_k$ in (97) is then less than unity in absolute value, and we may introduce the abbreviation

$$1 - \frac{M\omega^2 h}{2T} = \cos\alpha. \tag{99}$$

The general solution of (97), with the notation of (99), is then given by (47b) in the form

$$y_k = c_1 \cos \alpha k + c_2 \sin \alpha k. \tag{100}$$

The end condition $y_0 = 0$ requires that

$$c_1 = 0, \tag{101}$$

while the second end condition $y_{N+1} = 0$ leads to the condition

$$c_2 \sin \alpha(N + 1) = 0. \tag{102}$$

Unless α has a value for which $\sin \alpha(N+1) = 0$, the only solution of (102) is $c_2 = 0$, in which case (100) reduces to the *trivial* solution $y_k \equiv 0$, and no deflection can exist.

However, if it happens that

$$\alpha(N+1) = n\pi \qquad (n = 1, 2, 3, \cdots), \tag{103}$$

equation (102) is then satisfied identically, and c_2 is arbitrary. That is, if we write

$$\alpha_n = \frac{n\pi}{N+1} \qquad (n = 1, 2, 3, \cdots), \tag{104}$$

we conclude that no deflection occurs unless $\alpha = \alpha_n$, in which case the present linearized theory gives only the *shape* of the corresponding deflection curve y_n, specified by the ordinates

$$y_{n,k} = C_n \sin \frac{n\pi k}{N+1} \qquad (k = 0, 1, \cdots, N+1), \tag{105}$$

where C_n is undetermined. Since N and k are integral, it is seen that *the values* $n = 1, 2, 3, \ldots, N$ *lead to all possible distinct deflection modes.* All other integral values of n lead either to *no* deflection ($n = 0, N+1, 2N+2, \ldots$) or to deflection modes identical with those just listed.

If we denote the value of ω corresponding to α_n by ω_n, equation (99) gives

$$1 - \frac{Mh\omega_n^2}{2T} = \cos \alpha_n$$

or

$$\omega_n^2 = \frac{2T}{Mh}(1 - \cos \alpha_n) = \frac{4T}{Mh} \sin^2 \frac{\alpha_n}{2}. \tag{106}$$

Hence, using (104), we obtain the *critical speeds* ω_n in the form

$$\omega_n = 2\sqrt{\frac{T}{Mh}} \sin \frac{n\pi}{2(N+1)} \qquad (n = 1, 2, \cdots, N). \tag{107}$$

The number N of distinct critical speeds and deflection modes is seen to be equal to the number of distinct masses present.

With $k = x_k/h$ and $N + 1 = L/h$, equations (107) and (105) can be written in the form

$$\omega_n = 2\sqrt{\frac{T}{Mh}} \sin \frac{n\pi h}{2L} \qquad \left(n = 1, 2, \cdots, \frac{L}{h} - 1\right) \tag{108a}$$

and
$$y_n(x_k) = C_n \sin \frac{n\pi x_k}{L}. \tag{108b}$$

In the limiting case when $h \to 0$, if we write $\rho = M/h$, we have

$$\omega_n = \lim_{h\to 0} \frac{2}{h}\sqrt{\frac{T}{\rho}} \sin \frac{n\pi h}{2L} = \frac{n\pi}{L}\sqrt{\frac{T}{\rho}}, \tag{109}$$

where ρ is a limiting uniform mass density. The number of relevant values of n in (108a) increases without limit as $h \to 0$, in accordance with the fact that a rotating string of uniform mass density possesses an infinite set of critical speeds.

From (108b) we see that the segments of the string in the nth deflection mode are chords of the curve representing

$$y = C \sin \frac{n\pi x}{L},$$

and that there are exactly N distinct deflection modes of this type, each corresponding to one of the N critical values of ω given by (107).

If we write

$$\lambda = \frac{Mh\omega^2}{T}, \tag{110}$$

we conclude that the problem which consists of the linear homogeneous difference equation

$$\delta^2 y_k + \lambda y_k = 0 \tag{111a}$$

and the homogeneous boundary conditions

$$y_0 = 0, \qquad y_{N+1} = 0 \tag{111b}$$

determines N *characteristic values* of the quantity λ,

$$\lambda_n = 4 \sin^2 \frac{n\pi}{2(N+1)} \qquad (n = 1, 2, \cdots, N), \tag{111c}$$

and N corresponding *characteristic functions*

$$\phi_{n,k} = \sin \frac{n\pi k}{N+1} \qquad (n = 1, 2, \cdots, N). \tag{111d}$$

We have supposed that the speed ω is such that the inequality (98) is satisfied, that is, that $\lambda < 4$. The characteristic values (111c) are in accordance with this assumption. If we assume that

(98) is *not* satisfied, so that $\lambda \geqq 4$, the general solution of (111a) can be expressed by (47c) in the form

$$y_k = (-1)^k(c_1 e^{\alpha k} + c_2 e^{-\alpha k})$$

when $\lambda > 4$, where α is a real constant defined by

$$\cosh \alpha = 1 + \tfrac{1}{2}(\lambda - 4),$$

and is of the form

$$y_k = (-1)^k(c_1 + c_2 k)$$

when $\lambda = 4$. In either case it is readily verified that the end conditions (111b) can be satisfied only if $c_1 = c_2 = 0$, and hence $y \equiv 0$. That is, there are no real characteristic values of λ in addition to those given by (111c).

As a *third example*, we suppose again that N beads of equal mass M are attached to a stretched string with negligible mass, and study possible *free vibrations* of the system in a plane. That is, we suppose that no physical external force acts on the system, but that the separate beads are each given certain initial displacements and velocities in a plane at the time $t = 0$, and we investigate the motion of the system at all following times.

In order to obtain the equations of motion, we may replace the force f_k in equation (84) by the inertia force $-M\,\partial^2 y_k/\partial t^2$, so that (84) becomes

$$y_{k+1} - 2y_k + y_{k-1} = \frac{Mh}{T}\frac{\partial^2 y_k}{\partial t^2}. \tag{112}$$

We first seek the *natural modes* of free vibration, in each of which all the beads are vibrating in phase with a common frequency, and write, for the kth bead,

$$y_k = A_k \cos(\omega t + \beta) \tag{113}$$

where β is a constant phase angle, and A_k is the amplitude associated with the kth bead.

If (113) is introduced into (112), and the resulting common time factor is cancelled, there follows

$$A_{k+1} - 2A_k + A_{k-1} = -\frac{M\omega^2 h}{T} A_k. \tag{114}$$

If the ends $x = 0$ and $x = (N + 1)h$ of the string are restrained from motion, we have also the boundary conditions

$$A_0 = 0, \qquad A_{N+1} = 0. \tag{115}$$

But (114) and (115) are equivalent to (111a,b). Hence, from the preceding results, it follows that solutions of the form (113) exist if and only if ω takes on one of the N permissible values

$$\omega_n = 2\sqrt{\frac{T}{Mh}} \sin \frac{n\pi}{2(N+1)} \qquad (n = 1, 2, \cdots, N), \tag{116}$$

in which case the corresponding amplitude of vibration of the kth bead is given by an expression of the form

$$A_{n,k} = C_n \sin \frac{n\pi k}{N+1}. \tag{117}$$

It follows also that any mode of vibration in which the kth bead vibrates according to the law

$$y_{n,k} = C_n \sin \frac{n\pi k}{N+1} \cos (\omega_n t + \beta_n) \qquad (n = 1, 2, \cdots, N) \tag{118}$$

satisfies the basic equation (112) and the prescribed boundary conditions for arbitrary constant values of C_n and β_n. From the linearity of the problem, the same is true of any linear combination of such expressions, say

$$y_k = \sum_{n=1}^{N} C_n \sin \frac{n\pi k}{N+1} \cos (\omega_n t + \beta_n). \tag{119}$$

But this expression contains $2N$ arbitrary constants $C_1, \ldots, C_N$ and $\beta_1, \ldots, \beta_N$ which presumably can be chosen in such a way that the displacements and velocities of each of the N beads all take on prescribed values when $t = 0$. The determination of the constants is treated in Section 3.9.

The limiting form of (112), as $h \to 0$ and M/h tends toward a uniform linear mass density ρ, is seen to be the *wave equation*

$$\frac{\partial^2 y}{\partial x^2} = \frac{\rho}{T} \frac{\partial^2 y}{\partial t^2},$$

as would be expected.

It is of some interest to investigate the formulation of the general problem of the discretely loaded string when the assumption of uniform tension T is abandoned and replaced by the assumption

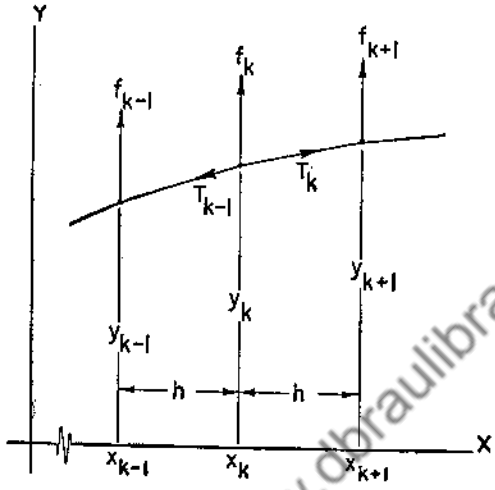


FIGURE 3.3

that the tension is constant only between successive points of force application. If the tension in the segment $x_k x_{k+1}$ is denoted by T_k (Figure 3.3), equation (21) is unchanged,

$$y_{k+1} - y_k = h \tan \phi_k, \tag{120}$$

whereas (22) takes the modified form

$$T_k \sin \phi_k - T_{k-1} \sin \phi_{k-1} + f_k = 0. \tag{121}$$

Again assuming small slope angles, we again approximate (120) in the form

$$\sin \phi_k \approx \frac{1}{h}(y_{k+1} - y_k),$$

and introduce this result into (121) to obtain the equation

$$T_k(y_{k+1} - y_k) - T_{k-1}(y_k - y_{k-1}) = -h f_k.$$

This equation can be written in several equivalent forms, such as

$$T_k y_{k+1} - (T_k + T_{k-1})y_k + T_{k-1}y_{k-1} = -h f_k, \tag{122a}$$

$$\Delta(T_{k-1} \Delta y_{k-1}) = -h f_k, \tag{122b}$$

and $$\nabla(T_k \triangle y_k) = -h f_k. \tag{122c}$$

3.7. Properties of sums and differences. We have seen that, in the solution of *difference* equations, the operators $\triangle$ and Σ^k are, to a certain extent, analogous to the differential and integral operators which relate to *differential* equations. In this section, we examine this analogy more closely.

It has been shown that if $\triangle y_k = f_k$, then $y_k = \Sigma^k f_{n-1} + c$ or $y_k = \Sigma^{k-1} f_n + C$, where the summation with respect to the dummy variable n extends from a convenient lower limit, say M, to the indicated upper limit. This result is analogous to the statement that if $\frac{dy}{dx} = f(x)$, then $y(x) = \int^x f(\xi)\, d\xi + c$, where the integration with respect to the dummy variable ξ extends from a convenient lower limit, say a, to the upper limit x.

It follows by substitution that

$$\sum_M^{k-1} \triangle y_n = y_k + c. \tag{123}$$

To determine the "constant of summation" we may set $k = M + 1$ in (123) and so obtain $\triangle y_M = y_{M+1} + c$ or

$$c = -y_M. \tag{124}$$

Thus (123) takes the form

$$\sum_M^{k-1} \triangle y_n = y_k - y_M. \tag{125}$$

This result is also easily verified directly, since the left-hand member is given by

$$(y_{M+1} - y_M) + (y_{M+2} - y_{M+1}) + \cdots + (y_{k-1} - y_{k-2}) + (y_k - y_{k-1}),$$

and this sum evidently "telescopes" into the result of (125).

With a change in notation, equation (125) can also be written in the form

$$\sum_M^N \triangle y_k = y_{N+1} - y_M, \tag{126}$$

which is seen to be analogous to the relation

$$\int_a^b \frac{dy}{dx}\,dx = y(b) - y(a).$$

It should be carefully noticed, however, that the right-hand member of (126) is *not* $y_N - y_M$, as might have been formally expected.

Corresponding to the product formula $d(uv)/dx = u\,dv/dx + v\,du/dx$, we find that

$$\begin{aligned}\Delta(u_k v_k) &= u_{k+1}v_{k+1} - u_k v_k \\ &\equiv v_{k+1}(u_{k+1} - u_k) + u_k(v_{k+1} - v_k),\end{aligned}$$

and hence we may write

$$\Delta(u_k v_k) = u_k\,\Delta v_k + v_{k+1}\,\Delta u_k. \tag{127}$$

From the symmetry of the left-hand member, it follows that u and v may also be interchanged in the right-hand member.

If we sum the equal members of this relation with respect to k from M to N, there follows

$$\sum_M^N \Delta(u_k v_k) = \sum_M^N u_k\,\Delta v_k + \sum_M^N v_{k+1}\,\Delta u_k.$$

But, according to (126), the left-hand member is given by

$$u_{N+1}v_{N+1} - u_M v_M \equiv \Big[u_k v_k\Big]_M^{N+1},$$

and hence the preceding results can be transposed into the form

$$\sum_M^N u_k\,\Delta v_k = \Big[u_k v_k\Big]_M^{N+1} - \sum_M^N v_{k+1}\,\Delta u_k. \tag{128}$$

This result is the formula for *summation by parts*, and is analogous to the familiar formula for *integration* by parts. It will be particularly useful in the developments of Section 3.9.

To illustrate the explicit use of (128), we consider the sum

$$\sum_0^N k\,r^k \qquad (r \neq 1),$$

where r is a constant. If we write

$$u_k = k, \qquad \Delta v_k = r^k,$$

we have also

$$\Delta u_k = 1, \qquad v_k = \sum_0^{k-1} r^n + c = \frac{r^k - 1}{r - 1} + c = \frac{r^k}{r - 1} + C,$$

since Σr^k is a geometric series. Taking $C = 0$, for convenience, the use of (128) leads to the result

$$\begin{aligned}\sum_0^N k\, r^k &= \left[\frac{k\, r^k}{r - 1}\right]_0^{N+1} - \frac{1}{r - 1}\sum_0^N r^{k+1} \\ &= \frac{(N + 1)r^{N+1}}{r - 1} - \frac{1}{r - 1}\left[\frac{r^{N+2} - r}{r - 1}\right] \\ &= \frac{1}{(r - 1)^2}\left[N\, r^{N+2} - (N + 1)r^{N+1} + r\right] \qquad (r \neq 1).\end{aligned}$$

In particular, we may proceed to the limit $N \to \infty$ if $|\, r\,| < 1$, and so obtain also the result

$$\sum_0^\infty k\, r^k = \frac{r}{(r - 1)^2} \qquad (|\, r\,| < 1).$$

It is of interest to notice that the same result is obtainable as follows:

$$\sum_0^N k\, r^k = r\sum_0^N k\, r^{k-1} = r\,\frac{d}{dr}\sum_0^N r^k = r\,\frac{d}{dr}\left(\frac{r^{N+1} - 1}{r - 1}\right).$$

We remark that we have arbitrarily chosen to deal with the operator Δ. Corresponding results can be obtained in terms of ∇.

3.8. Special finite sums. In this section we list certain frequently occurring finite sums which can be expressed in closed form.

First, the sum of a finite geometric series

$$\sum_1^K r^k = \frac{r^{K+1} - r}{r - 1} \qquad (r \neq 1), \tag{129}$$

where r is a constant, has already been considered. If we set $r = e^{i\alpha}$, where α is a real constant, (129) becomes

$$\sum_1^K e^{ik\alpha} = \frac{e^{i(K+1)\alpha} - e^{i\alpha}}{e^{i\alpha} - 1} \equiv \left[\frac{e^{iK\alpha/2} - e^{-iK\alpha/2}}{e^{i\alpha/2} - e^{-i\alpha/2}}\right] e^{i\left(\frac{K+1}{2}\right)\alpha}$$

$$= \frac{\sin\frac{K}{2}\alpha}{\sin\frac{1}{2}\alpha} e^{i\left(\frac{K+1}{2}\right)\alpha} \qquad (\alpha \neq 0, \pm 2\pi, \cdots). \quad (130)$$

Hence, by equating real and imaginary parts in (130), we obtain the results

$$c_K(\alpha) \equiv \sum_1^K \cos k\alpha = \frac{\sin\frac{K}{2}\alpha \cos\frac{K+1}{2}\alpha}{\sin\frac{1}{2}\alpha} \qquad (\alpha \neq 0, \pm 2\pi, \cdots) \quad (131)$$

and

$$s_K(\alpha) \equiv \sum_1^K \sin k\alpha = \frac{\sin\frac{K}{2}\alpha \sin\frac{K+1}{2}\alpha}{\sin\frac{1}{2}\alpha} \qquad (\alpha \neq 0, \pm 2\pi, \cdots). \quad (132)$$

It may be noticed in all cases that, if the lower limit differs from unity, the obvious relation

$$\sum_M^K \phi_k \equiv \sum_1^K \phi_k - \sum_1^{M-1} \phi_k \quad (133)$$

is useful.

From the results (131) and (132), many other sums may be obtained. For example, we notice that

$$\sum_1^K k \sin k\alpha = -\frac{d}{d\alpha} c_K(\alpha), \qquad \sum_1^K k \cos k\alpha = \frac{d}{d\alpha} s_K(\alpha), \quad (134a,b)$$

and so forth. Also, we have the relations

$$\sum_1^K \sin k\alpha \sin k\beta = \frac{1}{2}\sum_1^K [\cos k(\alpha - \beta) - \cos k(\alpha + \beta)]$$

$$= \frac{1}{2}[c_K(\alpha - \beta) - c_K(\alpha + \beta)], \quad (135a)$$

and, similarly,

$$\sum_1^K \sin k\alpha \cos k\beta = \frac{1}{2}[s_K(\alpha + \beta) + s_K(\alpha - \beta)] \qquad (135b)$$

and

$$\sum_1^K \cos k\alpha \cos k\beta = \frac{1}{2}[c_K(\alpha + \beta) + c_K(\alpha - \beta)]. \qquad (135c)$$

If we notice that

$$c_K(0) = K, \qquad s_K(0) = 0, \qquad (136)$$

we may derive from these forms the further results

$$\sum_1^K \sin^2 k\alpha = \frac{K}{2} - \frac{1}{2}c_K(2\alpha) = \frac{K}{2} - \frac{\sin K\alpha \cos (K+1)\alpha}{2 \sin \alpha} \qquad (137a)$$

and

$$\sum_1^K \cos^2 k\alpha = \frac{K}{2} + \frac{1}{2}c_K(2\alpha) = \frac{K}{2} + \frac{\sin K\alpha \cos (K+1)\alpha}{2 \sin \alpha}. \qquad (137b)$$

For a sum of the form

$$S_k(p) = 1^p + 2^p + 3^p + \cdots + k^p, \qquad (138)$$

where p is a nonnegative integer, we notice that $S_k(p)$ satisfies the difference equation

$$S_k - S_{k-1} = k^p \qquad (k \geqq 2) \qquad (139)$$

and the initial condition $S_1 = 1$. The homogeneous solution of (139) is merely $S_k^{(H)} = c$, and the method of undetermined coefficients then shows that a particular solution is a linear combination of the terms k^{p+1}, k^p, . . . , k^2, k. It follows that (138) *can be expressed as a polynomial of degree* $(p + 1)$ *in* k,

$$S_k(p) = c + A_1 k + A_2 k^2 + \cdots + A_{p+1}k^{p+1}. \qquad (140)$$

The A's can be determined by substitution into (139), after which c is determined by the initial condition $S_1 = 1$. In this way, results in the cases $p = 2$, 3, and 4 are obtained in the form

$$1^2 + 2^2 + 3^2 + \cdots + k^2 = \tfrac{1}{6}k(1 + k)(1 + 2k), \tag{141a}$$

$$1^3 + 2^3 + 3^3 + \cdots + k^3 = \tfrac{1}{4}k^2(1 + k)^2, \tag{141b}$$

$$1^4 + 2^4 + 3^4 + \cdots + k^4 = \tfrac{1}{30}k(1 + k)(1 + 2k)(3k^2 + 3k - 1). \tag{141c}$$

The summation of (138) can also be affected by use of the results of Problem 6.

Products of the form $k(k - 1)(k - 2) \cdots (k - p)$ are of frequent occurrence in the solution of difference equations. We use here the abbreviations

$$k^{(m)} = k(k - 1)(k - 2) \cdots (k - m + 1) \equiv \frac{\Gamma(k + 1)}{\Gamma(k - m + 1)} \tag{142}$$

and

$$k^{(-m)} = \frac{1}{(k + 1)(k + 2) \cdots (k + m)} \equiv \frac{\Gamma(k + 1)}{\Gamma(k + m + 1)}, \tag{143}$$

where m is a positive integer, and notice that $k^{(m)}$ is then a polynomial of degree m in k, which vanishes when $k = 0$, whereas $k^{(-m)}$ is the *reciprocal* of an mth degree polynomial in k.* To see the importance of such functions, and the usefulness of the abbreviations, we calculate the differences $\Delta k^{(m)}$ and $\Delta k^{(-m)}$ as follows:

$$\begin{aligned}\Delta k^{(m)} &= [(k + 1)k(k - 1) \cdots (k - m + 2)] \\ &\qquad - [k(k - 1) \cdots (k - m + 2)(k - m + 1)] \\ &= [(k + 1) - (k - m + 1)][k(k - 1) \cdots (k - m + 2)] \\ &= m\,k^{(m-1)},\end{aligned} \tag{144a}$$

* While the notation (142) is rather conventional, several different interpretations of (143) are in use, some of which lack the consistency of the notation used here. If the Gamma function definitions are taken to be the basic ones, then (143) is obtained from (142) by replacing m by $-m$, as the abbreviated notations suggest.

$$\begin{aligned}\Delta k^{(-m)} &= \frac{1}{(k+2)(k+3)\cdots(k+m)(k+m+1)} \\ &\qquad - \frac{1}{(k+1)(k+2)(k+3)\cdots(k+m)} \\ &= \left(\frac{1}{k+m+1} - \frac{1}{k+1}\right)\frac{1}{(k+2)(k+3)\cdots(k+m)} \\ &= \frac{-m}{(k+1)(k+2)\cdots(k+m+1)} \\ &= -m\,k^{(-m-1)}. \end{aligned} \tag{144b}$$

Thus we see that $k^{(m)}$ is a polynomial in k of degree m, related to the difference operator Δ just as the power x^m is related to the differential operator d/dx, and that $k^{(-m)}$ as defined in (143) is analogous in the same sense to the inverse power x^{-m}. These polynomials are sometimes known as *factorial powers* of k.

If m is *nonintegral*, then the definition involving the Gamma function is to be used. In this more general case, it is readily verified that (144a) and (144b) are still true for any nonnegative value of m. We may notice that $k^{(0)} = 1$.

By making use of (126), we may convert the difference formulas (144a,b) to the summation formulas

$$\sum_{k=M}^{N} k^{(m)} = \frac{1}{m+1}\sum_{k=M}^{N} \Delta k^{(m+1)} = \left[\frac{k^{(m+1)}}{m+1}\right]_M^{N+1} \tag{145a}$$

and

$$\sum_{k=M}^{N} k^{(-m)} = \frac{1}{-m+1}\sum_{k=M}^{N} \Delta k^{(-m+1)} = \left[\frac{k^{(-m+1)}}{-m+1}\right]_M^{N+1} \qquad (m \neq 1). \tag{145b}$$

These formulas can also be written in the explicit forms

$$\sum_{k=M}^{N} k(k-1)\cdots(k-m+1) = \left[\frac{k(k-1)\cdots(k-m)}{m+1}\right]_M^{N+1}, \tag{146a}$$

$$\sum_{k=M}^{N} \frac{1}{(k+1)(k+2)\cdots(k+m)} = -\frac{1}{m-1}\left[\frac{1}{(k+1)(k+2)\cdots(k+m-1)}\right]_M^{N+1} \qquad (m \neq 1). \tag{146b}$$

More generally, if we consider any *linear* function of k,

$$f_k = ak + b, \tag{147}$$

where a and b are constants, and write

$$\left.\begin{aligned} f_k^{(m)} &= f_k f_{k-1} \cdots f_{k-m+1}, \\ f_k^{(-m)} &= \frac{1}{f_{k+1} f_{k+2} \cdots f_{k+m}} \end{aligned}\right\}, \tag{148a,b}$$

we may verify that, in consequence of the fact that $\triangle f_k = a =$ constant, equations (144a,b) generalize to the forms

$$\triangle f_k^{(m)} = a\, m\, f_k^{(m-1)}, \qquad \triangle f_k^{(-m)} = -a\, m\, f_k^{(-m-1)}, \tag{149a,b}$$

and, similarly, equations corresponding to (145a,b) are modified only in that the right-hand members are divided by the constant a. Thus, the generalizations of (146a,b) take the form

$$\sum_{k=M}^{N} f_k f_{k-1} \cdots f_{k-m+1} = \left[\frac{f_k f_{k-1} \cdots f_{k-m}}{a(m+1)}\right]_M^{N+1} \tag{150a}$$

and

$$\sum_{k=M}^{N} \frac{1}{f_{k+1} f_{k+2} \cdots f_{k+m}} = -\frac{1}{a(m-1)}\left[\frac{1}{f_{k+1} f_{k+2} \cdots f_{k+m-1}}\right]_M^{N+1} \qquad (m \neq 1). \tag{150b}$$

To illustrate the use of these formulas, we notice that the sum

$$S_n = 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + \cdots + n(n+1) \tag{151a}$$

is of the form (145a), with $m = 2$, $M = 2$, and $N = n + 1$, and hence there follows

$$S_n = \sum_{2}^{n+1} k^{(2)} = \left[\frac{k^{(3)}}{3}\right]_2^{n+2} = \frac{(n+2)(n+1)n}{3} - \frac{2 \cdot 1 \cdot 0}{3} = \tfrac{1}{3}n(n+1)(n+2). \tag{151b}$$

The sum

$$S_n = \frac{1}{1 \cdot 4 \cdot 7} + \frac{1}{4 \cdot 7 \cdot 10} + \cdots + \frac{1}{(3n-2)(3n+1)(3n+4)} \tag{152a}$$

is of the form (150b) with $f_k = 3k - 5$, $m = 3$, $M = 1$, and $N = n$, and hence there follows

$$S_n = -\frac{1}{3 \cdot 2}\left[\frac{1}{(3k-2)(3k+1)}\right]_1^{n+1} = \frac{1}{24} - \frac{1}{6}\frac{1}{(3n+1)(3n+4)}. \tag{152b}$$

We may notice that, as $n \to \infty$, the sum becomes a convergent *infinite series*, with the sum $\frac{1}{24}$.

The preceding results can also be used in connection with the methods of partial fractions in certain other more general cases. To illustrate this fact, we consider the sum

$$S_n = \frac{1}{2 \cdot 4} + \frac{1}{3 \cdot 5} + \frac{1}{4 \cdot 6} + \cdots + \frac{1}{(n+1)(n+3)}. \tag{153a}$$

The kth term $1/(k+1)(k+3)$ cannot be put in the form $1/f_{k+1}f_{k+2}$. However, if we write

$$\begin{aligned}\frac{1}{(k+1)(k+3)} &= \frac{k+2}{(k+1)(k+2)(k+3)} \\ &= \frac{1}{2}\left[\frac{(k+1)+(k+3)}{(k+1)(k+2)(k+3)}\right] \\ &= \frac{1}{2}\left[\frac{1}{(k+1)(k+2)} + \frac{1}{(k+2)(k+3)}\right],\end{aligned}$$

each term in brackets in the last member is of this form, and we obtain, by the preceding methods,

$$2S_n = \sum_1^n k^{(-2)} + \sum_2^{n+1} k^{(-2)} = \frac{5}{6} - \frac{1}{1}\left(\frac{1}{n+2} + \frac{1}{n+3}\right),$$

and hence

$$\frac{1}{2 \cdot 4} + \frac{1}{3 \cdot 5} + \cdots + \frac{1}{(n+1)(n+3)} = \frac{5}{12} - \frac{2n+5}{2(n+2)(n+3)}. \tag{153b}$$

As $n \to \infty$ we have the additional result

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)(k+3)} = \frac{5}{12}$$

It should be noticed that in (146b) and (150b) the case $m = 1$ is excluded. The sum

$$\sum_{1}^{n} \frac{1}{ak+b} = \frac{1}{a+b} + \frac{1}{2a+b} + \cdots + \frac{1}{na+b} \qquad (a \neq 0)$$

cannot be expressed in closed form in terms of elementary functions. However, it is expressible in terms of a certain *tabulated* function. To obtain the desired result, we notice first that from the relation

$$(1+\alpha)(2+\alpha)\cdots(n+\alpha) = \frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}$$

there follows, by logarithmic differentiation with respect to the parameter α,

$$\frac{1}{1+\alpha} + \frac{1}{2+\alpha} + \cdots + \frac{1}{n+\alpha} = \frac{\Gamma'(n+\alpha+1)}{\Gamma(n+\alpha+1)} - \frac{\Gamma'(\alpha+1)}{\Gamma(\alpha+1)}$$

The so-called *Psi function*, defined by the relation

$$\Psi(x) = \frac{\Gamma'(x+1)}{\Gamma(x+1)}, \tag{154}$$

is a tabulated function.* With this notation, the preceding result takes the form

$$\frac{1}{1+\alpha} + \frac{1}{2+\alpha} + \cdots + \frac{1}{n+\alpha} = \Psi(n+\alpha) - \Psi(\alpha).$$

More generally, if both sides of this equation are divided by a constant a, and α is replaced by b/a, there follows finally

$$\sum_{1}^{n} \frac{1}{ak+b} = \frac{1}{a+b} + \frac{1}{2a+b} + \cdots + \frac{1}{na+b}$$

$$= \frac{1}{a}\left[\Psi\left(n+\frac{b}{a}\right) - \Psi\left(\frac{b}{a}\right)\right]. \tag{155}$$

* In some references, the definition (154) is replaced by the definition $\Psi(x) = \Gamma'(x)/\Gamma(x)$.

The number $-\Psi(0)$ is known as *Euler's constant,* and is often denoted by γ,

$$\Psi(0) = \Gamma'(1) = -\gamma = -0.5772157 \cdots . \tag{156}$$

Thus, in particular, the result of setting $a = 1$ and $b = 0$ in (155) is of the form

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} = \Psi(n) + \gamma. \tag{157}$$

3.9. Characteristic-value problems. Any linear homogeneous difference equation of second order can be put into the form

$$\triangle(p_{k-1} \triangle y_{k-1}) + s_k y_k = 0 \tag{158}$$

by a suitable choice of the functions p_k and s_k. This form is particularly convenient for the purposes of this section, and is the form to which the formulation of many physical problems leads in a natural way [see equation (122b)]. It is also expressible in the more symmetrical form

$$\nabla(p_k \triangle y_k) + s_k y_k = 0, \tag{158'}$$

as well as in the expanded form

$$p_k y_{k+1} - (p_k + p_{k-1}) y_k + p_{k-1} y_{k-1} + s_k y_k = 0. \tag{158''}$$

We suppose now that the coefficient s_k is expressed in the form $q_k + \lambda r_k$, where λ is a parameter which may take on different constant values in a given problem. Equation (158) then takes the form

$$\triangle(p_{k-1} \triangle y_{k-1}) + (q_k + \lambda r_k) y_k = 0. \tag{159}$$

As in the case of the analogous differential equations, and as in the second example of Section 3.6, we speak of the problem consisting of (159) and homogeneous boundary conditions prescribed for two *different* integral values of k as a *homogeneous boundary-value problem.* In such a problem, no nontrivial solution exists, in general, unless λ takes on one of a set of *characteristic values* λ_1, $\lambda_2, \ldots$, whereas if this is the case, say $\lambda = \lambda_n$, the conditions of the problem are satisfied by an expression of the form $y_k = C \phi_{n,k}$ where C is an arbitrary constant. The function $\phi_{n,k}$ is known as the *characteristic function* corresponding to λ_n. We show next that such functions possess properties analogous to those associated with

the corresponding functions which arise in the solution of linear differential equations and sets of linear algebraic equations (see Chapter 1).

For this purpose, suppose that the origin has been so chosen that (159) holds for $k = 1, 2, \ldots, N$, and that suitable homogeneous conditions (yet to be specified) are prescribed when $k = 0$ and $k = N + 1$. Let λ_m and λ_n be two distinct characteristic values of λ, with corresponding characteristic functions $\phi_{m,k}$ and $\phi_{n,k}$. Then the two equations

$$\left.\begin{aligned} \Delta(p_{k-1}\,\Delta\phi_{m,k-1}) + (q_k + \lambda_m r_k)\phi_{m,k} &= 0, \\ \Delta(p_{k-1}\,\Delta\phi_{n,k-1}) + (q_k + \lambda_n r_k)\phi_{n,k} &= 0 \end{aligned}\right\} \tag{160a,b}$$

are satisfied. If we multiply (160a) by $\phi_{n,k}$ and (160b) by $\phi_{m,k}$, and subtract the respective results, we obtain the relation

$$\begin{aligned}(\lambda_m - \lambda_n) r_k \phi_{m,k}\phi_{n,k} \\ = \phi_{m,k}\,\Delta(p_{k-1}\,\Delta\phi_{n,k-1}) - \phi_{n,k}\,\Delta(p_{k-1}\,\Delta\phi_{m,k-1}).\end{aligned} \tag{161}$$

By summing the equal members of (161) with respect to k from $k = 1$ to $k = N$, there follows also

$$\begin{aligned}(\lambda_m - \lambda_n) \sum_{k=1}^{N} r_k \phi_{m,k}\phi_{n,k} \\ = \sum_{k=1}^{N} \phi_{m,k}\,\Delta(p_{k-1}\,\Delta\phi_{m,k-1}) - \sum_{k=1}^{N} \phi_{n,k}\,\Delta(p_{k-1}\,\Delta\phi_{m,k-1}).\end{aligned} \tag{162}$$

The sums on the right can be transformed, by the formula (128) for summation by parts, to the form

$$\begin{aligned}\Big[\phi_{m,k}\,p_{k-1}\,\Delta\phi_{n,k-1}\Big]_{k=1}^{k=N+1} - \sum_{k=1}^{N} p_k\,\Delta\phi_{m,k}\,\Delta\phi_{n,k} \\ - \Big[\phi_{n,k}\,p_{k-1}\,\Delta\phi_{m,k-1}\Big]_{k=1}^{k=N+1} + \sum_{k=1}^{N} p_k\,\Delta\phi_{n,k}\,\Delta\phi_{m,k}.\end{aligned}$$

The two sums in this last expression cancel identically, and the summed parts can be combined to give either of the equivalent forms

$$(\lambda_m - \lambda_n) \sum_{k=1}^{N} r_k \phi_{m,k}\phi_{n,k} = \begin{cases} \Big[p_{k-1}(\phi_{m,k}\,\Delta\phi_{n,k-1} - \phi_{n,k}\,\Delta\phi_{m,k-1})\Big]_{k=1}^{k=N+1} \\ \Big[p_{k-1}(\phi_{n,k}\phi_{m,k-1} - \phi_{m,k}\phi_{n,k-1})\Big]_{k=1}^{k=N+1} \end{cases} \tag{163}$$

The second form is merely an expansion of the first one. From a consideration of this result, it follows that the sum (163) vanishes, in particular, if the prescribed boundary conditions are of the form

$$\left.\begin{aligned} y_0 = 0 \quad \text{or} \quad p_0 \,\Delta y_0 = 0 \quad \text{or} \quad y_0 + \alpha \,\Delta y_0 = 0, \\ y_{N+1} = 0 \quad \text{or} \quad p_N \,\Delta y_N = 0 \quad \text{or} \quad y_{N+1} + \alpha \,\Delta y_N = 0 \end{aligned}\right\}. \tag{164}$$

Further, the condition at the lower limit of the range may be omitted if $p_0 = 0$, while the condition at the upper limit may be omitted if $p_N = 0$.

In analogy with the terminology of Chapter 1, we say that two functions f_k and g_k are *orthogonal* over the range $k = 1, 2, \ldots, N$ if

$$\sum_1^N f_k g_k = 0, \tag{165a}$$

and are *orthogonal relative to the weighting function* r_k if

$$\sum_1^N r_k f_k g_k = 0. \tag{165b}$$

From the preceding results, we conclude that *two characteristic functions of the difference equation* (159), *satisfying the same homogeneous boundary conditions for* $k = 0$ *and* $k = N + 1$, *and corresponding to distinct characteristic values of* λ, *are orthogonal over the range* $k = 1, 2, \ldots, N$ *relative to the weighting function* r_k. In particular, if the coefficient of $\lambda\, y_k$ in (159) is *unity*, the characteristic functions are *simply* orthogonal, with weighting function unity.

In physical problems, the functions p_k and r_k are *positive* over the range $k = 1, 2, \ldots, N$. As is shown in the following section, the problem then leads to N *real* characteristic numbers, and to a corresponding orthogonal set of N characteristic functions which are *linearly independent* in the sense that no nontrivial linear combination of these N functions vanishes for each of the N relevant values of k.* In consequence of this fact, any function f_k which is defined for $k = 1, 2, \ldots, N$ can be expressed as a linear combination of

* In unusual cases, the characteristic numbers may not be distinct, and two or more characteristic functions may then correspond to the same characteristic number. These functions can be orthogonalized by the Schmidt procedure of Section 1.12.

the N characteristic functions,

$$f_k = \sum_{n=1}^{N} A_n \phi_{n,k} \qquad (k = 1, 2, \cdots, N). \tag{166}$$

By this statement we mean that the N constants of combination can be determined so that the linear combination takes on the same values as f_k at the N relevant points.

To evaluate the constants, we may require that the difference between the two members of (166) be orthogonal (relative to r_k) to each characteristic function, over the relevant range in k. Thus, if we multiply both sides of (166) by $r_k \phi_{m,k}$ and sum the results, we obtain the condition

$$\sum_{k=1}^{N} r_k f_k \phi_{m,k} = \sum_{n=1}^{N} A_n \left(\sum_{k=1}^{N} r_k \phi_{m,k} \phi_{n,k} \right).$$

But, in view of the orthogonality of the system, all the inner sums on the right are zero except that one for which $m = n$, and we obtain the equation

$$A_n \sum_{k=1}^{N} r_k \phi_{n,k}^2 = \sum_{k=1}^{N} r_k f_k \phi_{n,k} \qquad (n = 1, 2, \cdots, N) \tag{167}$$

which determines each coefficient in (166) as the ratio of two calculable sums.

More generally, if we think of k as a *continuous variable*, so that $f(k)$ is defined, say, in the *interval* $(0, N + 1)$, then the right-hand member of (166) affords an *approximation* to f, in the sense that, with the coefficients given by (167), this function agrees with f at the N points of the domain $(k = 1, 2, \ldots, N)$ for which the generating difference equation is valid.

3.10. Matrix notation. In this section, we investigate the relationship between the discussion of the preceding section and the corresponding discussions in Chapter 1. We again consider the difference equation (158), and write it, for present purposes, in the expanded form (158″):

$$p_k y_{k+1} - (p_k + p_{k-1}) y_k + p_{k-1} y_{k-1} + (q_k + \lambda r_k) y_k = 0. \tag{168}$$

If we introduce the abbreviation

$$a_k = p_k + p_{k-1} - q_k, \tag{169}$$

this equation can be written also in the form

$$-p_{k-1}y_{k-1} + a_k y_k - p_k y_{k+1} = \lambda r_k y_k \quad (k = 1, 2, \cdots, N). \tag{170}$$

If these equations are written out explicitly, they take the form

$$\begin{aligned}
-p_0 y_0 + a_1 y_1 - p_1 y_2 &= \lambda r_1 y_1, \\
-p_1 y_1 + a_2 y_2 - p_2 y_3 &= \lambda r_2 y_2, \\
-p_2 y_2 + a_3 y_3 - p_3 y_4 &= \lambda r_3 y_3, \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
-p_{N-1} y_{N-1} + a_N y_N - p_N y_{N+1} &= \lambda r_N y_N.
\end{aligned} \tag{171}$$

In addition, two homogeneous boundary conditions are to be prescribed. The possible forms listed in equation (164) are comprised by the conditions

$$y_0 = \mu_1 y_1, \qquad \mu_2 y_N = y_{N+1}, \tag{172a,b}$$

where μ_1 and μ_2 are appropriately chosen constants.

These two conditions permit the elimination of y_0 and y_{N+1} from equations (171), leading to a homogeneous set of N linear algebraic equations in the N unknown quantities $y_1, y_2, \cdots, y_N$. If, in accordance with the terminology of Chapter 1, we consider these quantities as the components of an N-dimensional vector $\mathbf{y}$, the resultant set of equations can be written in the matrix form

$$\mathbf{a}\,\mathbf{y} = \lambda\,\mathbf{r}\,\mathbf{y}, \tag{173}$$

where the matrix $\mathbf{a}$ is of the form

$$\mathbf{a} = \begin{bmatrix}
a_1' & -p_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
-p_1 & a_2 & -p_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & -p_2 & a_3 & -p_3 & \cdots & 0 & 0 & 0 \\
0 & 0 & -p_3 & a_4 & \cdots & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & \cdots & -p_{N-2} & a_{N-1} & -p_{N-1} \\
0 & 0 & 0 & 0 & \cdots & 0 & -p_{N-1} & a_N'
\end{bmatrix}. \tag{174}$$

with

$$a_1' = a_1 - \mu_1 p_0, \qquad a_N' = a_N - \mu_2 p_N, \tag{175}$$

and where $\mathbf{r}$ is a *diagonal* matrix, $[r_i\, \delta_{ij}]$.

We notice that $\mathbf{a}$ is a symmetric matrix, and that the requirement that $r_1, r_2, \ldots, r_N$ be *positive* leads to the fact that the matrix $\mathbf{r}$ is *positive definite.* Hence the results of Section 1.25 are directly applicable, and the statements made in the preceding section can be established.

It is clear, from the form of (171), that if $p_0 = 0$ the quantity y_0 is not involved in those equations, so that a homogeneous relation (172a) is then not needed. In this case, there follows $a_1' = a_1$ in (174). A similar statement applies in the case when $p_N = 0$, as was discovered by a different approach in the preceding section.

3.11. The vibrating loaded string. In Section 3.6, it was shown that for small free vibrations of a string of length $L = (N+1)h$, with beads of equal mass M attached at the points $x = h, 2h, \ldots, Nh$, and with fixed ends, the deflection y_k at a point $x_k = kh$ may be compounded from N normal modes, in the form

$$y_k = \sum_{n=1}^{N} C_n \sin \frac{n\pi k}{N+1} \cos(\omega_n t + \beta_n), \tag{176}$$

where

$$\omega_n = 2\sqrt{\frac{T}{Mh}} \sin \frac{n\pi}{2(N+1)}. \tag{177}$$

In consequence of the results of Section 3.9, the amplitude functions

$$\phi_{n,k} = \sin \frac{n\pi k}{N+1} \tag{178}$$

are orthogonal over the range $k = 1, 2, \ldots, N$, with weighting function unity:

$$\sum_{k=1}^{N} \sin \frac{m\pi k}{N+1} \sin \frac{n\pi k}{N+1} = 0 \qquad (m \neq n). \tag{179}$$

This result follows from the fact that (111a) is identified with (159)

by taking $p_k = r_k = 1$ and $q_k = 0$, and can be independently verified by making use of equation (135a) of Section 3.8.

Also, in the case when $m = n$, equation (137a) gives the result

$$\sum_{k=1}^{N} \sin^2 \frac{n\pi k}{N+1} = \frac{N}{2} - \frac{\sin \dfrac{nN\pi}{N+1} \cos n\pi}{2 \sin \dfrac{n\pi}{N+1}}$$

or, after an elementary reduction,

$$\sum_{k=1}^{N} \sin^2 \frac{n\pi k}{N+1} = \frac{N+1}{2}. \tag{180}$$

To complete the solution of the physical problem, it is necessary to determine the $2N$ constants C_n and β_n so that the prescribed initial *deflections* and *velocities* of the N beads are assumed by the solution when $t = 0$. We denote these prescribed values as follows:

$$y_k\Big|_{t=0} = d_k, \qquad \frac{\partial y_k}{\partial t}\Big|_{t=0} = v_k. \tag{181a,b}$$

The requirement that (176) satisfy these conditions then takes the form

$$\left.\begin{aligned} d_k &= \sum_{n=1}^{N} (C_n \cos \beta_n) \sin \frac{n\pi k}{N+1}, \\ v_k &= \sum_{n=1}^{N} (-\omega_n C_n \sin \beta_n) \sin \frac{n\pi k}{N+1} \end{aligned}\right\}$$

$$(k = 1, 2, \cdots, N). \tag{182a,b}$$

If equations (182a,b) are compared with (166), equation (167) shows that the constants must satisfy the equations

$$C_n \cos \beta_n \left(\sum_{k=1}^{N} \sin^2 \frac{n\pi k}{N+1} \right) = \sum_{k=1}^{N} d_k \sin \frac{n\pi k}{N+1},$$

$$-\omega_n C_n \sin \beta_n \left(\sum_{k=1}^{N} \sin^2 \frac{n\pi k}{N+1} \right) = \sum_{k=1}^{N} v_k \sin \frac{n\pi k}{N+1},$$

and hence, making use of (180), we may write

$$A_n \equiv C_n \cos \beta_n = \frac{2}{N+1} \sum_{k=1}^{N} d_k \sin \frac{n\pi k}{N+1}, \tag{183a}$$

$$B_n \equiv -C_n \sin \beta_n = \frac{2}{\omega_n (N+1)} \sum_{k=1}^{N} v_k \sin \frac{n\pi k}{N+1}. \tag{183b}$$

With this notation, the required solution (176) can be written in the form

$$y_k = \sum_{n=1}^{N} \sin \frac{n\pi k}{N+1} (A_n \cos \omega_n t + B_n \sin \omega_n t). \tag{184}$$

As a simple example, we consider the case when only *two* masses are present ($h = L/3$), and suppose that at the instant $t = 0$ the first mass is released *from rest* with an initial deflection d, while the second mass is initially *at rest* in an *undeflected* position. In this case we have the following data:

$$N = 2; \quad d_1 = d, \quad d_2 = 0; \quad v_1 = v_2 = 0. \tag{185}$$

Equations (183a,b) then give the results

$$A_1 = \frac{2}{3} d \sin \frac{\pi}{3} = \frac{d}{\sqrt{3}}, \quad A_2 = \frac{2}{3} d \sin \frac{2\pi}{3} = A_1, \quad B_1 = B_2 = 0,$$

and the solution (184) takes the form

$$y_k = \frac{d}{\sqrt{3}} \left(\sin \frac{\pi k}{3} \cos \omega_1 t + \sin \frac{2\pi k}{3} \cos \omega_2 t \right) \quad (k = 1, 2), \tag{186}$$

where, in accordance with (177),

$$\omega_1 = \sqrt{\frac{T}{Mh}}, \qquad \omega_2 = \sqrt{\frac{3T}{Mh}}. \tag{187}$$

By setting k successively equal to 1 and 2, the displacements of the two masses at any time t are thus obtained in the form

$$y_1 = \frac{d}{2} (\cos \omega_1 t + \cos \omega_2 t), \quad y_2 = \frac{d}{2} (\cos \omega_1 t - \cos \omega_2 t). \tag{188}$$

The loaded string has been chosen as a model, in the illustrations of the preceding theory of this chapter, so that the various *types* of problems which most frequently arise in many other fields may be motivated and investigated as simply as possible.

3.12. Linear equations with variable coefficients. The general homogeneous linear difference equation of the *first order* can be written in the form

$$y_{k+1} - a_k y_k = 0. \tag{189}$$

If (189) is valid when $k = 0$, we may obtain successively the results

$$y_1 = y_0 a_0, \quad y_2 = y_0 a_0 a_1, \quad y_3 = y_0 a_0 a_1 a_2,$$

and hence, by induction,

$$y_k = y_0(a_0 a_1 a_2 \cdots a_{k-1}). \tag{190}$$

The coefficient of y_0 in (190) is in the form of a *product* of k terms, and is conventionally abbreviated in the form

$$\prod_{n=0}^{k-1} a_n \equiv \prod_{n=1}^{k} a_{n-1} \equiv a_0 a_1 a_2 \cdots a_{k-1}. \tag{191}$$

If we notice that any fixed number of factors in (190) could be incorporated with y_0, to form a new arbitrary constant, it follows that *the general solution of the equation*

$$y_{k+1} - a_k y_k = 0 \quad \text{or} \quad (E - a_k)y_k = 0 \tag{192}$$

is of the form

$$y_k = C \prod^k a_{n-1}, \tag{193}$$

with the convention that *the symbol* $\prod^k$ *indicates the formation of the product of those factors for which the relevant dummy index takes on all integral values from some permissible fixed integral lower limit to the variable upper limit* k.

In illustration, we consider the equation

$$(k+1)y_{k+1} + (k+2)y_k = (k+1)(k+2) \qquad (k = 1, 2, \cdots). \tag{194}$$

With $a_k = -(k+2)/(k+1)$, equation (193) leads to the general homogeneous solution

$$y_k^{(H)} = C \prod_1^k \left(-\frac{n+1}{n} \right)$$

$$= C(-1)^k \frac{2}{1} \cdot \frac{3}{2} \cdot \cdots \frac{k}{k-1} \cdot \frac{k+1}{k}$$

$$= (-1)^k C(k+1).$$

With this result, the method of variation of parameters [equation (75)] leads to a particular solution of the nonhomogeneous equation, in the form

$$y_k^{(P)} = (-1)^k (k+1) \sum_1^k (-1)^n \frac{n+1}{n+1}$$

$$= (-1)^k (k+1)[-1 + 1 - \cdots + (-1)^k].$$

The sum in brackets is zero if k is even, and -1 if k is odd, and hence is conveniently expressible in the form $(\cos k\pi - 1)/2$. Thus the complete solution of (194) can be written in the form

$$y_k = (k+1) \left(C + \frac{\cos k\pi - 1}{2} \right) \cos k\pi$$

or, with $C = (c+1)/2$,

$$y_k = \frac{k+1}{2} (c \cos k\pi + 1) \qquad (k = 1, 2, \cdots). \tag{195}$$

No general method exists for solving linear difference equations of higher order (with variable coefficients) in terms of finite sums and products. While infinite series solutions (involving *factorial* powers of k) can be obtained in many cases,* they are of limited usefulness and are not discussed here. Instead, three special methods of occasional usefulness are outlined.

a. *Reduction of Order.* In case *one* homogeneous solution, say $y_k^{(H)} = u_k$, can be found by inspection or otherwise, an equation of lower order can be obtained for the determination of a second homogeneous solution. For if we write $v_k = y_k/u_k$ and attempt to determine v_k, the resultant equation must be satisfied by $v_k =$ constant or $\Delta v_k = 0$. Hence the equation will be of reduced order if

*See Reference 1.

the new unknown $V_k = \triangle v_k = \triangle(y_k/u_k)$ is introduced. In particular, if $y_k =$ constant is a homogeneous solution, the order of the equation is reduced by the substitution $V_k = \triangle y_k$.

b. *Factorization.* If a linear difference equation can be written in such a form that the relevant difference operator is factored, say, in the form

$$(E - b_k)(E - a_k)y_k = \phi_k, \tag{196}$$

then the general solution can be obtained by solving two *successive* equations of the first order, since (196) is equivalent to the two simultaneous equations

$$(E - b_k)u_k = \phi_k \quad \text{or} \quad u_{k+1} - b_k u_k = \phi_k \tag{197a}$$

and

$$(E - a_k)y_k = u_k \quad \text{or} \quad y_{k+1} - a_k y_k = u_k. \tag{197b}$$

Thus (197a) first determines u_k, after which y_k is determined from (197b).

c. *Substitution.* In some cases it is possible to rearrange an equation so that it takes the form

$$a\, f_{k+2}y_{k+2} + b\, f_{k+1}y_{k+1} + c\, f_k y_k = \phi_k, \tag{198a}$$

where a, b, and c are constants. The substitution $u_k = f_k y_k$ obviously reduces the equation to one with constant coefficients. Similarly, the substitution $u_k = y_k/f_k$ reduces the equation

$$a\, f_k f_{k+1}y_{k+2} + b\, f_k f_{k+2}y_{k+1} + c\, f_{k+1} f_{k+2}y_k = \phi_k \tag{198b}$$

to such a form. In illustration, the substitution $u_k = y_k/(k+1)$ is seen to be appropriate in dealing with (194).

3.13. Approximate solution of ordinary differential equations. In the remainder of this chapter, we depart from the consideration of the possibility of obtaining *explicit* solutions of difference equations, and indicate applications of the fact that they can be solved by step-by-step methods or considered as sets of simultaneous linear algebraic equations. The principal applications to be treated are related to problems governed by *partial differential equations*. However, in order to motivate the basic procedures, we consider first analogous problems governed by *ordinary* differential equations.

As a first example, we consider the simple problem of determining the solution of the differential equation

$$\frac{d^2y}{dx^2} = x \qquad (x > 0) \tag{199}$$

which satisfies the *initial conditions*

$$y(0) = 0, \qquad y'(0) = 1. \tag{200}$$

It will be seen that the methods to be used are readily generalized to the treatment of more involved problems.

In order to obtain an *approximate* solution to the problem, we first replace the differential equation by a finite difference approximation, the simplest of which is of the form

$$\frac{y(x+h) - 2y(x) + y(x-h)}{h^2} = x. \tag{201}$$

Accordingly, we replace the initial conditions by the requirements

$$y(0) = 0, \qquad \frac{y(h) - y(0)}{h} = 1. \tag{202}$$

As the increment h tends to zero, the new problem tends to the original one, and it is reasonable to expect that the *solution* of the new problem tends, at the same time, to the *solution* of the original one. Thus it may be expected that, for sufficiently small values of h, the solution of (201) which satisfies (202) will afford a satisfactory approximation to the required solution.

If we require that (201) be satisfied at the successive points $x_1 = h$, $x_2 = 2h$, . . . , $x_k = kh$, . . . , and notice that (202) determines y at the initial points $x_0 = 0$ and $x_1 = h$, the difference equation can be written in the form

$$y_{k+1} - 2y_k + y_{k-1} = k\,h^3 \qquad (k = 1, 2, \cdots), \tag{203}$$

where $y_k \equiv y(x_k) \equiv y(kh)$, and the initial conditions (202) take the form

$$y_0 = 0, \qquad y_1 - y_0 = h. \tag{204}$$

While this problem can be solved explicitly by the methods of Sections 3.4 and 3.5, in the form

$$y_k = \left(h - \frac{h^3}{6}\right)k + \frac{h^3}{6}k^3, \tag{205}$$

this situation does not ordinarily occur. However, in the present case, we may notice that (205) is equivalent to the equation

$$y(x_k) = x_k + \frac{1}{6}x_k^3 - \frac{h^2}{6}x_k, \tag{205'}$$

while the exact solution of the original problem is of the form

$$y(x) = x + \tfrac{1}{6}x^3. \tag{206}$$

Thus it follows that the solution (205') does indeed tend to (206) as h tends to zero, and also that the ratio of the error associated with the approximation (205') to the exact value of the solution at any point x_k is less than $h^2/6$.

In the absence of such information, we would merely assign a convenient numerical value to h, say $h = 0.1$, and generate approximate values of y at the chosen points by step-by-step calculation. Thus, from (204), we then obtain

$$y_0 = 0, \qquad y_1 = 0.1.$$

By writing (203) in the form

$$y_{k+1} = 2y_k - y_{k-1} + 0.001k,$$

there then follows

$$y_2 = 0.2 - 0 + 0.001 = 0.201,$$

$$y_3 = 0.402 - 0.1 + 0.002 = 0.304,$$

and so forth. An estimate of the accuracy attained would be afforded by comparing these results with the results of a second series of calculations, based on the halved spacing $h = 0.05$ or on the doubled spacing $h = 0.2$.

More efficient methods of integrating differential equations approximately, when *initial conditions* are prescribed, involve the use of differences of higher order, and may be found in the literature.* The preceding method was presented here principally for the reason that it is analogous to methods, discussed in following sections, which are of frequent use in the solution of certain initial-value problems associated with *partial* differential equations.

* See References 2 and 3.

As a second example, we consider the nonhomogeneous *boundary value* problem consisting of the differential equation

$$\frac{d^2y}{dx^2} + y = 0 \qquad (0 < y < 1), \tag{207}$$

and the *end* conditions

$$y(0) = 0, \quad y(1) = 1. \tag{208}$$

We replace this problem by the difference equation

$$y_{k+1} - 2y_k + y_{k-1} + h^2y_k = 0 \qquad (k = 1, 2, \cdots, N) \tag{209}$$

and the end conditions

$$y_0 = 0, \qquad y_{N+1} = 1, \tag{210}$$

where $y_k \equiv y(kh)$ and $(N + 1)h = 1$. Two possible procedures are now evident. *First*, we may determine the values $y_2, y_3, \cdots, y_N$ successively (as before), in terms of the specified value y_0 and the *unknown* value y_1, and determine y_1 finally in such a way that $y_{N+1} = 1$; that is, the value y_1 may be carried through the calculation as a literal parameter, and determined at the end of the calculation. *Alternatively*, we may treat (209) as a set of N linear algebraic equations in N unknown quantities, y_0 and y_{N+1} being given by (210), and solve this set of equations by any of several standard methods. This procedure is particularly well suited to the use of modern automatic calculators. For the sake of simplicity, we here take $N = 3$, so that $h = \frac{1}{4}$. The three relevant equations then take the form

$$\left.\begin{aligned} -\tfrac{31}{16}y_1 + \quad y_2 \qquad\quad &= 0, \\ y_1 - \tfrac{31}{16}y_2 + \quad y_3 &= 0, \\ y_2 - \tfrac{31}{16}y_3 &= -1 \end{aligned}\right\}, \tag{211}$$

from which the values $y_1 \doteq 0.2943$, $y_2 \doteq 0.5702$, and $y_3 \doteq 0.8104$ may be obtained as approximations to the required ordinates at $x = \frac{1}{4}, \frac{1}{2}$, and $\frac{3}{4}$. Approximate values of intermediate ordinates can be obtained by polynomial interpolation, or by merely plotting the calculated ordinates and joining them by a smooth curve. The true ordinates are found from the exact solution $y = (\sin x)/(\sin 1)$ to be $y_1 \doteq 0.2940$, $y_2 \doteq 0.5698$, and $y_3 \doteq 0.8102$. It happens, again, that in this case the explicit solution of the approximate

formulation can be obtained by the methods of Section 3.4, and convergence to the exact solution as $h \to 0$ can be explicitly established.

An iterative method, which avoids the direct solution of equations (211), is outlined in Section 3.19.

As a third example, we consider the *characteristic-value* problem consisting of the differential equation

$$\frac{d^2y}{dx^2} + \lambda y = 0 \qquad (0 < y < 1), \tag{212}$$

and the end values

$$y(0) = 0, \qquad y(1) = 0, \tag{213}$$

and which is accordingly replaced by the problem

$$y_{k+1} - 2y_k + y_{k-1} + \lambda h^2 y_k = 0 \qquad (k = 1, 2, \cdots, N) \tag{214}$$

with

$$y_0 = 0, \qquad y_{N+1} = 0. \tag{215}$$

The new problem then comprises a characteristic-value problem of the type considered in Section 1.11, and can be solved directly or by the methods of matrix iteration described in Section 1.23.

In particular, if we take $N = 3$, the equations corresponding to (214) take the form

$$\left.\begin{aligned} 2y_1 - y_2 &= \lambda h^2 y_1, \\ -y_1 + 2y_2 - y_3 &= \lambda h^2 y_2, \\ -y_2 + 2y_3 &= \lambda h^2 y_3 \end{aligned}\right\}. \tag{216}$$

The explicit solution of the problem consisting of (214) and (215) was obtained in Section 3.6, and is specified by equations (111a–d) with λ replaced by λh^2. Thus, with the present notation, when N interior points are chosen the problem determines N distinct characteristic values of λ, and N corresponding characteristic functions, which are of the following form:

$$\lambda_n = \frac{4}{h^2} \sin^2 \frac{n\pi h}{2}, \qquad \phi_{n,k} = \sin n\pi x_k \qquad (n = 1, 2, \cdots, N). \tag{217}$$

It is easily seen that, for small values of the spacing h, these results approximate the first N characteristic numbers and functions of the exact problem:

$$\lambda_n = n^2\pi^2, \qquad \phi_n(x) = \sin n\pi x \qquad (n = 1, 2, \cdots). \tag{218}$$

By noticing that $h = 1/(N+1)$, we may obtain from (217) the following table which indicates the rate of convergence to λ_n, as a function of the number N of interior points:

N \ n	1	2	3	4
1	8.00	—	—	—
2	9.00	27.00	—	—
3	9.37	32.00	54.63	—
. . .	. . .	. . .	. . .	. . .
10	9.80	38.40	83.49	141.4
. . .	. . .	. . .	. . .	. . .
20	9.87	39.26	87.51	153.6
. . .	. . .	. . .	. . .	. . .
∞	9.87	39.48	88.83	157.9

In the remainder of this chapter, we indicate the application of similar methods to the approximate solution of certain problems governed by *partial* differential equations.

3.14. The one-dimensional heat-flow equation. Transient heat flow in a homogeneous medium, in which the temperature T depends upon one rectangular coordinate x and upon time t, is governed by the partial differential equation

$$\frac{\partial T}{\partial t} = \alpha^2 \frac{\partial^2 T}{\partial x^2}, \tag{219}$$

where α^2 is a constant known as the thermal diffusivity of the conducting medium. This equation can be considered as the formal limit, as the increments h_x and h_t tend to zero, of the difference equation

$$\frac{T(x, t+h_t) - T(x, t)}{h_t} = \alpha^2 \frac{T(x+h_x, t) - 2T(x, t) + T(x-h_x, t)}{h_x^2}. \tag{220}$$

After a rearrangement, this relation can be put in the form

$$T(x, t+h_t) = \alpha^2 \frac{h_t}{h_x^2} T(x+h_x, t) + \left(1 - 2\alpha^2 \frac{h_t}{h_x^2}\right) T(x, t) + \alpha^2 \frac{h_t}{h_x^2} T(x-h_x, t). \tag{221}$$

This equation will retain its form as the increments h_x and h_t tend to zero if and only if h_t is taken to be proportional to h_x^2. A particularly convenient choice of the relationship between the two spacings is seen to be

$$h_x^2 = 2\alpha^2 h_t, \tag{222}$$

in consequence of which (221) takes the form

$$T(x, t + h_t) = \tfrac{1}{2}[T(x + h_x, t) + T(x - h_x, t)]. \tag{223}$$

This relation states that, within the framework of the approximate formulation, the temperature at a point x, at time $t + h_t$, is the *average* of the temperatures at the two *neighboring* points at the time t, if the spacings satisfy (222).

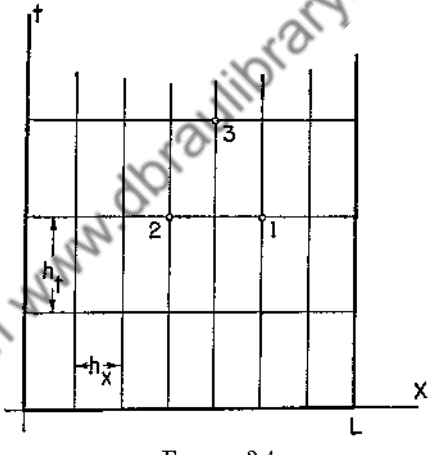


FIGURE 3.4

In order to illustrate the use of this formulation, we consider the solution of the problem in which, initially, the temperature distribution in a homogeneous rod of length L varies linearly from 100° at one end ($x = 0$) to 150° at the other end ($x = L$). At the instant $t = 0$, we suppose that the temperatures at the ends are suddenly reduced to 0° and maintained at that temperature thereafter. The resultant temperature distribution in the rod is then to be determined, as a function of distance x from one end and time t measured from the instant of change. It is convenient, for present purposes, to introduce a fictitious xt-plane (Figure 3.4), in which points corresponding to successive positions and times are indicated as the vertices of a network of squares or rectangles. If N interior division points are taken in the x-direction, this means that the spacing h_x is such that $(N + 1)h_x = L$, and the corresponding actual time increment h_t is then determined by (222). For the three numbered points in Figure 3.4, equation (223) gives the relation

$$T_3 = \tfrac{1}{2}(T_1 + T_2), \tag{224}$$

and since T is prescribed when $t = 0$ (and along the end boundaries $x = 0$ and $x = L$), successive use of this formula determines approximate values of T at following times.

If, for simplicity, we take $N = 4$, the following array is very easily obtained in this way:

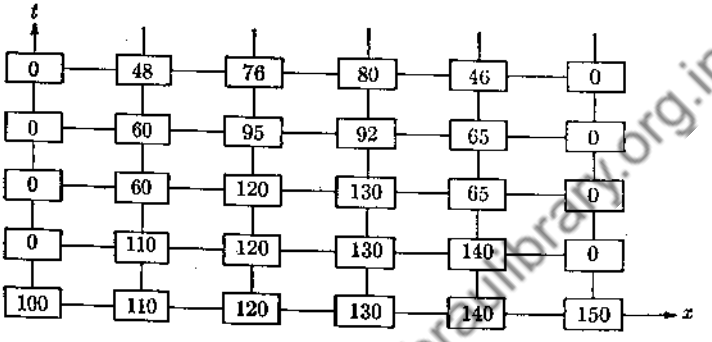


FIGURE 3.5

However, we may notice that here the temperatures predicted for $t = h_t$ are exactly those which are actually *prescribed* along the rod *immediately after* the change. Thus it appears that the approximate solution so obtained lags the exact solution by about one time interval. In fact, since the end temperatures are required to be zero throughout the first time interval (except *at* $t = 0$), it may be suggested that in the difference-equation formulation the initial values at the ends be taken to be zeros. Accordingly, the calculations in the present case would differ from those given in Figure 3.5 only in that the initial row of entries (at the foot of the diagram) would be deleted and the time origin would be moved upward by one unit.

Still, if the initial end values were indeed replaced by zeros, the error would clearly be overcorrected, since then we would obtain an approximate solution to a problem for which the initial temperature distribution, before the abrupt change, departs continuously from the originally prescribed one near the ends of the rod and vanishes at the two ends, the end temperatures then merely being *maintained* at zero thereafter.

The difficulty stems from the fact that here the prescribed limit of $T(x, t)$ as x and t tend to zero is 100 if the approach is made along the x-axis and 0 if the approach is made along the t-axis, and a

similar statement applies to the end $x = L$. A reasonable compromise between the two extremes suggested above consists in replacing the initial values at the ends by the *average* of the two limits approached in the x- and t-directions.

This modified procedure leads to the following calculated results:

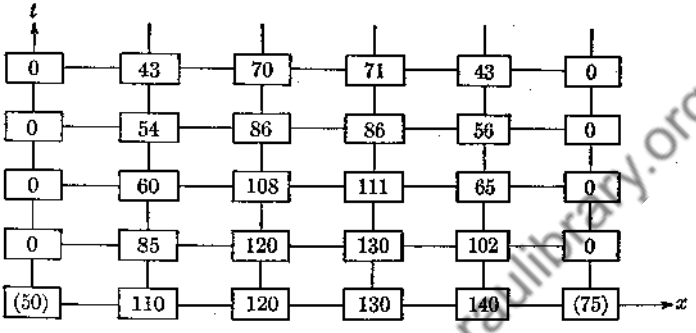


FIGURE 3.6

The exact solution of the problem considered can be obtained in the form

$$T(x, t) = \frac{100}{\pi} \sum_{n=1}^{\infty} \frac{2 - 3\cos n\pi}{n} \sin \frac{n\pi x}{L} e^{-\frac{n^2\pi^2\alpha^2 t}{L^2}},$$

from which true values corresponding to the preceding approximations are found as follows:

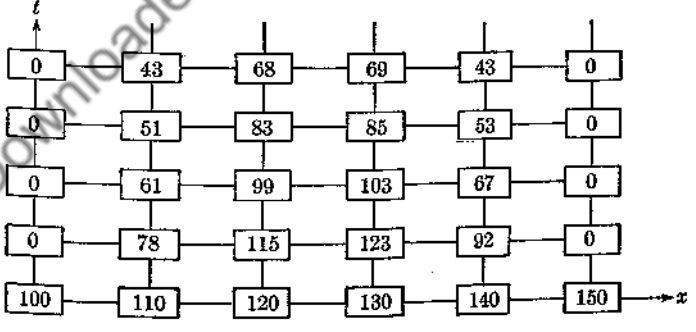


FIGURE 3.7

It is seen that, even with the use of only four intermediate division points, reasonably accurate results are obtained with very little labor in Figure 3.6.

For a rod of length $L = 1$ ft and diffusivity $\alpha^2 = 0.01$ sq ft per hr, there follows $h_x = 0.2$ ft, and from (222), we find that the time scale is then given by $h_t = 2$ hr.

In place of prescribing the temperature T at an end of a rod, one might prescribe the rate at which heat flows through that end. If this rate of heat flow is denoted by Q, the relevant end condition would then be of the form

$$\frac{dT}{dx} = -\frac{Q}{K A}, \tag{225a}$$

where K is the thermal conductivity of the material and A the cross-sectional area of the rod. Here Q is positive if the flow is in the positive x-direction. Thus, if heat were introduced into the end $x = 0$ at a prescribed rate Q, an end condition relevant to the difference-equation formulation would be of the form

$$T_1 - T_0 = -\frac{Q\,h}{K\,A} = h\left(\frac{dT}{dx}\right)_0. \tag{225b}$$

It is important to notice that no increase in complexity is introduced in the *approximate* calculation if we modify the problem in such a way that the prescribed temperature (or rate of heat loss) at an end of the rod *varies with time* in an arbitrarily specified manner.

3.15. The two-dimensional heat-flow equation. Transient flow of heat in a homogeneous, isotropic medium, in which the temperature depends upon two rectangular coordinates x and y, and upon time t, is governed by the differential equation

$$\frac{\partial T}{\partial t} = \alpha^2\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right). \tag{226}$$

In the usual problem, the temperature T is prescribed as a function of x and y over a two-dimensional region at the time $t = 0$, and the temperature distribution along the closed boundary is prescribed for all time $t > 0$. The resultant temperature distribution is then required as a function of x, y, and t.

If we introduce the increments h_x, h_y, and h_t, and proceed as in the preceding section, it is readily verified that when these incre-

ments satisfy the relation

$$h_x^2 = h_y^2 = 4\alpha^2 h_t \tag{227}$$

the approximating difference equation reduces to the convenient form

$$T(x, y, t + h_t) = \tfrac{1}{4}[T(x + h_x, y, t) + T(x - h_x, y, t) + T(x, y + h_y, t) + T(x, y - h_y, t)]. \tag{228}$$

Thus, if we consider the pattern of Figure 3.8, equation (228) states that, when the increments satisfy (227), the temperature at

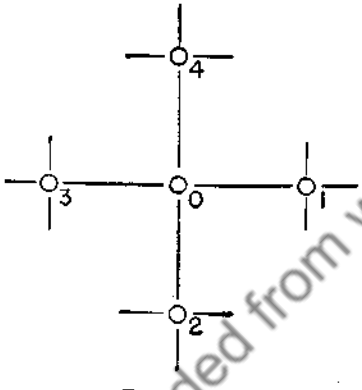


FIGURE 3.8

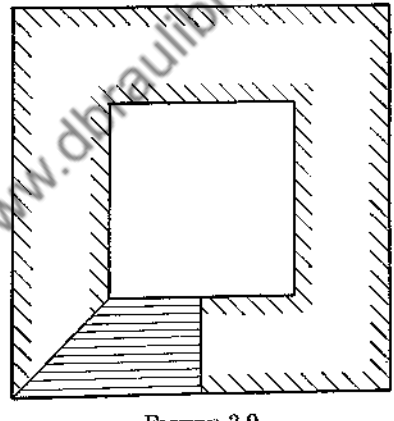


FIGURE 3.9

point 0, at time $t + h_t$, is merely the *average* of the temperatures at the four adjacent points 1, 2, 3, and 4 at the time t. This relation can be written in the abbreviated form

$$T_0\Big|_{t+h_t} = \frac{1}{4}(T_1 + T_2 + T_3 + T_4)\Big|_t. \tag{229}$$

In order to illustrate the numerical treatment of actual problems, we consider the region indicated in Figure 3.9. Initially, we suppose that all points in the region are at the temperature 100°. Then, we suppose that the temperatures along the inner boundary are abruptly raised to 300° and maintained at that temperature,

while the temperatures along the outer boundary are maintained at 100°. (As will be seen, the more interesting case in which the inner temperature is raised continuously, in a specified way, from 100° to 300° can be treated just as easily.) From the physical symmetry, it is clear that we may restrict attention to the shaded portion of the region indicated in Figure 3.9.

At the time $t = 0$, we take as the temperature along the inner boundary the average of the initial value 100° and the immediately following value 300°, so that the initial diagram appears as follows (with the indicated choice of division points):

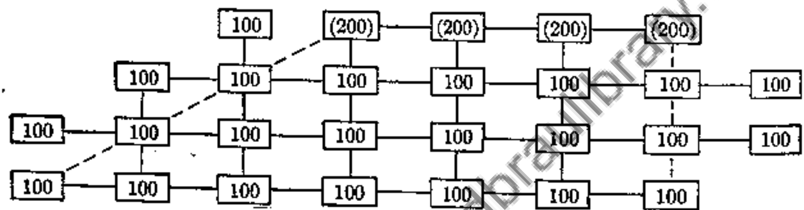


FIGURE 3.10

The diagonal line at the left, and the last vertical line at the right, are to be lines of symmetry, and the values at the indicated relevant points outside the region under consideration are to be determined accordingly.

The approximate temperature distribution after the time increment h_t, determined in terms of the physical spacing by (227), is obtained by the averaging process (229) as follows:

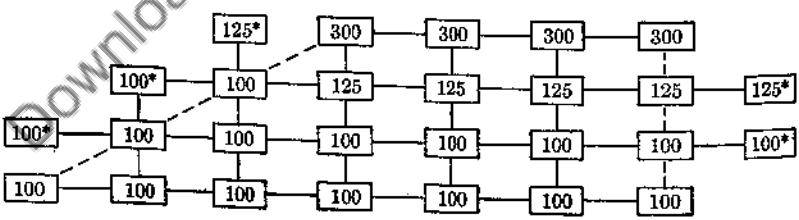


FIGURE 3.11

The starred values are obtained by symmetry, after the interior values have been obtained by the averaging process. The tem-

peratures along the inner (upper) boundary, from this stage onward, are given their true prescribed values. After a second averaging, the approximate distribution when $t = 2h_t$ is obtained as follows:

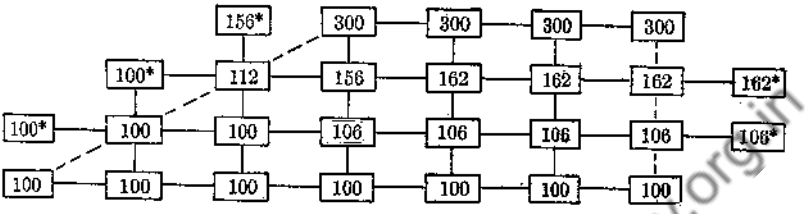


FIGURE 3.12

If the outside dimension of the region in Figure 3.9 were, say, 4 ft, the spacing $h_x = h_y$ would be $\frac{1}{3}$ ft, and, for a material of diffusivity $\alpha^2 = 0.01$ sq ft per hr, the corresponding time increment would be $h_t = 2.77$ hr.

It is of some interest to consider a physical basis for the actual difference-equation formulation of the heat-flow problem. For this purpose, the material of the relevant region can be considered as possessing two properties: that of heat *conduction* and that of heat *absorption*. The (uniform) thickness of the physical body, in the direction perpendicular to the planes in which heat flow occurs, is denoted by b. In place of studying the actual continuous body, we substitute for it a network of point masses interconnected by conducting rods (Figure 3.13), associating with each interior vertex point an effective mass $\rho\, h^2 b$, where ρ is the mass density of the material and h is the net spacing, and associating with each interior rod the effective cross-sectional area $h\, b$. Thus, in Figure 3.13, the material of the shaded square surrounding the point

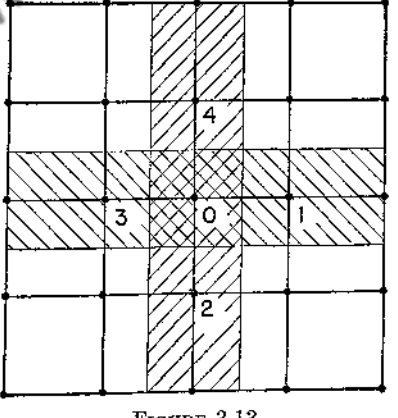


FIGURE 3.13

0 is associated with 0, in so far as heat absorption is concerned, while the four rods connecting 0 with adjacent points are considered to comprise the material in the respective shaded strips, in so far as heat conduction is concerned.

The rate at which heat is conducted in the positive x-direction, along any rod, is given by the quantity $-K\,A(\partial T/\partial x)$, where K is the thermal conductivity of the material and A is the cross-sectional area of the rod. In particular, in order to obtain the rate of heat flow from 1 to 0, we replace $\partial T/\partial x$ in the connecting rod by the constant value $(T_1 - T_0)/h$ and write $A = b\,h$, and so find that the rate of flow from 1 to 0 is given by $K\,b(T_1 - T_0)$. By considering the other rods leading to the point 0 in a similar way, we deduce that the rate Q_0 at which heat is being conducted to the point 0 at any instant is given by the expression

$$Q_0 = K\,b(T_1 + T_2 + T_3 + T_4 - 4T_0). \tag{230}$$

But also the rate of increase of the temperature of the mass $\rho\,b\,h^2$ associated with the point 0 is given by

$$\frac{\partial T_0}{\partial t} = \frac{Q_0}{s\,\rho\,b\,h^2},$$

where s is the specific heat of the material, and hence there follows (to a first approximation)

$$Q_0 = \frac{s\,\rho\,b\,h^2}{h_t}\left(T_0\Big|_{t+h_t} - T_0\Big|_t\right), \tag{231}$$

where h_t is a time spacing. If we recall that the diffusivity α^2 is defined by $\alpha^2 = K/s\rho$, and choose h_t in such a way that (227) is satisfied, the result of equating the right-hand members of (230) and (231) is precisely the difference equation (229).

These physical considerations are frequently useful in interpreting the results of the approximate analysis. Thus, for example, it is of interest to determine the approximate rate of heat loss (as a function of time) through the shaded region of Figure 3.9. With the approximate data of Figure 3.12, the rate of heat flow through the rods extending from the inner boundary of one-eighth of the

entire region after $2h_t$ units of time is given by

$$[(300 - 156) + (300 - 162) + (300 - 162) + \tfrac{1}{2}(300 - 162)]Kb = 489Kb.$$

Only one-half the flow through the right-hand rod is considered, since only a half-strip is associated with that rod in the region under consideration. However, the material to be associated with the rod extending from the inner corner to the point at temperature 156 in Figure 3.12 consists of the diagonal half-square at its left and the vertical half-square at its right, so that the flow through that rod receives full weighting. The (approximate) total rate of heat loss from the interior of the complete region is then eight times this result, or $3912Kb$, at the time $t = 2h_t$. Finally, the so-called *thermal resistance* at that instant, defined as the ratio of the constant temperature difference between the inner and outer faces to the total rate of heat loss, is given by $0.0512/Kb$.

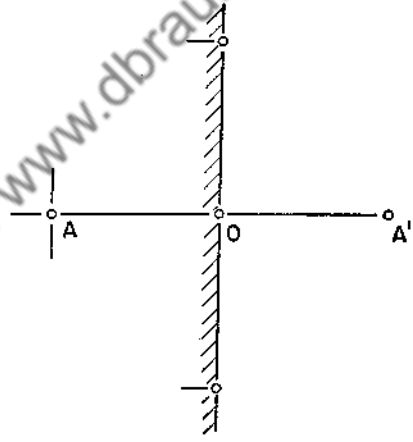


FIGURE 3.14

In place of prescribing the temperature itself over the boundary of the physical region, when $t > 0$, one might prescribe the rate of heat flow normal to all or part of the boundary. The preceding discussion indicates the modifications necessary in such cases. In particular, at a point on an *insulated* boundary, there must be no net flow *outward* in the direction normal to the boundary. This situation can be achieved by associating with each point A *adjacent* to such a boundary point 0 a symmetrically placed image point A' *at the same temperature* (Figure 3.14). Thus, for example, the procedure of Figures 3.10 to 3.12 would also be employed to solve the problem in which the right- and left-hand boundaries are *insulated* boundaries, rather than lines of symmetry; that is, the solutions of the two problems are identical. It should be noticed that

the requirement that the temperatures at A and 0 be equal, so that there shall be no flow in the rod joining these points, is *not* physically appropriate (unless the boundary points adjacent to 0 happen to be at the same temperature as 0) since the possibility of flow from A to 0 and thence along *boundary* rods (or conversely) should not be excluded. The difference between the solution corresponding to this requirement and that corresponding to the preferred one would, however, tend to zero as the spacings were continually reduced.

Unless the boundary of the relevant region is of a special form (such as a square, a rectangle with sides commensurable with a convenient spacing, or a figure bounded by a portion of such a boundary and one or more suitable lines making an angle of 45° with the remainder of the boundary), it is usually impossible to construct a square net in such a way that its boundary points all coincide with points of the actual boundary. Methods of determining appropriate values to be assigned to outer points of the net which do not fall on the boundary (and taking into account conditions involving the normal derivative of the unknown function in such cases) are considered in Section 3.18. In some cases, it is convenient to transform the problem into one involving boundaries which are rectilinear (or nearly so) by the use of conformal mapping. An example of this procedure is given in Section 3.19.

3.16. Laplace's equation in two dimensions. Steady-state flow of heat in a homogeneous, isotropic medium, in which the temperature depends only upon position specified by the rectangular coordinates x and y, is governed by *Laplace's equation:*

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0. \tag{232}$$

In this case, the values of the temperature T or of the normal derivative $\partial T/\partial n$ are usually prescribed at all points of the closed boundary B of a region R, and the temperatures at internal points of R are required. The former problem is known as the *Dirichlet problem*, the latter as the *Neumann problem*.

If equal spacings are taken in the x- and y-directions,

$$h_x = h_y = h, \tag{233}$$

the approximating difference equation takes the form

$$T(x, y) = \tfrac{1}{4}[T(x + h, y) + T(x - h, y) + T(x, y + h) + T(x, y - h)], \quad (234)$$

or, with the notation of Figure 3.15,

$$T_0 = \tfrac{1}{4}(T_1 + T_2 + T_3 + T_4). \quad (235)$$

Thus, when equal spacings are chosen, the difference equation requires that the temperature at any interior point be the average of the temperatures at the four adjacent points.

It is known that the solution of any transient heat-flow problem, in which the prescribed boundary temperatures do not vary with time, tends in time toward a solution of Laplace's equation which takes on the prescribed boundary values, and that only one such solution can exist. This fact indicates that if we start with *any* assumed temperature distribution at interior net points of the region R, and repeat the averaging process corresponding to (229) sufficiently often, the results of successive averaging processes will tend to the solution of the difference-equation formulation of the steady-state problem.

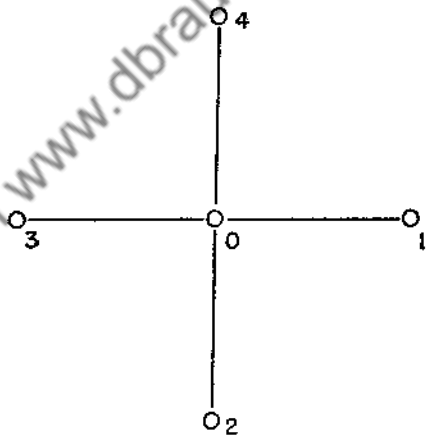


FIGURE 3.15

Frequently it is possible to *guess* the required steady-state distribution to a fair degree of accuracy, and to improve the approximation by a sequence of averaging processes until a repetition of values indicates that satisfactory accuracy has been obtained. Unfortunately, the convergence of this method is usually slow. A more flexible method, which often permits a very great decrease in the amount of necessary calculation, is outlined in Section 3.17.

However, in order to illustrate the method just described, we consider the determination of the steady-state solution of the prob-

lem dealt with in the preceding section. The result of such a series of calculations is presented in the following table:

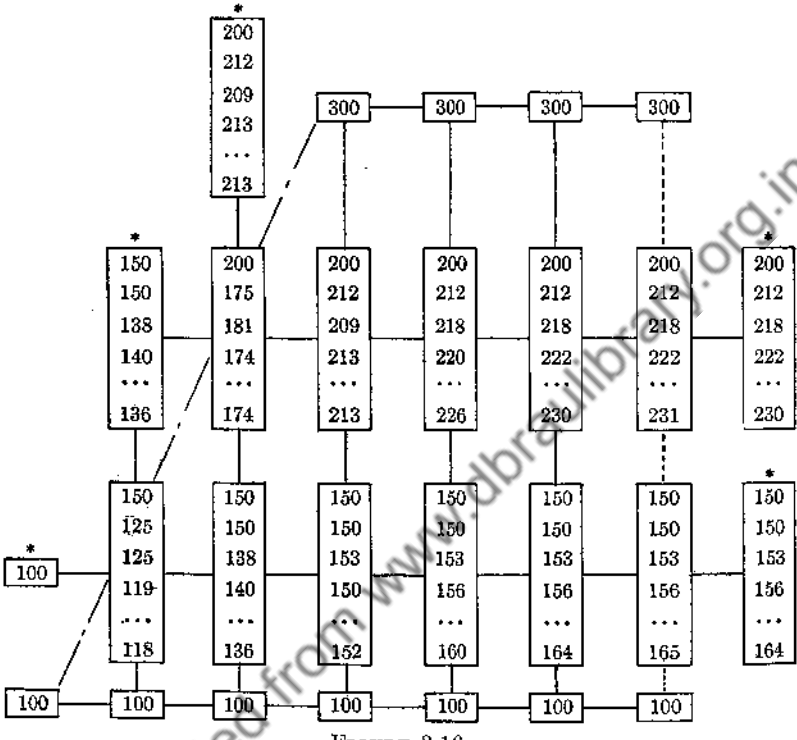


FIGURE 3.16

The entry at the top of each column denotes an initial "guess," while succeeding entries denote the results of successive averaging processes. The omitted entries correspond to the results of eight such cycles of operations. In each case, the entry in a given cycle is the average of the four neighboring entries in the *preceding* cycle. Here, a total of twelve cycles was required before all entries repeated themselves to the three significant figures retained.

If the procedure is modified in such a way that the entry in a given cycle is the average of the *most recently* calculated values of the four neighboring entries (so that entries in a cycle affect certain succeeding entries in the same cycle) the rate of convergence is increased. Here the rate of convergence depends upon the *order* in which the new entries are made. In particular, by proceeding first from left to right along the upper row of interior points, and

then from left to right along the lower row (entering values at exterior points immediately after symmetrically placed interior values have been calculated), we reduce the requisite number of cycles to six in this way.

It may be expected that further reduction of labor would result if it were possible, at each stage of the process, to determine those entries which differ most from the actual limiting values, and to concentrate primarily on improving those entries. A procedure which tends to accomplish this purpose, and which possesses certain other additional advantages, is described in the following section.

3.17. Relaxation methods and Laplace's equation. As applied to the difference equation (234) or (235), the so-called "relaxation" method associates with each interior node of the square net a "residual" R, defined by the equation

$$R_0 = T_1 + T_2 + T_3 + T_4 - 4T_0. \tag{236}$$

The difference equation (234) or (235) then requires that the residual at each *interior* point vanish. Suppose that an initially estimated value of T is associated with each net point, and that the corresponding residuals are also tabulated. If now at any interior net point 0 the estimate T_0 is modified by a certain amount, and the estimates at all other points are unchanged, it follows from (236) that the residual at that point is *decreased* by *four times* that amount, whereas the residuals at the neighboring point 1, 2, 3, and 4 are *increased* by that amount itself. Thus with the difference equation under consideration, we have the "relaxation pattern" indicated in Figure 3.17, which specifies changes in *residuals* corresponding to a unit increase in the *estimate* T_0.

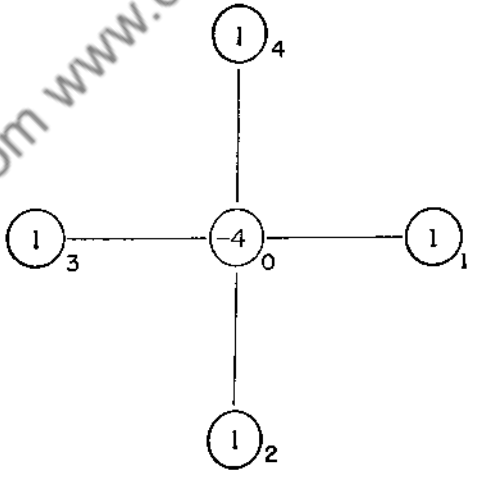


FIGURE 3.17

In general terms, the *relaxation method* then consists in con-

sidering, at each stage of the calculation, that point whose residual is of greatest numerical value, and modifying the estimate of the associated value of T at that point in such a way that the magnitude of that maximum residual is decreased.

In particular, one might remove ("liquidate") the residual at the point 0 *completely* by adding to T_0 exactly one-fourth of the residual R_0. However, this procedure is rarely useful except near the end of the over-all calculation, since new residuals will be introduced, in general, when the neighboring entries are subsequently modified. Since the algebraic reduction of a residual at a given point is accompanied by an algebraic increase in the neighboring residuals, it is usually advisable to "overrelax" a point 0 (that is, to add more than one-fourth R_0, so that the residual at 0 changes its sign) when the predominating residuals at the neighboring points are of the same sign as R_0, and to "underrelax" 0 otherwise. In this way, we tend to cause the residuals to differ in sign from point to point, as is usually desirable for the purpose of rapid convergence.

It may be noticed that, unless the relaxed point is adjacent to a boundary, the *algebraic sum* of the residuals is unchanged. However, since residuals are not calculated for boundary points, the net over-all residual *is* modified when a point adjacent to a boundary is relaxed. Thus it is apparent that, when the residuals are predominantly of one sign, the net over-all residual can be decreased only by effectively moving residuals to the boundary.

A useful physical interpretation of the relaxation process, as applied to the steady-state heat-flow problem, is obtained by comparing equations (236) and (230). Since the residual R_0 at any interior net point 0 is identical with the quantity Q_0/Kb, it follows that the residual at 0 is proportional to the rate of which heat *would be* conducted to 0 if the temperatures at the net points were in accordance with the estimated values. Thus the presence of a negative residual at an interior point would correspond to the presence of a "heat source" at that point. The relaxation process can accordingly be visualized, in this application, as essentially "balancing out" interior sources and sinks and moving *excess* sources (or sinks) to the *boundary* of the region.

In order to illustrate the process, we consider the situation in which T is prescribed as zero along one foot of an isosceles right triangle, and as 100° along the remainder of the boundary. The

temperatures at interior points (in the steady state) are required. For simplicity, we introduce only the three interior points A, B, and C of Figure 3.18. Initial estimates are made at each of these

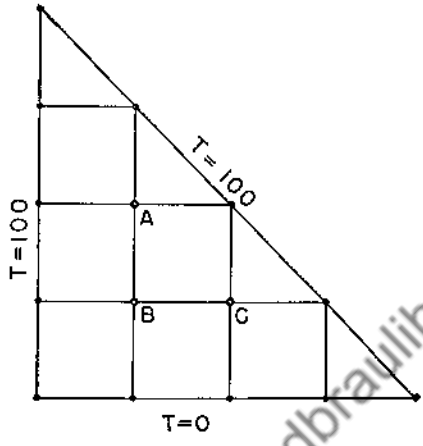


FIGURE 3.18

points, and the corresponding residuals (indicated in parentheses) are calculated, as is shown in Figure 3.19. A typical sequence of relaxations at the three interior points is indicated below.

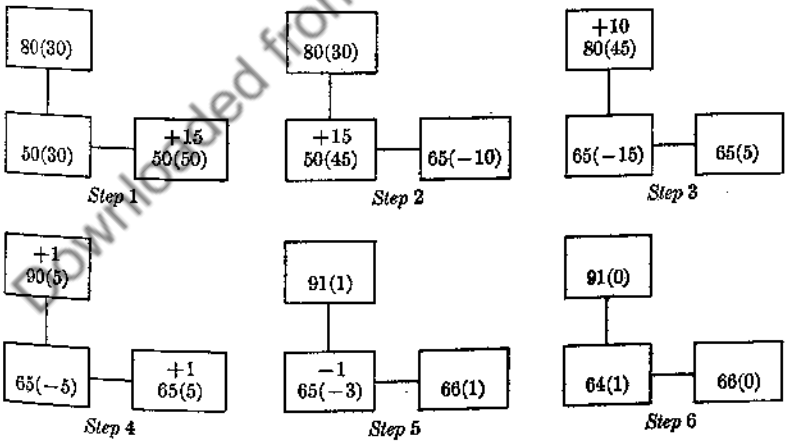


FIGURE 3.19

In each case, the amount added to a particular entry is tabulated immediately above that entry.

In Step 1, we notice that the point C has the largest residual (50). Since the neighboring residual is of the same sign, we *over*relax C by adding 15, to obtain the array of Step 2. At this stage, B has the largest residual (45). Since the larger of the neighboring residuals is of the same sign, we again overshoot zero. In the third step, we *under*relax A, to obtain the array of Step 4. Here we simultaneously relax A and C, and after one further step, arrive at the final results. The final residuals are then checked by equation (236), in order to expose possible errors in the intermediate calculations. In case the existence of such errors is discovered, the process is merely continued until the corrected residuals are removed (to within the tolerance adopted).

In actual practice, the successive steps are usually carried out on a single diagram, the successive entries (or corrections) and residuals at each point being arranged in a column. Since intermediate calculations are of no ultimate interest, it is often more convenient to enter fixed boundary values in ink, and interior values and residuals in pencil, and to alter these entries by erasure as the calculation proceeds. It is desirable, in a lengthy calculation, to check the residuals completely from time to time (taking into account the fixed boundary values), in order to avoid prolonged propagation of numerical errors.

The great advantage of the relaxation method over the averaging process discussed in the preceding section consists in its *flexibility*. While the successive steps could indeed be prescribed in a fixed manner, in such a way that the method becomes identical with the averaging process, even a limited amount of practice permits one to discover special devices and "rules of thumb" which tend to improve the rate of convergence of the iterations. The possibility of overshooting or undershooting zero residuals, and of concentrating at each stage on that point at which the difference equation is least nearly satisfied, is particularly valuable.

A special technique, which is very useful when the residuals at points in a certain region are predominantly of the same sign, is usually referred to as *block relaxation*. In this operation, *all* entries corresponding to a chosen connected set of interior net points are *simultaneously* modified by a certain amount. For example it is easily verified that if all entries in the block indicated in Figure 3.20 are *increased by unity*, the *residuals* at these points and at neighbor-

ing points are modified as shown. It is convenient to speak of a net line which extends from a point in a block to a point *outside* the block as a "free line" of that block. With this terminology, the following rule of *block relaxation* is readily established in the general case under consideration:

If each entry in a block is increased by unity, the residual at any point of the block is decreased by the number of free lines leading from

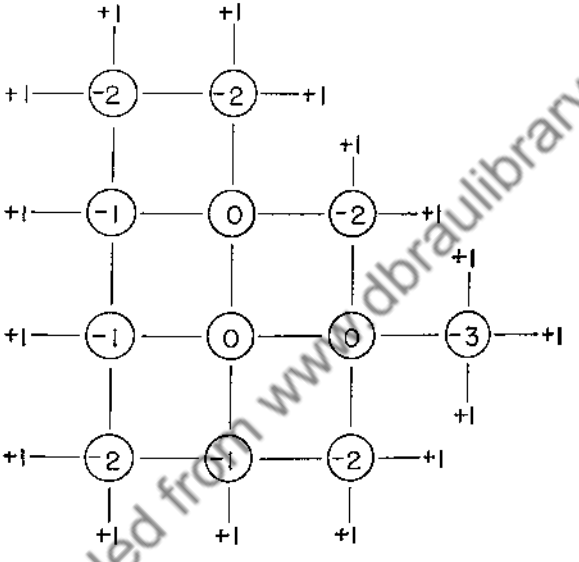


FIGURE 3.20

that point, and the residual at any point adjacent to the block is increased by the number of lines which join that point to points in the block.

In particular, if *all* interior points in the region under consideration are relaxed as a single block, so that the "free lines" lead only to boundary points (at which residuals are not calculated), the increase of each entry by unity is accompanied merely by the decrease of the residual at any point by the number of lines joining that point to boundary points. The net decrease in the over-all residual is thus equal to the total number N of lines joining boundary points to interior points.

It is often convenient to initiate the relaxation process by first calculating the net over-all residual ΣR_i, corresponding to the

original estimates, and then increasing the estimate at each interior point by approximately $(\Sigma R_i)/N$, so that the new *mean* residual is approximately zero.

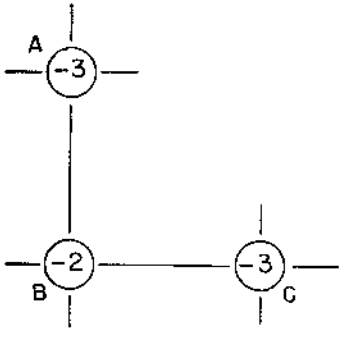


FIGURE 3.21

Thus, in the example of Figure 3.18, the initial total residual is $30 + 30 + 50 = 110$ and $N = 8$, so that $(\Sigma R_i)/N = 14$ to two significant figures. The block relaxation pattern, giving the changes in *residuals* corresponding to a *unit* overall increase in the entries, is given by Figure 3.21. Thus, if each interior entry in Figure 3.18 is increased by 14, the residuals at the points A, B, and C are decreased by 42, 28, and 42, respectively, and the new array of entries and residuals is as given in Figure 3.22. Two further relaxations then lead to the final result (when only two-figure accuracy is required).

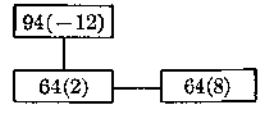


FIGURE 3.22

In applying relaxation methods, it is usually desirable to start the process with relatively few interior points, and to proceed successively to finer nets by adding new points (either throughout the region or only in areas where rapid transitions occur) after each series of relaxations.

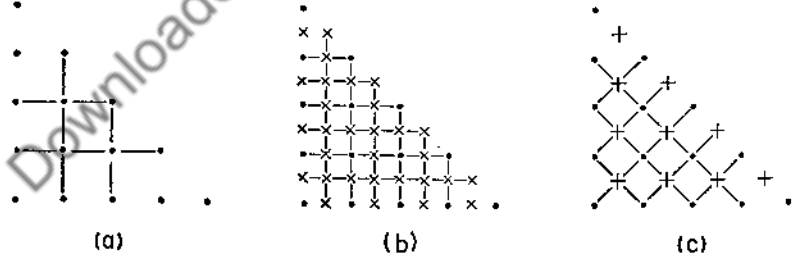


FIGURE 3.23

Thus, in the illustrative example of this section, the initial net (Figure 3.23a) could be refined by halving the spacing throughout (Figure 3.23b), and hence obtaining a similar net with 21 interior

points, or by merely introducing an additional point at the center of each initial square (Figure 3.23c).

The new net obtained by the latter method, which contains 9 interior points, is seen to be diagonal to the original net. However, since Laplace's equation is invariant under a rotation of the coordinate system, the basic relation (236) can be applied equally well to the new net. In any case, the calculated approximations at the initially chosen points may serve as starting estimates in the following process, and a starting estimate at each added point can be obtained by graphical interpolation or, in the case of the second refinement, by calculating the average of the four (previously approximated) adjacent values.

By judiciously combining the basic averaging process with the more flexible relaxation procedure, considerable labor can often be avoided. In particular, we may notice that the point B in Figure 3.18 is at the center of a square of side $2h$, all vertices of which lie on the boundary. If the value of T at the right-angle corner is taken to be 50 (the mean of the two limits approached along the edges), the temperature at B may be estimated as the average of the temperatures at *diagonally* adjacent points: $(100 + 100 + 50 + 0)/4 = 62.5$. Corresponding estimates at A and C are then immediately found, as the average of values at adjacent points, to be 91 and 66, respectively. With these starting values, a single relaxation leads to the final (two-figure) result for this net.

It may be remarked that the possibility of using triangular and hexagonal nets, as well as taking into account corrections to the approximation of an nth derivative by an nth difference, has been considered in the literature (see Reference 5).

In the following section, a brief account of the treatment of irregular boundaries is given.

3.18. Treatment of boundary conditions. In most boundary-value problems governed by Laplace's equation, the boundary condition specifies either the unknown function, say T, or the normal derivative $\partial T/\partial n$ at each point of the boundary. In more complicated cases, a linear combination of these two quantities may be prescribed, or T may be prescribed along part of the boundary and $\partial T/\partial n$ along the remainder. In this section we consider the treatment of such conditions with reference to the approximate formulation of the problems in terms of difference equations.

Boundary-value problems of the *first kind*, in which the unknown function is prescribed along the complete closed boundary of a region, ordinarily present no essential difficulties (except at corners, where difficulties already discussed may exist) when the boundary points of the net coincide with actual boundary points. Boundaries for which this last situation does not exist may be called *irregular boundaries*, and boundary points of the net which do not lie on the true boundary may be termed *irregular points* of the net. It is with the treatment of such points that we are here concerned. While more or less elaborate methods of dealing with these points have been proposed, experience indicates that the use of relatively simple methods is to be preferred, since the corresponding inaccuracies tend to disappear as the net is refined.

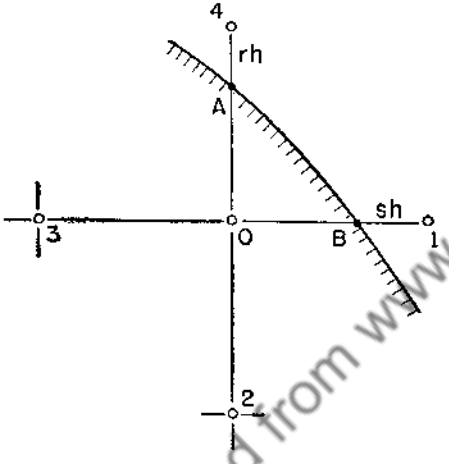


FIGURE 3.24

In a region such as that of Figure 3.24, which is adjacent to a boundary, the calculation of the residual at the interior point 0 may involve the values of T at one or more irregular points such as points 1 and 4. If we suppose that the spacing h is sufficiently small that the function T is nearly a linear function of x and y in that region, we may obtain approximate values of T_1 and T_4 by linear extrapolation (or interpolation, if these points lie *inside* the true boundary), in the form

$$T_4 = T_A + \frac{r}{1-r}(T_A - T_0), \tag{237}$$

$$T_1 = T_B + \frac{s}{1-s}(T_B - T_0). \tag{238}$$

Here the ratios r and s are to be taken as *negative* if the points 4 and 1, respectively, lie *inside* the true boundary. The values T_A and T_B are assumed to be prescribed. Two possible relaxation

procedures are then suggested, of which the first is usually to be preferred.

In the first place, we may estimate T initially at *all* points of the chosen net. In the absence of further information, the value of T at an irregular point may be taken as the prescribed value at the nearest boundary point. The boundary points of the net are then held fixed, and the inner points are relaxed until their residuals are liquidated. At this stage, corrected estimates at irregular points may be obtained by the use of extrapolation. The resultant residuals at interior points are then again liquidated, and the process is repeated until no further changes occur. Usually only a small amount of recalculation is needed, and it may be preferable to apply the correction only after transition to a finely spaced net has been made.

Alternatively, in the case of the configuration of Figure 3.24, we may use (237) and (238) to eliminate T_4 and T_1 from the expression for the residual at the point 0, to reduce the expression

$$R_0 = T_1 + T_2 + T_3 + T_4 - 4T_0 \tag{239}$$

to the form

$$R_0 = \frac{1}{1-r} T_A + \frac{1}{1-s} T_B + T_2 + T_3 - \left(4 + \frac{r}{1-r} + \frac{s}{1-s}\right) T_0. \tag{240}$$

The points 1 and 4 may then be omitted from later consideration. Equation (240) permits the initial estimate of the residual at 0; it must then be noticed that whereas unit increases in T_2 and T_3 each lead to unit increases in R_0, as before, a unit increase in the estimated value of T_0 now leads to a decrease in R_0 given by $\left(4 + \frac{r}{1-r} + \frac{s}{1-s}\right)$, rather than 4. That is, in addition to modifying the initial calculation of residuals at points adjacent to irregular points of the net, we must form an array of modified relaxation patterns for certain points. Irregular points are then not involved in the subsequent relaxation.

In dealing with a boundary-value problem of the *second kind*, where the derivative of T normal to the boundary is prescribed at all points of a closed boundary, we encounter certain additional

difficulties. We consider first the simple case in which a portion of the boundary considered is straight, and such that boundary points of the net fall on it (Figure 3.25). With the notation of Figure 3.25, the basic condition to be satisfied at a boundary point 0 can be most simply approximated by the difference condition

$$T_0 - T_3 = h\left(\frac{\partial T}{\partial n}\right)_0. \tag{241}$$

If T were a linear function of distance along the line joining points 3 and 0, this condition would be in complete agreement with the exact one. More generally, if T_b and T_i represent respectively the values of T at a boundary point and at an interior point which lies on the normal to the boundary at the boundary point, and at a distance d from that point, the corresponding condition could be taken in the form

$$T_b - T_i = d\left(\frac{\partial T}{\partial n}\right)_b. \tag{242}$$

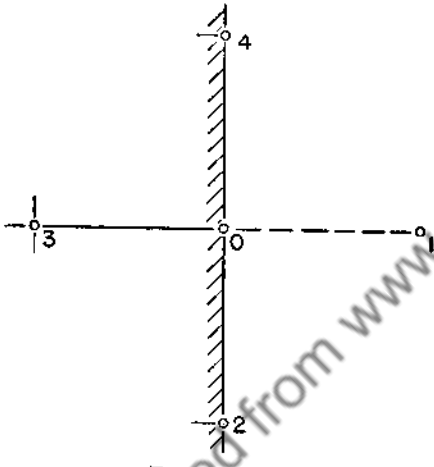


FIGURE 3.25

For irregular boundaries, two such *net* points are usually not available, and further approximations are needed, as will be discussed shortly.

For regular boundaries of special types, a condition which is both more nearly accurate and also more conveniently used than (241) can often be derived. We again restrict attention to the case in which the governing differential equation is Laplace's equation, so that the problem may be interpreted in terms of steady-state heat flow. We may notice first that since interior heat sources and sinks cannot be present, it follows that the *net* rate of flow of heat through the entire boundary must be zero, so that $\partial T/\partial n = -Q/KA$ must be so prescribed that its *mean value* along the entire boundary is zero. Also, it is clear from physical considerations that in this case the interior temperature distribution

is not uniquely determined, since any constant temperature may be added at all points without affecting the value of the normal derivative at the boundary. However, it is known that the solution *is* unique except for such an arbitrary additive constant. Thus, the temperature at any conveniently chosen interior point may be arbitrarily specified, after which the entire distribution is determinate.

For the boundary point 0 of Figure 3.25, the rate Q_0 at which heat flows *outward* from the boundary (through a "rod" of cross section $A = b\,h$) is given by the equation

$$\frac{Q_0}{K\,b} = (T_3 - T_0) - \frac{1}{2}(T_0 - T_4) - \frac{1}{2}(T_0 - T_2)$$

or

$$\frac{Q_0}{K\,b} = T_3 + \frac{1}{2}T_2 + \frac{1}{2}T_4 - 2T_0, \tag{243}$$

one-half the flows from 0 to 2 and to 4 being considered since only half-strips are associated with boundary rods. At boundary points where T is prescribed, equation (243) serves merely to determine the rate at which heat is taken away (or supplied, if Q_0 is negative) at points of the boundary, and does not enter into the actual determination of temperatures. However, at a boundary point where T is not prescribed, equation (243) may be taken as the *physically* motivated condition which requires that the rate of outward flow normal to the boundary at such a point take on a *prescribed* value Q_0. Also, since we have the requirement $Q_0 = -K\,b\,h(\partial T/\partial n)_0$, we may write (243) in the form

$$\frac{1}{2}(T_2 + 2T_3 + T_4 - 4T_0) = -h\left(\frac{\partial T}{\partial n}\right)_0. \tag{244}$$

If we now introduce a fictitious heat source or sink at the exterior point 1 of Figure 3.25, at a temperature T_1 such that

$$T_1 - T_3 = 2h\left(\frac{\partial T}{\partial n}\right)_0, \tag{245}$$

the requirement that (244) be satisfied is equivalent to the requirement that the conventional residual

$$R_0 = T_1 + T_2 + T_3 + T_4 - 4T_0 \tag{246}$$

shall vanish. Thus, we essentially extend the region of definition by introducing external points, and deal with points on the actual boundary as interior points of the extended region. Whereas the condition of equation (241) is exact only when T varies *linearly* near the boundary, the condition of (245) is exact also when T varies *parabolically*. In particular, for an *insulated* boundary, this procedure reduces to the introduction of *image* points which was discussed at the end of Section 3.15.

Clearly, the fictitious point need not be considered if we replace (246) by the definition

$$R_0 = T_2 + 2T_3 + T_4 - 4T_0 + 2h\left(\frac{\partial T}{\partial n}\right)_0, \tag{247}$$

and associate this residual with the *boundary* point 0. The prescribed quantity $h\,(\partial T/\partial n)_0$ then enters only into the *initial* calculation of R_0, and the subsequent relaxation pattern is modified only in that a unit increase in the estimated value of T_3 now leads to an increase of *two* units in R_0.

We may notice that the preceding equations define the residual at a boundary point where T is not prescribed as *twice* the net rate of heat flow (in units of Kb) into such a point. This definition merely weights the error committed in failing to satisfy (243) at a boundary point by a factor of two, and hence does not affect the array of approximate temperatures to which the relaxation process leads. It should be noticed that (247) is relevant only to Laplace's equation. However, (245) may be used to obtain similar definitions in other cases.

The success of the preceding procedure is seen to be a consequence of the fact that the line joining the points 3 and 0 is normal to the boundary. In the case of an irregular boundary, the generalization of this formulation usually leads to a rather elaborate and confusing relaxation pattern, and it is usually preferable to generalize the simpler condition (241) or (242), making up for the resultant loss in accuracy by proceeding to a finer net spacing. We next discuss one such generalization.

Figure 3.26 represents a situation in which one outer point of the net falls outside a curved boundary. A line is drawn through this point, *normal to the boundary*, and is extended until it intersects a mesh line, which lies completely inside the actual boundary, at the

point P. If this normal intersects the boundary at the point B, and if a linear variation of T along the normal is assumed, the temperatures at the points 1 and P are related (approximately) by the equation

$$T_1 - T_P = d\left(\frac{\partial T}{\partial n}\right)_B, \tag{248a}$$

where d is the distance between the points 1 and P. If also we assume a linear variation of T along the line joining the points

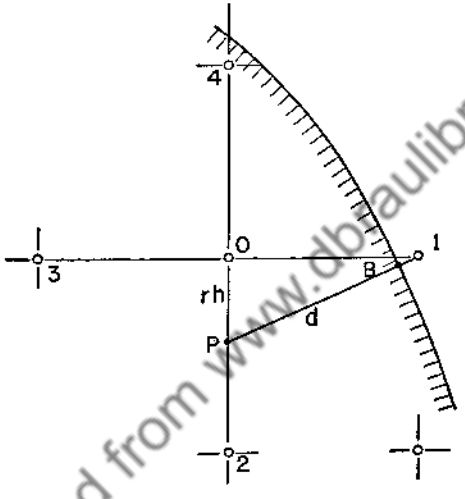


FIGURE 3.26

0 and 2, there follows also (by linear interpolation)

$$T_P = T_0 + r(T_2 - T_0), \tag{248b}$$

where r is the ratio of the distance from 0 to P to the spacing h. The quantity T_P can then be eliminated between (248a) and (248b), to give the relation

$$T_1 = (1 - r)T_0 + r\,T_2 + d\left(\frac{\partial T}{\partial n}\right)_B. \tag{249}$$

This relation is easily remembered if it is noticed that it is equivalent to (242), where $T_i = T_P$ is the weighted average of the values of T at the extremities of the mesh line intersected by the normal.

By introducing (249) into (246), we obtain an expression for the residual at the point 0, in the form

$$R_0 = (1 + r)T_2 + T_3 + T_4 - (3 + r)T_0 + d\left(\frac{\partial T}{\partial n}\right)_B. \quad (250)$$

The use of this expression permits omission of the irregular point 1 from further consideration. However, the initial calculation of R_0 and the subsequent relaxation patterns relevant to neighboring points are both modified.

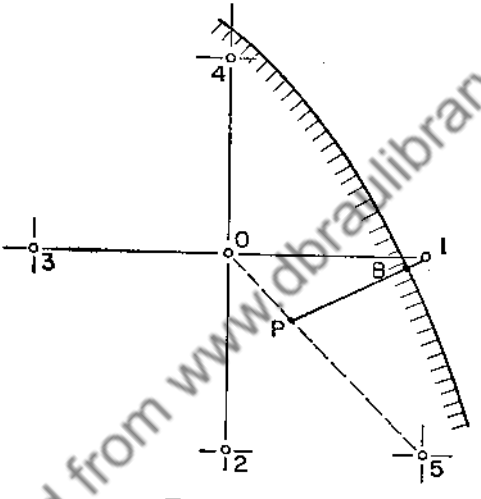


FIGURE 3.27

It is seen that the normal at the boundary point B may instead intersect the horizontal line extending to the right from point 2 in Figure 3.26, in which case (250) will be modified in such a way that the value of T at an additional point is involved. Further modification will occur if more than one of the net points adjacent to 0 are irregular points of the net.

While (248a) is exact if T is a linear function of the distance along the normal at B, it will be exact also for parabolic variation along that line if the point B happens to bisect the line connecting the points 1 and P, and will afford a good approximation to the true requirement if B *nearly* bisects that segment. It is clear that the normal could instead be terminated on a *diagonal* line, as is indicated in Figure 3.27. The corresponding modification of the

definition of R_0 is easily determined. While this procedure involves interpolation over somewhat shorter intervals, it tends to introduce a greater number of points into the relaxation patterns, and to further complicate the calculation.

In order to illustrate the use of the method outlined, we suppose that a portion of the boundary under consideration is a quadrant of a circle of radius $2h$ (Figure 3.28), and that along that arc it is prescribed that $h(\partial T/\partial n) = 50 \sin\theta$, where θ is the polar angle indicated in that figure. By constructing the indicated normals at the outer points 1 and 4, and proceeding as outlined above, approximate expressions relating T_1 and T_4 to interior values of T are easily obtained in the form

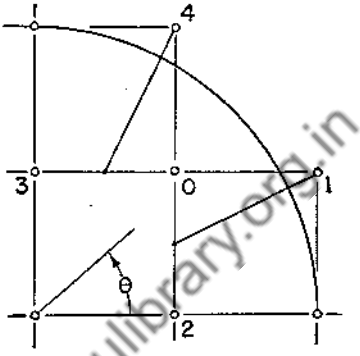


FIGURE 3.28

$$T_1 = \frac{1}{2}T_0 + \frac{1}{2}T_2 + \frac{\sqrt{5}}{2}\cdot 50 \cdot \frac{1}{\sqrt{5}} = \frac{1}{2}T_0 + \frac{1}{2}T_2 + 25,$$

$$T_4 = \frac{1}{2}T_0 + \frac{1}{2}T_3 + \frac{\sqrt{5}}{2}\cdot 50 \cdot \frac{2}{\sqrt{5}} = \frac{1}{2}T_0 + \frac{1}{2}T_3 + 50,$$

and the residual at the point 0 may accordingly be taken in the following form:

$$R_0 = (\tfrac{1}{2}T_0 + \tfrac{1}{2}T_2 + 25) + T_2 + T_3 + (\tfrac{1}{2}T_0 + \tfrac{1}{2}T_3 + 50) - 4T_0$$
$$= \tfrac{3}{2}T_2 + \tfrac{3}{2}T_3 - 3T_0 + 75.$$

The use of this definition implies the modification of the relaxation patterns corresponding to the points 0, 2, and 3. That is, we obtain from it the entries listed in Figure 3.29 for those patterns. The remaining entries in those patterns, and in other modified patterns, are to be obtained after other modified residuals have been defined.

3.19. Other applications of relaxation methods. In the difference-equation formulation of the Dirichlet problem, N uniformly spaced interior points of the relevant region are selected,

and the N unknown quantities in the resultant formulation are the values of the function T at those N points. Associated with each such point, the basic difference equation then affords a linear equation involving certain of those unknowns, so that the resultant problem then actually consists in solving a set of N linear algebraic equations in N unknowns. For the Dirichlet problem, not more than five unknown quantities are involved in any one equation; the presence of an adjacent boundary point reduces this number by

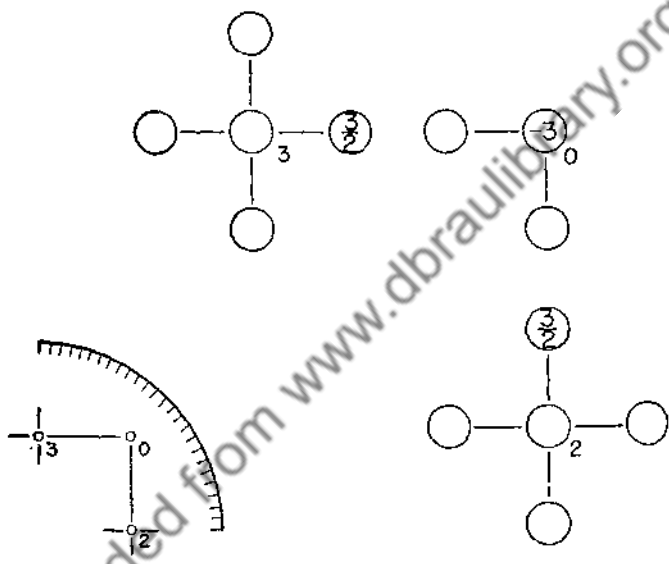


FIGURE 3.29

unity, and introduces a *known* quantity into the corresponding equation.

The relaxation method can then be considered as an iterative method of solving such a set of equations. Corresponding to an initial estimate of the N unknowns, a measure of the extent to which each equation fails to be satisfied (a "residual") is selected, and is associated with the point which gave rise to that equation. Next, a table listing changes in residuals due to a *unit* increase in the value of each unknown (a "relaxation pattern" corresponding to each point) is constructed. If the residual corresponding to the kth equation is the predominant one, the estimated value of the

kth unknown is modified (by use of this table) in such a way that this maximum residual is reduced, and the resultant modified array of residuals is determined. The process is repeated successively until the magnitudes of all residuals are within the tolerance adopted.

In the Dirichlet problem, the coefficient of the kth unknown in the kth equation is large relative to the coefficients of the remaining unknowns in that equation. The efficiency of the relaxation method, in dealing with this problem, is to a large extent a consequence of this situation.

The preceding summary essentially characterizes the application of relaxation methods to any problem which is specified (exactly or approximately) by a set of a finite number of linear algebraic equations in the same number of unknowns. While the precise procedure whereby the magnitudes of the several residuals are eventually liquidated is not specified, it is exactly in this flexibility that the power of the method lies. By arbitrarily prescribing the *order* and *nature* of successive steps, the relaxation process can be made to be equivalent to any one of several standard iterative methods for solving such sets of equations. However, in consequence of the simplicity of the basic ideas involved, only a moderate amount of experience and ingenuity leads to ability to vary the technique in an efficient way, in accordance with the peculiarities of the particular problem under consideration. While certain "ill-conditioned" sets of equations are apparently not amenable to *any* standard iterative methods of solution, the flexibility of the relaxation method often permits its application when standard methods fail.

In dealing with physical problems, the successive steps in the iterative process may be motivated by physical considerations, and one may take full advantage of information afforded by known solutions of similar problems, or of trends or peculiarities indicated by early stages of the calculation.

The treatment of problems governed by Laplace's equation is particularly straightforward because of the simplicity and uniformity of the relevant relaxation pattern. In the more general case, this pattern may vary from point to point in the region considered.

The application of relaxation methods to boundary-value problems governed by *ordinary* differential equations [see, for example,

equations (207) to (211) of Section 3.13] is clearly a one-dimensional specialization of the technique associated with problems governed by partial differential equations, and will not be considered explicitly here. However, a few further examples may be cited to indicate the scope of the applicability of these methods in the treatment of other types of problems.

In the approximate solution of *Poisson's equation* in rectangular coordinates,

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + f(x, y) = 0, \tag{251}$$

where the unknown function $\phi(x, y)$ is prescribed along the boundary of a region and $f(x, y)$ is a given function, the residual at an interior point 0 may be defined by the equation

$$R_0 = \phi_1 + \phi_2 + \phi_3 + \phi_4 - 4\phi_0 + h^2 f_0, \tag{252}$$

with the notation of Figure 3.17. By comparing (236) and (252), we see that the solution of such a problem differs from the solution of the Dirichlet problem only in that the known quantity $h^2 f_0$ is *initially* added to the calculated residual at each point. The relaxation *pattern* is unchanged, and the subsequent relaxation process is then carried out exactly as before.

In dealing with a *characteristic-value* problem, such as that of determining values of λ for which the equation

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \lambda\phi = 0 \tag{253}$$

possesses nontrivial solutions which vanish (or satisfy certain other homogeneous conditions) along the boundary of a given region, we are led in a similar way to define the residual

$$R_0 = \phi_1 + \phi_2 + \phi_3 + \phi_4 - (4 - \lambda h^2)\phi_0 \tag{254}$$

at each interior point. Various methods of making use of the relaxation process have been proposed in connection with problems of this type. Such methods consist essentially in first estimating the values of the *fundamental* characteristic function ϕ (corresponding to the smallest characteristic value of λ) at the net points, and in then determining a corresponding estimate of the relevant value of λ. This latter estimate may, for example, be taken as

the average value of the ratio $(4\phi_0 - \phi_1 - \phi_2 - \phi_3 - \phi_4)/h^2\phi_0$ at interior net points of the region. With this estimated value of λ, it will (in general) be impossible to liquidate completely all residuals. However, after a certain amount of relaxation a new estimate of λ may be calculated, and the process may be repeated until a value of λ is obtained for which all residuals are liquidated within the tolerance adopted. Since the characteristic functions are determinate only within an arbitrary multiplicative constant, and since the conditions of the problem are satisfied by the trivial solution for which $\phi = 0$ everywhere, convergence to this trivial solution can be averted by arbitrarily fixing the value of ϕ at a conveniently chosen interior point of the net. The approximate calculation of additional characteristic functions and values of λ (as well as the formulation of improved techniques) can be based on the orthogonality properties considered in Chapter 1 (see Sections 1.11 and 1.24).

The application of relaxation methods to the solution of boundary-value problems governed by linear differential equations of higher order involves more elaborate relaxation patterns. Thus, for example, corresponding to the *biharmonic equation*

$$\frac{\partial^4\phi}{\partial x^4} + 2\frac{\partial^4\phi}{\partial x^2\,\partial y^2} + \frac{\partial^4\phi}{\partial y^4} = 0 \tag{255}$$

the relaxation pattern associated with an interior point of the specified region is found to involve the residuals at *twelve* equally spaced neighboring points. With the notation of Figure 3.30, the residual at an inner point 0, *which is not adjacent to the boundary*, is readily found to be expressible in the form

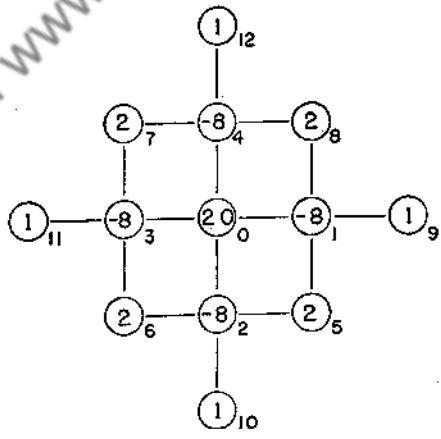


FIGURE 3.30

$$R_0 = 20\phi_0 - 8(\phi_1 + \phi_2 + \phi_3 + \phi_4) + 2(\phi_5 + \phi_6 + \phi_7 + \phi_8) + (\phi_9 + \phi_{10} + \phi_{11} + \phi_{12}). \tag{256}$$

Thus the changes in the residuals accompanying a *unit* increase in the estimated value of ϕ_0 at such a point are as indicated in Figure 3.30. At points adjacent to the boundary, the difference equation does not apply, and hence fails to define residuals. However, along the boundary of the relevant region, *two* conditions must be prescribed in the exact formulation of the problem. For example, the values of both ϕ and its normal derivative $\partial\phi/\partial n$ may be prescribed along the boundary. In the corresponding difference-equation formulation, one must prescribe boundary values of ϕ and also (essentially) relate boundary values to values at points which are adjacent to the boundary. In particular, along a regular straight boundary, the requirement that $\partial\phi/\partial n$ take on prescribed values can be conveniently taken into account by introducing fictitious external points which are images (relative to the boundary) of interior points adjacent to the boundary (Figure 3.31). If the value

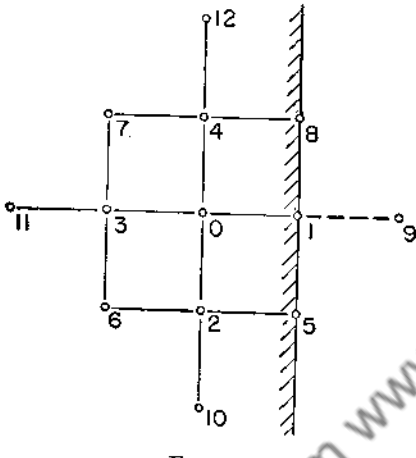


FIGURE 3.31

$$\phi_9 = \phi_0 + 2h\left(\frac{\partial\phi}{\partial n}\right)_1 \tag{257a}$$

is assigned to the external point in Figure 3.31, the point 0 (which is adjacent to the boundary) can then be treated as a completely interior point in the extended region.

A less nearly exact formulation, which is, however, more easily generalized to the treatment of irregular boundaries, clearly consists in calculating ϕ_0 directly from the approximate relation

$$\phi_1 - \phi_0 = h\left(\frac{\partial\phi}{\partial n}\right)_1. \tag{257b}$$

This relation is exact for linear variation in ϕ near the boundary, whereas (257a) is exact also for parabolic variation.

Partial differential equations of the general form

$$\frac{\partial}{\partial x}\left(\alpha \frac{\partial \phi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\beta \frac{\partial \phi}{\partial y}\right) = F \tag{258}$$

arise frequently in practice. The functions α, β, and F may depend only on the position coordinates x and y, or they may, in addition, involve the unknown function ϕ and its *first* partial derivatives ϕ_x and ϕ_y. Several methods of applying relaxation techniques to the approximate solution of problems governed by such equations have been proposed. A frequently useful one consists in first writing (258) in the expanded form

$$\alpha\, \phi_{xx} + \beta\, \phi_{yy} = F - \alpha_x \phi_x - \beta_y \phi_y,$$

and in then introducing the functions

$$f = \frac{\beta}{\alpha}, \qquad G = \frac{1}{\alpha}\,(F - \alpha_x \phi_x - \beta_y \phi_y), \tag{259a,b}$$

so that (258) takes the form

$$\phi_{xx} + f\, \phi_{yy} = G. \tag{260}$$

In those cases when the problem governed by this equation is a *boundary-value* problem over a certain region, the coefficient $f = \beta/\alpha$ is generally *positive* over that region (that is, the differential equation is of the "elliptic" type).

The result of replacing ϕ_{xx} and ϕ_{yy} by appropriate difference quotients in (260) is then of the form

$$\frac{\phi_1 - 2\phi_0 + \phi_3}{h_x^2} + f_0 \frac{\phi_2 - 2\phi_0 + \phi_4}{h_y^2} = G_0, \tag{261}$$

where f_0 and G_0 represent the values of f and G at the point 0. In the case when f and G depend only on x and y, these quantities are known; otherwise, their values at 0 depend upon the function ϕ which is to be determined. If equal spacings are taken, so that

$$h_x = h_y = h,$$

the residual at any interior net point 0 can be defined by the equation

$$R_0 = \phi_1 + f_0\phi_2 + \phi_3 + f_0\phi_4 - 2(1 + f_0)\phi_0 - h^2 G_0. \tag{262}$$

Corresponding to estimated values of ϕ at net points, the coefficients f_0 and G_0 are then calculated at each point. If values of partial derivatives of ϕ are involved in the calculation of these coefficients, these values are approximated by appropriate difference ratios. Initial residuals are calculated at each net point, by the use of (262), and are partially liquidated by use of the corresponding relaxation pattern of Figure 3.32. This pattern will, in general, vary from point to point in the net. After a certain amount of relaxation according to this pattern, revised values of f_0 and G_0 are

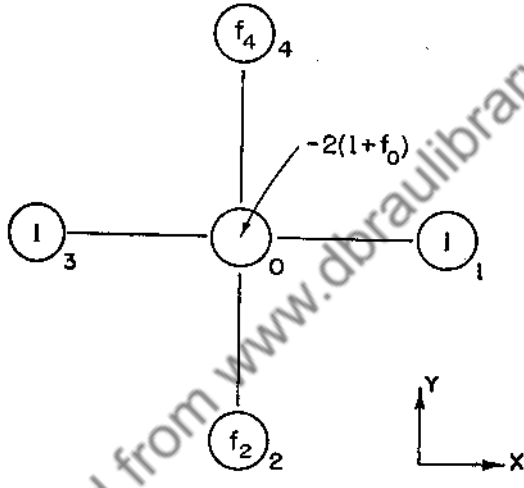


FIGURE 3.32

calculated at each point, modified residuals are calculated, and further relaxation is carried out according to the new pattern. The process is continued until no further revisions of f_0 and G_0 are required, and all residuals are satisfactorily liquidated.

When G involves *linear* terms in ϕ and/or its derivatives, it is often preferable to transpose such terms to the left in (260) and to replace partial derivatives by differences, thus obtaining a relaxation pattern which differs from that of Figure 3.32.

Many such applications of relaxation methods to the approximate solution of involved problems governed by one or more nonlinear partial differential equations can be found in the literature (see, for example, Reference 6).

When a two-dimensional problem is governed by Laplace's

equation, and the boundaries are of certain simple types, the use of conformal mapping is frequently advantageous. In particular, when part or all of the boundary consists of arcs of concentric circles (or of curves which are *nearly* circular), it is convenient to take the origin in the xy-plane at the center of the circles, and to associate with the xy-plane a new uv-plane according to the relations*

$$u = \log \frac{r}{a}, \qquad v = \theta, \tag{263}$$

where r and θ are the polar coordinates of the point (x, y) and a is a conveniently chosen positive constant. If we adopt the convention that $0 \leqq \theta < 2\pi$, it is then found, for example, that the region in the xy-plane bounded by two concentric circles of radii a and b, with center at the origin, is mapped into the rectangular region for which

$$0 \leqq v < 2\pi, \qquad 0 \leqq u \leqq \log \frac{b}{a}$$

in the uv-plane. The boundaries $r = a$ and $r = b$ map into the boundaries $u = 0$ and $u = \log (b/a)$, respectively; arcs of circles with center at the origin map into segments of lines parallel to the v-axis, and segments of radial lines into segments of lines parallel to the u-axis (see Figures 3.33a,b).

In terms of the new variables u and v defined by (263), Laplace's equation takes the form

$$\frac{\partial^2 \phi}{\partial u^2} + \frac{\partial^2 \phi}{\partial v^2} = 0. \tag{264}$$

If the values of ϕ are prescribed along a boundary in the xy-plane, the function ϕ must then take on these same values at corresponding points of the corresponding boundary in the uv-plane. Thus by introducing a square network in the new region, the usual relaxation methods can be conveniently employed to determine values of ϕ at interior points of that region, and hence at corresponding points in the original region. The relations

$$\frac{\partial \phi}{\partial r} = \frac{1}{a e^u} \frac{\partial \phi}{\partial u}, \qquad \frac{1}{r} \frac{\partial \phi}{\partial \theta} = \frac{1}{a e^u} \frac{\partial \phi}{\partial v} \tag{265}$$

* In the terminology of functions of a complex variable, if we write $z = x + i y$ and $w = u + i v$, the z-plane (with a "cut" along the positive x-axis) is mapped into a portion of the w-plane by the relation $w = \log (z/a)$.

permit the translation of conditions prescribing the normal derivative of ϕ along a circular or radial boundary in the original plane into corresponding conditions in the uv-plane.

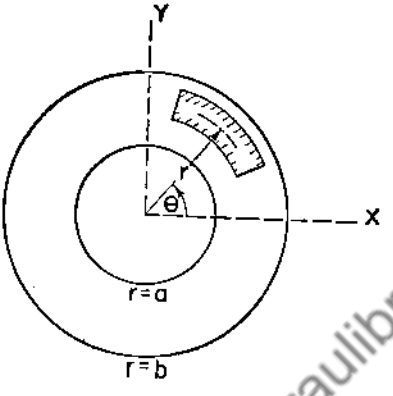


FIGURE 3.33a

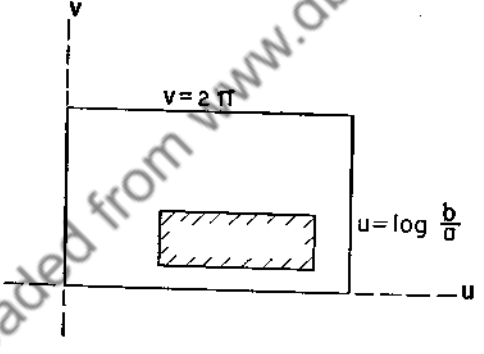


FIGURE 3.33b

Under the same mapping, it is found that *Poisson's equation*, in the form

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + f(x, y) = 0, \tag{266}$$

is transformed into the equation

$$\frac{\partial^2\phi}{\partial u^2} + \frac{\partial^2\phi}{\partial v^2} + a^2e^{2u}F(u, v) = 0, \tag{267}$$

where

$$F(u, v) = f(a\,e^u \cos v,\ a\,e^u \sin v). \tag{268}$$

3.20. Convergence of finite-difference approximations. In the remainder of this chapter we consider briefly certain questions which bear on the validity of considering the solution of a problem governed by a *difference* equation as an approximation to the solution of a related problem governed by a *differential* equation.

We suppose, first of all, that the exact problem, governed by a differential equation together with appropriate side conditions, *does indeed possess a unique solution.* In order to obtain an *approximate* solution of this problem, we replace the differential equation by a difference equation, in which derivatives with respect to the independent variables are replaced by difference quotients of the same order, relative to certain increments ("spacings") associated with the respective variables. The side conditions are expressed similarly, in terms of finite differences.

In particular, for a problem involving two independent variables the resultant difference equation is then valid at N interior vertices of a network of squares or rectangles which (exactly or approximately) covers a certain region of a plane in which the two variables are considered as rectangular coordinates. In this way, we obtain a set of N algebraic equations involving the values of the dependent variable at the N interior net points, certain of these equations generally involving prescribed quantities in consequence of the side conditions.

The two following important questions then arise:

1. Does this set of algebraic equations possess a (unique) solution?

2. Suppose that the chosen relationship between the spacings is retained, but that the spacings are indefinitely diminished in such a way that net points tend to become densely distributed everywhere inside the relevant region. Does the sequence of "approximate" solutions tend toward the solution of the exact problem?

It is clear that these questions are of more than theoretical interest. In particular, with reference to the second question, it is by no means inconceivable that the sequence of solutions corresponding to the successively refined nets might exist, and actually tend to a limiting function, but that this limit is *not* the solution to the *true* problem.

Unfortunately, complete answers to these questions are not

known, particularly in the case of nonlinear equations, and the engineer must rely in such cases upon his knowledge of the physical problem, to decide whether a function arrived at by successive refinement of a chosen net is to be accepted as an approximation to the required solution. However, in the case of *linear* problems, the theory is in a somewhat more satisfactory state, and certain known results* are summarized in the remainder of this section.

In many problems arising in practice, involving a linear partial differential equation of the second order, the relevant equation is a specialization of the form

$$\frac{\partial}{\partial x}\left(a\frac{\partial \phi}{\partial x}\right)+\frac{\partial}{\partial x}\left(b\frac{\partial \phi}{\partial y}\right)+\frac{\partial}{\partial y}\left(b\frac{\partial \phi}{\partial x}\right)+\frac{\partial}{\partial y}\left(c\frac{\partial \phi}{\partial y}\right)+d\phi=f, \tag{269}$$

where the prescribed coefficients may be constants or functions of the independent variables x and y. An equation of this form is said to be *self-adjoint*. Further, the equation is said to be of *elliptic*, *parabolic*, or *hyperbolic* type, in a given region, according as the discriminant $b^2 - ac$ is *negative*, *zero*, or *positive*, respectively, throughout that region. *Boundary-value* problems are essentially associated with *elliptic* equations ($b^2 < ac$), whereas *initial-value* problems are, in general, governed by *hyperbolic* equations ($b^2 > ac$) or *parabolic* equations ($b^2 = ac$). In the latter cases, one of the independent variables may often represent time while the other is a position coordinate.

For *boundary-value* problems governed by linear, self-adjoint equations of elliptic type, in which the function ϕ is prescribed along the boundary, it is known that the answers to the preceding questions are affirmative. That is, when the exact problem possesses a solution, the approximate problem can also be solved; further, for any arbitrarily fixed ratio of the spacings h_x and h_y, as the net is continually refined the solution of the approximate problem tends to the solution of the exact problem.

In the special case of the Dirichlet problem, in which the governing equation is that of Laplace, the answer to the first question can be obtained in a simple and instructive way, as follows, when equal spacings are taken in the x- and y-directions. As has been seen,

* See Reference 7.

the linear algebraic equations to be solved then require that the value of ϕ at any interior point of the net be the average of the values at the four adjacent points. From this fact it follows that *at no interior point of the net can ϕ take on a maximum or minimum value relative to neighboring points.* Hence, in this case, maxima and minima can occur only at boundary points. Suppose now that ϕ is prescribed as *zero* at points of a closed boundary. Then it is clear that ϕ must necessarily be zero at *all* points of the net, in consequence of the preceding result. But we have seen (Section 1.4) that if a homogeneous set of N linear algebraic equations in N unknowns possesses only the trivial solution, then the corresponding nonhomogeneous set of equations obtained by introducing nonzero right-hand members always possesses one and only one solution. Hence it follows that since the approximate Dirichlet problem has a unique (trivial) solution when zero boundary values are prescribed, it also has a unique solution in the general case. The treatment of the *convergence* question is more difficult (see Reference 7).

In the case of *initial-value* problems, the existence of a solution to the associated finite-difference problem is in general assured when the exact problem possesses a solution. However, in order that this solution tend toward the exact solution as the net is indefinitely refined, it is, in general, necessary that certain conditions involving the spacings be satisfied.

In particular, for a problem governed by the *hyperbolic* equation

$$V^2 \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial t^2} + a \frac{\partial \phi}{\partial x} + b \frac{\partial \phi}{\partial t} + c\,\phi = 0 \qquad (V > 0), \quad (270)$$

in which ϕ and $\partial\phi/\partial t$ are prescribed for $-\infty < x < \infty$ when $t = 0$, convergence of the finite-difference solution to the exact solution with successive net refinements has been established when the spacings h_x and h_t satisfy the relation

$$\frac{h_x}{h_t} > V. \qquad (271)$$

Further, it has been shown that convergence to the exact solution is, in general, *impossible* when $h_x < V\,h_t$. Certain facts bearing on this situation, in a special case, are discussed in the following section.

In more involved problems, the governing equation may, for example, be elliptic over parts of the region and hyperbolic over

other parts, and the nature of the equation in the neighborhood of a point may actually depend upon the behavior of the unknown solution in the neighborhood of that point. Problems of this type combine the features of boundary-value and initial-value problems, in which it may happen that part or all of a certain boundary is not specified, but must itself be determined. Very limited theoretical information is available as to the existence of solutions of either the exact or approximate problems, and as to the relationship between these solutions when they do exist. However, finite-difference methods often lead to "approximate solutions" to otherwise intractable problems, the validity of which can be checked empirically.

Relaxation methods are useful only in solving problems which are essentially of the boundary-value type. It should be noticed, however, that they are *needed* only in such problems since the difference equations associated with initial-value problems are, in general, readily solved by step-by-step methods. When an iterative method is used to solve the set of algebraic equations generated by the difference equation (by successive approximations) we encounter a third question as to whether this method itself converges to the solution of that set of equations. However, this question of convergence cannot be discussed with relation to the general *relaxation* procedure, since the explicit technique is not prescribed. When the set of equations possesses a solution, that solution *can* be obtained by *some* process of relaxation, and it remains only to discover some such process in a given case, by judicious trial and error.

3.21. The one-dimensional wave equation. In order to illustrate important aspects of the approximate solution of problems governed by hyperbolic equations, we suppose first that a function $\phi(x, t)$ is to satisfy the equation

$$V^2 \frac{\partial^2 \phi}{\partial x^2} - \frac{\partial^2 \phi}{\partial t^2} = 0 \tag{272}$$

for all positive t, $t > 0$, and for all values of x, $-\infty < x < +\infty$, where V is a positive constant, and where both ϕ and $\partial\phi/\partial t$ are prescribed when $t = 0$, according to the conditions

$$\phi(x, 0) = F(x), \qquad \phi_t(x, 0) = 0. \tag{273a,b}$$

The exact solution of this problem is easily seen to be of the form

$$\phi(x, t) = \tfrac{1}{2}[F(x - Vt) + F(x + Vt)]. \tag{274}$$

Before considering the corresponding finite-difference formulation of the problem, we may notice that this result is capable of an interesting interpretation. If we consider the point $P(x_1, t_1)$ in the xt-plane (Figure 3.34), equation (274) states that the value of ϕ at that point is the mean of the *prescribed* values of ϕ at the two points A and B on the x-axis for which $x = x_1 - Vt_1$ and $x = x_1 + Vt_1$, respectively. More generally, it is seen that the values of ϕ at all points in the shaded triangular section of Figure 3.34 depend upon the *prescribed* values of ϕ at points in the interval AB of the x-axis, and *only* upon those particular initial values. In the more general case, when arbitrary nonzero initial values of $\partial\phi/\partial t$ are also prescribed along the line $t = 0$, it is found that again only those values prescribed at points in the interval AB influence the solution in the region PAB (see Problem 88).

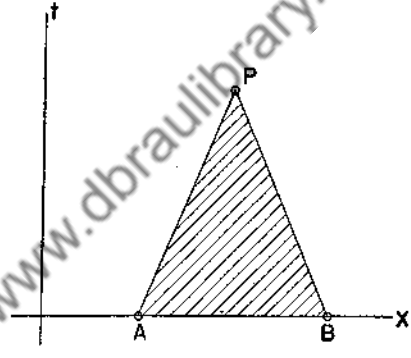


FIGURE 3.34

We may notice further that one of the terms in (274) remains constant along any line $x - Vt =$ constant, while the other term is constant along any line $x + Vt =$ constant. These two families of lines are known as the *characteristics* of the differential equation (272). Thus, the region PAB of Figure 3.34 is the region bounded by the initial line $t = 0$ and by the two characteristics which pass through the point P. We may speak of it as the *region of determination* for the point P, relevant to the differential equation (272).

Suppose now that (272) is replaced by the difference equation

$$V^2 \frac{\phi(x + h_x, t) - 2\phi(x, t) + \phi(x - h_x, t)}{h_x^2} - \frac{\phi(x, t + h_t) - 2\phi(x, t) + \phi(x, t - h_t)}{h_t^2} = 0, \tag{275}$$

where h_x and h_t are spacings relevant to x and t. If we relate these spacings by the equation

$$h_x = \kappa V h_t, \tag{276}$$

where κ is a constant, and associate with the point (x, t) and the four adjacent points the indices indicated in Figure 3.35, equation (275) can be rewritten in the abbreviated form

$$\kappa^2\phi_4 = \phi_1 + \phi_3 - \kappa^2\phi_2 + 2(\kappa^2 - 1)\phi_0. \tag{277}$$

With respect to the corresponding rectangular network in the xt-plane, equation (273a) prescribes ϕ at points on the initial line

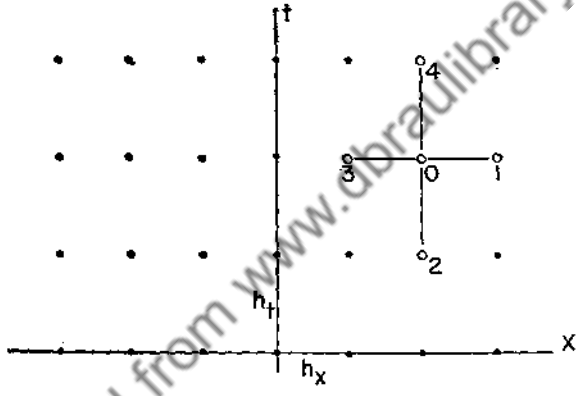


FIGURE 3.35

$t = 0$, after which equation (273b), expressed as an appropriate difference relation, serves to determine ϕ at points of the parallel line $t = h_t$. From this stage onward, (277) determines values of ϕ at points of the lines $t = 2h_t$, $3h_t$, and so forth, by step-by-step calculation, the values along each line depending only upon values in the *two* preceding lines.

It is easily seen that the calculated value of ϕ at the point P of Figure 3.36 depends only upon the calculated values at the indicated points of that figure, all of which lie within or on the boundary of the region bounded by the line $t = 0$ and the two lines $x - \kappa Vt =$ constant and $x + \kappa Vt =$ constant which pass through P. This region may be called the *region of determination* for the point P, relevant to the difference equation (275) with $h_x = \kappa V h_t$. For

any net, no matter how fine, of which P is a net point and for which (276) holds, this region is invariant, once the ratio κ has been chosen.

Suppose now that the ratio κ is *less than unity*. Then the region of determination for the *difference* equation lies completely inside the corresponding region for the *differential* equation which governs the exact problem, and it follows that the solution of the difference equation takes on a value at P which does not depend upon prescribed values of ϕ at points of the interval AB (Figure 3.34) which are at and near the ends of that interval. Since this situation violates (274) and continues to hold in the limit as the net spacings

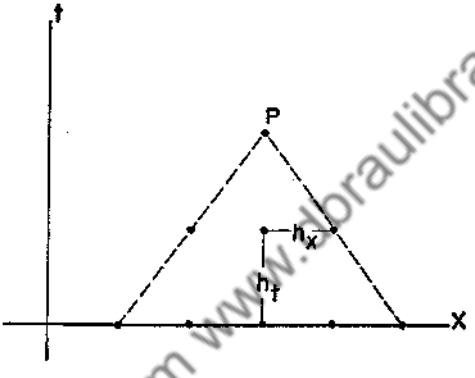


FIGURE 3.36

tend to zero, it follows that the solution of the difference equation does *not*, in general, converge to the solution (274) of the true problem when $\kappa < 1$, that is, when $h_x < V h_t$. However, it is known that convergence to the true solution *does* follow when $\kappa \geqq 1$, that is, when the spacings satisfy the condition

$$h_x \geqq V h_t, \tag{278}$$

so that the region of determination for the *difference* equation *includes or coincides with* the corresponding region for the *differential* equation (see Reference 7).

The same statement applies in the more general case when (272) governs a problem in which ϕ *and* $\partial\phi/\partial n$ are prescribed along any initial line or curve *which does not coincide with a characteristic.* In addition, ϕ *or* $\partial\phi/\partial n$ may be prescribed along one or two lines or curves, originating at points of the initial curve and of semi-infinite

extent on one side of the initial curve, but not intersecting each other, the solution then being required in the semi-infinite "curvilinear strip" bounded by these curves.

The variable t was used in (272) to represent one of the independent variables because of the fact that, in many problems governed by that equation and its generalizations, that variable is identified with time while the variable x represents distance. The constant V then has the dimensions of a velocity. However, in other cases (in which the more general conditions mentioned in the preceding paragraph may often apply) both variables may represent distances, so that the xt-plane is truly a physical plane.

In an interpretation of the first type, ϕ may represent the displacement of a point of a uniform string which is executing small free vibrations in a plane. The constant V^2 is then the ratio of the tension (assumed to be large and uniform) to the linear density of the string. The variables x and t then represent distance, measured along the string, and time, respectively. At the time $t = 0$, the displacement ϕ and velocity $\partial\phi/\partial t$ of each point are prescribed, while at the ends $x = a$ and $x = b$ the deflection (or slope) is prescribed as a function of t. The case considered explicitly was that in which the string is considered to be of infinite length, and in which the string is released from a prescribed initial position with zero initial velocity.

The simplest permissible choice of the spacing ratio is that for which $h_x = V h_t$, so that $\kappa = 1$ and equation (277) takes the form

$$\phi_4 = \phi_1 + \phi_3 - \phi_2, \tag{279}$$

with the notation of Figure 3.35, and does not involve the value ϕ_0. The corresponding net in the xt-plane then has the property that the sides of the parallelogram determined by the four relevant points lie on the characteristics of the governing differential equation. The advantage of this situation follows from the known fact that irregularities in the solution of (272) can exist only along its characteristics. (In consequence of this possibility, it is found that a characteristic cannot, in general, be taken as an initial curve in the formulation of a problem, if that problem is to possess a unique solution.)

Analogous situations exist in the case of problems governed by other hyperbolic equations. For an equation of the form

$$a\,\phi_{xx} + 2b\,\phi_{xy} + c\,\phi_{yy} + d\,\phi_x + e\,\phi_y + f\,\phi = g, \qquad (280)$$

the characteristics are those curves in the xy-plane which satisfy the differential equation

$$a\,(dy)^2 - 2b\,dx\,dy + c\,(dx)^2 = 0. \qquad (281)$$

Unless (280) is hyperbolic or parabolic ($b^2 \geqq ac$), this equation cannot have real solutions. However, for a hyperbolic equation two distinct families of curves are determined; for a parabolic equation the two families become coincident. The region of determination for a point P, in an initial-value problem, is then bounded by the two characteristics which pass through P, and by the arc of the initial curve which they intercept. In order to insure convergence of a finite-difference approximation, the spacings should be so related that the region of determination for each point P, relevant to the difference equation, nowhere lies interior to the corresponding region for the differential equation.

It is often convenient to transform a linear problem of hyperbolic type by choosing new independent variables u and v in such a way that the characteristics of the transformed equation are the straight lines u = constant and v = constant. The introduction of a rectangular net in a plane in which u and v are rectangular coordinates (the sides of the rectangles being parallel to those axes) then serves to make the regions of determination for net points, relevant to the difference equation, coincide with the regions which are appropriate to the differential equation. For this purpose, in the case of (280) we may set

$$u = P(x, y), \qquad v = Q(x, y), \qquad (282a)$$

where $P(x, y)$ = constant and $Q(x, y)$ = constant represent two independent solutions of (281). It is then easily shown that (280) takes the form

$$\phi_{uv} + A\,\phi_u + B\,\phi_v + C\,\phi = D, \qquad (282b)$$

where A, B, C, and D are functions of the new independent variables u and v (see Problems 90 and 91). The use of (282a) permits the mapping of the relevant boundaries and associated prescribed conditions from the original xy-plane to the new uv-plane.

Thus, for example, if we make the change in variables $u = Vt - x$ and $v = Vt + x$ in equation (272) the new equation takes

the form $\partial^2\phi/\partial u\,\partial v = 0$, with the characteristics u = constant and v = constant. The initial line $t = 0$ maps into the line $u + v = 0$, and the strip $[t > 0, a < x < b]$ maps into the diagonal strip bounded by the lines $u + v = 0$, $v - u = 2a$, and $v - u = 2b$ (Figure 3.37). If equal spacings are taken in the u- and v-directions, the relevant difference equation takes the form $\phi_1 - \phi_2 + \phi_3 - \phi_4 = 0$. The two prescribed conditions at the end of the strip, as well as the single condition prescribed along each side, are transformed (point by point) so as to apply to the net points on the boundary of the new configuration. The solution is then obtained in the usual way, the value of ϕ obtained at a net point being also the required value at the corresponding point of the original configuration.

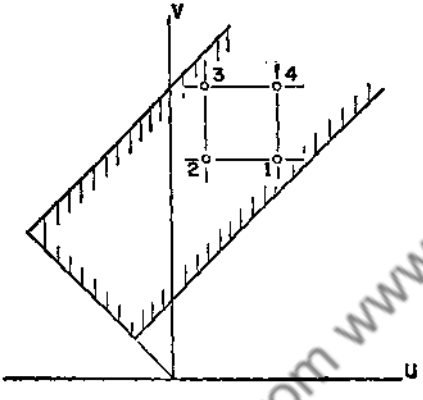


FIGURE 3.37

It is clear that little is gained by the transformation in the special case just considered, since the new net is merely diagonal to the square net which would have been obtained in an xy-plane in which y is identified with Vt. However, in the more general case, in which the characteristics may be curves, rather than straight lines, it is often extremely important that the "frontier" of the advancing calculation lie on a characteristic, in order that proper account can be taken of possible irregularities which are propagated only along characteristics. This situation is brought about by the method outlined above.

3.22. Instability. In addition to the question of convergence of a finite-difference approximation, as the mesh spacings tend to zero, there is a further difficulty relevant to initial-value problems, the nature of which may be illustrated by the considerations which follow.

Suppose that we require the function $T(x, t)$ which satisfies the differential equation

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} \qquad (0 < x < \pi, \quad t > 0), \tag{283}$$

together with the initial condition

$$T(x, 0) = \sin rx, \tag{284}$$

where r is an integer, and the two end conditions

$$T(0, t) = 0, \qquad T(\pi, t) = 0 \qquad (t > 0). \tag{285a,b}$$

The exact solution of this problem is given by the expression

$$T(x, t) = e^{-r^2 t} \sin rx. \tag{286}$$

Corresponding to the differential equation (283), we may consider the difference equation

$$\frac{T(x, t + h_t) - T(x, t)}{h_t} = \frac{T(x + h_x, t) - 2T(x, t) + T(x - h_x, t)}{h_x^2} \tag{287}$$

where h_x and h_t are spacings in the x- and t-directions. If we write the coordinates of the mesh points in the form

$$x_m = m\, h_x, \qquad t_n = n\, h_t \tag{288}$$

and introduce the abbreviation

$$\kappa = \frac{h_x^2}{h_t}, \tag{289}$$

this equation can be written in the form

$$\kappa(T_{m,n+1} - T_{m,n}) = T_{m+1,n} - 2T_{m,n} + T_{m-1,n} \tag{290}$$

where

$$T_{m,n} \equiv T(m\, h_x, n\, h_t). \tag{291}$$

If there are $M - 1$ division points along the x-axis, the requirement that (284) be satisfied at these points becomes

$$T_{m,0} = \sin \frac{mr\pi}{M} \qquad (m = 1, 2, \cdots, M - 1), \tag{292}$$

whereas the conditions (285a,b) take the form

$$T_{0,n} = 0, \qquad T_{M,n} = 0 \qquad (n = 1, 2, \cdots). \tag{293}$$

In the remainder of this section, we obtain the explicit solution of this problem, and compare it with the solution (286) of the true problem. The method of solution is completely analogous to the

method of "separation of variables" which is often useful in dealing with partial differential equations.

If we write

$$T_{m,n} = f_m g_n, \tag{294}$$

where f is independent of n, and g independent of m, equation (290) can be rewritten in the form

$$\kappa \frac{g_{n+1} - g_n}{g_n} = \frac{f_{m+1} - 2f_m + f_{m-1}}{f_m}. \tag{295}$$

The equal members of (295) are clearly independent of both m and n. If we denote their common value by the constant $-\lambda$, we obtain the following difference equations which must be satisfied by f and g:

$$f_{m+1} - (2 - \lambda) f_m + f_{m-1} = 0, \tag{296}$$

$$g_{n+1} - \left(1 - \frac{\lambda}{\kappa}\right) g_n = 0. \tag{297}$$

Reference to equations (111a–d) of Section 3.6 shows that (296) possesses a nontrivial solution for which $f_0 = f_M = 0$, in accordance with the requirement (293), if and only if the constant λ takes on one of the $M - 1$ distinct values

$$\lambda_s = 4 \sin^2 \frac{s\pi}{2M}, \tag{298}$$

where s is an integer, in which case that solution is of the form

$$f_m = C \sin \frac{ms\pi}{M}, \tag{299}$$

where C is an arbitrary constant. It is clear that the initial condition (292) will be satisfied if we take

$$s = r, \qquad C = 1, \tag{300}$$

and require that

$$g_0 = 1. \tag{301}$$

Accordingly, with the choice $\lambda = 4 \sin^2 (r\pi/2M)$, the appropriate solution of (297) is found to be

$$g_n = \left(1 - \frac{4}{\kappa} \sin^2 \frac{r\pi}{2M}\right)^n, \tag{302}$$

and the required solution $T_{m,n} = f_m g_n$ is determined in the form

$$T_{m,n} = \left(1 - \frac{4}{\kappa}\sin^2\frac{r\pi}{2M}\right)^n \sin\frac{mr\pi}{M}. \tag{303}$$

In order to compare this solution with the solution (286) of the exact problem, we replace m by x_m/h_x, n by t_n/h_t, and $M\,h_x$ by π, in accordance with (288), and so rewrite (303) in the more explicit form

$$T(x_m, t_n) = \left(1 - \frac{4}{\kappa}\sin^2\frac{rh_x}{2}\right)^{t_n/h_t} \sin rx_m. \tag{304}$$

If we write

$$h_x = h, \qquad h_t = \frac{h^2}{\kappa}, \tag{305}$$

in accordance with (289), the coefficient of $\sin rx_m$ in (304) can be expressed in the form

$$\left(1 - \frac{4}{\kappa}\sin^2\frac{rh}{2}\right)^{t_n/h^2}, \tag{306}$$

after which elementary considerations show that this quantity tends to the limit e^{-r^2t} as h approaches zero, *for any fixed positive value of the spacing ratio* κ.

Hence, it follows that the solution (304) of the difference-equation formulation of the problem does indeed tend to the solution (286) of the true problem as the spacings tend to zero, for any fixed value of the spacing ratio $\kappa = h_x^2/h_t$. The same result clearly obtains when the right-hand member of (284) is replaced by a *finite* sum of terms for which r takes on different integral values.

However, if we examine (304) more closely, we may notice that if h_x is so chosen that

$$\kappa < 2\sin^2\frac{rh_x}{2} \tag{307}$$

it follows that the quantity inside parentheses in (304) is a negative quantity *with absolute value greater than unity*, so that, in this case, the coefficient of $\sin rx_m$ will oscillate with ever-increasing amplitude as t_n/h_t takes on increasing integral values. Thus it follows that, when (307) is satisfied, the "approximate solution" will oscillate with increasing amplitude as t increases, and hence will fail

completely to approximate the true solution, which decreases exponentially with t for any fixed value of x.

Further, in the case when $\kappa = 2 \sin^2 (rh_x/2)$, the solution (304) takes the form

$$T(x_m, t_n) = (-1)^n \sin rx_m,$$

and hence oscillates with constant amplitude from point to point in the t-direction along the net.

While it is true that, for any prescribed fixed value of r, and for any chosen fixed value of the ratio κ, a process of successively refining the net will eventually lead to a nonoscillatory approximation which tends toward the exact solution with continued refinement, it is clearly desirable to choose κ once and for all in such a way that the inequality (307) is reversed for any finite values of r and h_x. This situation is attained if and only if we require that the ratio be such that

$$\kappa = \frac{h_x^2}{h_t} \geqq 2, \tag{308}$$

the equality sign being permissible because of the fact that the coefficient of $4/\kappa$ in (303) is unity only when the factor $\sin (mr\pi/M)$ vanishes.

By superimposing solutions of the form (303), we find that the expression

$$T_{m,n} = \sum_{r=1}^{M-1} C_r \left(1 - \frac{4}{\kappa} \sin^2 \frac{r\pi}{2M}\right)^n \sin \frac{mr\pi}{M} \tag{309}$$

is the solution of the difference equation (290), subject to the end conditions (293), and to the initial condition

$$T_{m,0} = f_m \equiv \sum_{r=1}^{M-1} C_r \sin \frac{mr\pi}{M} \quad (m = 1, 2, \cdots, M-1), \tag{310}$$

where reference to equation (180) shows that

$$C_r = \frac{2}{M} \sum_{m=1}^{M-1} f_m \sin \frac{mr\pi}{M}. \tag{311}$$

In particular, if we require that

$$f_m = \delta_{m_0 m} \qquad (0 < m_0 < M), \tag{312}$$

so that $T'_{m,0}$ vanishes except when $m = m_0$, and is unity for that value of m, equation (311) leads to the determination

$$C_r = \frac{2}{M} \sin \frac{m_0 r\pi}{M},$$

and hence the corresponding solution (309) is obtained in the form

$$T_{m,n} = \frac{2}{M} \sum_{r=1}^{M-1} \left(1 - \frac{4}{\kappa} \sin^2 \frac{r\pi}{2M}\right)^n \sin \frac{m_0 r\pi}{M} \sin \frac{mr\pi}{M}. \quad (313)$$

The term for which $r = M - 1$ involves the factor

$$\left[1 - \frac{4}{\kappa} \sin^2 \frac{(M-1)\pi}{2M}\right]^n = \left(1 - \frac{4}{\kappa} \cos^2 \frac{\pi}{2M}\right)^n,$$

which is unbounded in absolute value as $n \to \infty$ unless

$$\kappa \geqq 2 \cos^2 \frac{\pi}{2M}. \quad (314)$$

When (314) is satisfied, it is easily seen that all terms in (313) tend exponentially to zero as $n \to \infty$. (In the case of equality, the single term for which $r = M - 1$ oscillates with constant amplitude.)

Thus we may conclude that the presence of an initial *numerical inaccuracy* (such as a round-off error) at $m = m_0$ will lead to an error in the approximate solution which increases exponentially in magnitude with n (that is, with time) when and only when (314) is violated. In this case, the approximate procedure is said to be *unstable;* when (314) is satisfied, the procedure is said to be *stable.* As the net is continually refined ($M \to \infty$), the stability criterion (314) tends to the requirement (308).

In the case of a general initial condition of the form (310), in which the right-hand member may tend to an infinite series of the form

$$T(x, 0) = f(x) \equiv \sum_{r=1}^{\infty} C_r \sin rx,$$

as the net is continually refined, the question of *convergence* of the sequence of solutions to the solution of the true problem, as the spacings tend to zero, is of some difficulty when (308) is violated. It appears that such convergence usually does not obtain when the procedure is unstable. However, the solution (303), corresponding

to the special one-term initial function (292), illustrates the fact that this correlation between instability and lack of convergence is not perfectly general. Also, we may recall that, in that special case, the *exact* solution of the difference equation oscillates about the exact solution of the approximated differential equation with unbounded amplitude as n increases, when (307) is satisfied. When the condition

$$2 \sin^2 \frac{rh_x}{2} \leqq \kappa < 2 \sin^2 \frac{(M-1)h_x}{2}$$

is satisfied, the procedure is still *unstable*, but increasing oscillation of the *exact* solution is *not* present. Thus it follows that neither ultimate lack of convergence (as $M \to \infty$) nor infinite oscillation of the exact solution (as $n \to \infty$) is inevitably implied by instability. However, by definition, instability *does* imply a tendency for the effect of any numerical inaccuracy to increase unboundedly as $n \to \infty$.

When the spacing ratio satisfies (308), both stability (for a given mesh fineness) and convergence (with continued mesh refinement) are obtained when $T(x, 0)$ is prescribed in a regular way.

For the difference equation (221), an obvious change of variables shows that the stability criterion (308) is replaced by the condition $h_x^2 \geqq 2\alpha^2 h_t$. It may be noticed that the choice made in equation (222) is in accordance with this requirement.

3.23. Stability criteria. In this section, we first obtain a criterion for stability of a five-point difference equation of the form

$$w_{m+1,n} - 2a\, w_{m,n} + b^2 w_{m-1,n} = c(w_{m,n+1} + d\, w_{m,n-1}), \quad (315)$$

where a, b, c, and d are real *constants*, under the assumption that end conditions of the rather general form

$$w_{0,n} = \mu_1 w_{1,n} + u_n, \qquad w_{M,n} = \mu_2 w_{M-1,n} + v_n$$
$$(0 \leqq \mu_1 \leqq b^{-1}, \quad 0 \leqq \mu_2 \leqq b) \quad (316)$$

are imposed when $m = 0$ and $m = M$, and initial conditions are imposed when $n = 0$ and $n = 1$ (or merely at $n = 0$ if $d = 0$). It will be convenient to refer to end conditions which are special forms of those listed in (316) as *proper* end conditions for (315). The solution is required for positive n and for $0 \leqq m \leqq M$. Many difference-equation approximations to linear second-order differ-

ential equations with constant coefficients are taken in this form. As will be seen, the procedure can be easily generalized to certain more involved linear equations with constant coefficients. Also, if the coefficients depend upon m and/or n, it may be possible to divide the region into subregions in such a way that the coefficients may be replaced by constant average values in each subregion, and to apply the criterion to each subregion separately. Needless to say, this last procedure is heuristic, and *cannot* be guaranteed to be valid.

In studying the propagation of a numerical inaccuracy, we must replace any nonhomogeneous end conditions by corresponding homogeneous ones. Thus, for example, if w is *prescribed* as a function of n along the boundaries $m = 0$ and $m = M$, the propagated error (from any source) must *vanish* along those boundaries.

Proceeding as in the preceding section, we seek solutions of (315) of the product form

$$w_{m,n} = f_m g_n,$$

and find that f_m and g_n must accordingly satisfy the relation

$$\frac{f_{m+1} - 2a f_m + b^2 f_{m-1}}{f_m} = c \frac{g_{n+1} + d g_{n-1}}{g_n} = -\lambda, \tag{317}$$

where λ is an arbitrary constant. Hence there must follow

$$f_{m+1} - (2a - \lambda) f_m + b^2 f_{m-1} = 0 \tag{318}$$

and

$$g_{n+1} + \frac{\lambda}{c} g_n + d\, g_{n-1} = 0. \tag{319}$$

Since only b^2 appears in (315), and b has been assumed to be real, there is no loss of generality in taking b to be nonnegative,

$$b \geqq 0. \tag{320}$$

In order to obtain the solution of (318) in a convenient form, it is desirable to write

$$2a - \lambda \equiv 2b \cos \alpha, \tag{321}$$

so that (318) takes the form

$$f_{m+1} - 2b f_m \cos \alpha + b^2 f_{m-1} = 0,$$

and the methods of Section 3.4 lead to the solution

$$f_m = b^m(c_1 \sin \alpha m + c_2 \cos \alpha m). \tag{322}$$

The imposition of *proper homogeneous* end conditions for $m = 0$ and $m = M$ then determines $M - 1$ permissible values of the parameter α, say $\alpha = \alpha_r$ $(r = 1, 2, \ldots, M - 1)$, and, for each such α_r, relates the coefficients c_1 and c_2. For present purposes, it is not necessary to effect this determination explicitly. However, it is important to notice that (in virtue of the results of Problem 55) all permissible values of α are *real*, and that no solutions of (318) which are independent of those so obtained can satisfy the prescribed end conditions.

For each such value of α_r, equation (321) determines the corresponding value of the separation constant λ,

$$\lambda_r = 2(a - b \cos \alpha_r), \tag{323}$$

and the corresponding function g_n is determined from the equation obtained by introducing (323) into (319),

$$g_{n+1} + \frac{2}{c}(a - b \cos \alpha_r)g_n + d\, g_{n-1} = 0. \tag{324}$$

The general solution of (324) can be expressed in the form

$$g_n^{(r)} = A_r\, {}_1\beta_r{}^n + B_r\, {}_2\beta_r{}^n, \tag{325}$$

where ${}_1\beta_r$ and ${}_2\beta_r$ are the roots of the equation

$$\beta^2 + \frac{2}{c}(a - b \cos \alpha_r)\beta + d = 0, \tag{326}$$

if those roots are distinct.

The most general solution of (315), subject to the prescribed homogeneous end conditions, is then of the form

$$w_{m,n} = \sum_{r=1}^{M-1} (A_r\, {}_1\beta_r{}^n + B_r\, {}_2\beta_r{}^n) f_m^{(r)}, \tag{327}$$

where $f_m^{(r)}$ is a convenient multiple of the appropriate form of (322). The coefficients A_r and B_r, which may be complex, are determined finally by initial error distributions when $n = 0$ and $n = 1$, or for any two consecutive values of n. When $d = 0$, only one nonzero root of (326) is obtained, and the one resulting set of coefficients in (327) is determined by a single initial error distribution. When the

roots happen to be *equal*, the second term in parentheses in (327) is replaced by $B_r\, n\, {}_1\beta_r{}^n$.

Thus it follows that *any error distribution is of the form* (327), *appropriately modified in the case of equal roots.* In order that no error distribution shall grow exponentially in magnitude as $n \to \infty$, it is necessary and sufficient that the constants ${}_1\beta_r$ and ${}_2\beta_r$ be not larger than unity in absolute value. This statement applies also in the case of equal roots. If $\beta = +1$ or -1 happens to be a *double* root, the contents of the parentheses are replaced by $A_r + B_r n$ or $(-1)^n(A_r + B_r n)$, and *linear* instability (in which the error may grow linearly with n) may be present. Hence we obtain the following *stability criterion:*

The difference equation (315) *is stable, for arbitrarily prescribed proper* end conditions, if and only if*

(1) *the roots of the equation*

$$\beta^2 + \frac{2}{c}(a - b\cos\alpha)\beta + d = 0 \tag{328}$$

cannot exceed unity in absolute value for any real values of α, *and*

(2) *neither* $\beta = +1$ *nor* $\beta = -1$ *may be a repeated root.*

If (1) *is satisfied, but either* $\beta = +1$ *or* $\beta = -1$ *may be a repeated root, then* (315) *may be linearly unstable.*

We may notice that this criterion is independent of the nature of the end conditions, so long as they are *proper* ones. For any *specific* end conditions of this type, this requirement is slightly conservative, since then only those values of α for which (322) can satisfy the corresponding homogeneous end conditions need be considered in (328).

Linear instability can occur only if $d = 1$ and the equation $a - b\cos\alpha = \pm c$ can be satisfied by a permissible real value of α. It is usually not troublesome in numerical work.

If, in illustration, we apply this criterion to equation (290) of the preceding section, we have

$$a = 1 - \frac{\kappa}{2}, \quad b = 1, \quad c = \kappa, \quad d = 0,$$

* The restrictions in (316) may be replaced by the condition $0 \leqq \mu_1 \leqq M\,b^{-1}/(M-1)$ if $\mu_2 = 0$ or by $0 \leqq \mu_2 \leqq M\,b/(M-1)$ if $\mu_1 = 0$ [see Problem 55(f)]. When the restrictions on μ_1 and μ_2 are violated, a *complex* value of α may be permissible and the criterion obtained then may not be valid.

and (328) takes the form

$$\beta^2 + \frac{2}{\kappa}\left(1 - \frac{\kappa}{2} - \cos\alpha\right)\beta = 0. \tag{329}$$

Clearly, *linear* instability cannot exist. The nonzero root of (329) is given by

$$\beta = 1 - \frac{2}{\kappa}(1 - \cos\alpha), \tag{330}$$

and the stability criterion then requires that

$$-1 \leqq 1 - \frac{2}{\kappa}(1 - \cos\alpha) \leqq 1, \tag{331}$$

for all real α. The right-hand inequality is nonrestrictive for any $\kappa > 0$, whereas the left-hand inequality gives

$$\kappa \geqq 1 - \cos\alpha. \tag{332}$$

If this condition is to hold for *all* real α, there must follow $\kappa \geqq 2$, in accordance with (308). In the case of the specific end conditions which prescribe $w_{m,n}$ when $m = 0$ and $m = M$, the permissible values of α were found in the preceding section to be of the form

$$\alpha_r = \frac{r\pi}{M} \qquad (r = 1, 2, \cdots, M - 1). \tag{333}$$

The right-hand member of (332) then takes on its maximum permissible value when $r = M - 1$, and (332) then reduces to the criterion (314).

As a second application of this criterion, we consider the result of replacing the differential equation $\partial^2 w/\partial x^2 + \partial w/\partial x = \partial w/\partial t$ by the difference equation

$$\frac{w_{m+1,n} - 2w_{m,n} + w_{m-1,n}}{h_x^2} + \frac{w_{m+1,n} - w_{m,n}}{h_x} = \frac{w_{m,n+1} - w_{m,n}}{h_t}, \tag{334}$$

with the usual abbreviation $w_{m,n} \equiv w(m\,h_x, n\,h_t)$. If we write

$$\kappa \equiv \frac{h_x^2}{h_t}, \qquad h \equiv h_x, \tag{335}$$

this equation can be written in the form

$$(1 + h)w_{m+1,n} - (2 + h - \kappa)w_{m,n} + w_{m-1,n} = \kappa\, w_{m,n+1}, \tag{336}$$

and is identified with (315) by writing

$$a = \frac{2 + h - \kappa}{2(1 + h)}, \quad b = \frac{1}{\sqrt{1 + h}}, \quad c = \frac{\kappa}{1 + h}, \quad d = 0. \tag{337}$$

The nonzero root of (328) is then found to be

$$\beta = 1 - \frac{2 + h - 2\sqrt{1 + h}\cos\alpha}{\kappa}. \tag{338}$$

In view of the inequality $2 + h \geqq 2\sqrt{1 + h}$, the requirement $\beta \leqq 1$ is nonrestrictive; the requirement $\beta \geqq -1$ leads to the inequality

$$\kappa \geqq 1 + \frac{h}{2} - \sqrt{1 + h}\cos\alpha. \tag{339}$$

If the difference equation is to be stable for *all* proper end conditions, the right-hand member is maximized when $\cos\alpha = -1$, and the desired stability criterion becomes

$$\kappa \geqq 1 + \frac{h}{2} + \sqrt{1 + h}. \tag{340}$$

If the end conditions prescribe w when $m = 0$ and $m = M$, permissible values of α are of the form $\alpha_r = r\pi/M$, as before, and the maximum permissible value of the right-hand member of (340) corresponds to $\cos\alpha = \cos\alpha_{M-1} = -\cos(\pi/M) = -\cos(\pi h/L)$, where $L \equiv Mh$ is the length of the range in x. Thus the sharper criterion in this specific case is of the form

$$\kappa \geqq 1 + \frac{h}{2} + \sqrt{1 + h}\cos\frac{\pi h}{L}. \tag{341}$$

For reasonably small values of the ratio h/L, (341) differs only slightly from (340).

It should be pointed out that there also exists in the literature* a rather widely used criterion, often associated with the name of von Neumann, for testing the stability of linear difference equations with constant coefficients. This procedure consists in first directly assuming a solution of the difference equation in the form

$$w_{m,n} = \beta^n e^{im\phi}. \tag{342}$$

* See, for example, Reference 11.

Substitution of this assumption into the difference equation, and subsequent cancellation of the resultant common factor $\beta^n e^{im\phi}$, then leads to an equation which must be satisfied by the parameters β and ϕ. The von Neumann criterion for stability is the requirement that it be impossible to satisfy this equation by any real or complex value of β for which $|\beta| > 1$, when ϕ takes on all *real* values.

If this procedure is applied to equation (315), the necessary relationship involving β and ϕ is obtained in the form

$$c\left(\beta + \frac{d}{\beta}\right) = e^{i\phi} - 2a + b^2 e^{-i\phi}$$

or, after a rearrangement,

$$\beta^2 + \frac{1}{c}(2a - e^{i\phi} - b^2 e^{-i\phi})\beta + d = 0. \tag{343}$$

The von Neumann criterion for stability of (315) is thus the requirement that the roots of this equation be not larger than unity in absolute value, for all real values of ϕ.

When $b = 1$, it is seen that equations (343) and (328) become formally identical. Thus (excluding the consideration of *linear* instability) the two criteria are identical in this special case. When $b \neq 1$, it can be shown that the von Neumann criterion is *conservative* in the present case; that is, this criterion will predict instability when instability exists, but it may also predict instability when the equation is actually stable.*

As a specific illustration, we again consider the difference equation (336). With the data of (337), the nonzero root of equation (343) is found to be

*The criterion was proposed by von Neumann for difference equations (with constant coefficients) in which the range of m is infinite ($-\infty < m < +\infty$), and *it provides a condition which is indeed necessary and sufficient in such cases,* since then only product solutions of the form (342), *with ϕ real,* can remain finite as $m \to \pm\infty$. In those cases when the net of definition is confined to a strip, and end conditions (of various types) are imposed along the lateral boundaries of the strip, *complex* values of ϕ may be admissible and the criterion should be applied with some caution. In particular, if "proper" end conditions relevant to a stable formulation were replaced by "improper" ones (in the present terminology), the new formulation would generally become unstable whereas the von Neumann criterion would, of course, continue to predict stability.

$$\beta = 1 - \frac{2(1 - \cos\phi) + h(1 - e^{i\phi})}{\kappa}$$

$$= \frac{1}{\kappa}\left[\kappa - (2 + h)(1 - \cos\phi) + i\,h \sin\phi\right],$$

and the requirement $|\beta|^2 \leqq 1$ takes the form

$$\frac{1}{\kappa^2}\left\{[\kappa - (2 + h)(1 - \cos\phi)]^2 + h^2 \sin^2\phi\right\} \leqq 1. \qquad (344)$$

The quantity on the left is easily shown to take on its maximum value (for fixed h and κ) when $\phi = \pi$. Hence the von Neumann criterion is found to be

$$\left|1 - \frac{2(2 + h)}{\kappa}\right| \leqq 1 \qquad \text{or} \qquad \kappa \geqq 2 + h. \qquad (345)$$

If h is small relative to unity, this restriction differs only slightly from the correct restriction (340), as may be seen from the expansion $\sqrt{1 + h} = 1 + \frac{1}{2}h - \frac{1}{8}h^2 + \cdots$. We may verify also that (345) is indeed conservative in this case, by noticing that $\sqrt{1 + h} \leqq 1 + \frac{1}{2}h$.

The fact that (345) and (340) differ by little when h is small might have been anticipated from the fact that the coefficient b defined in (337) differs by little from unity when h is small. This situation always exists when (315) is an approximation to a *differential equation.* In cases when b differs more appreciably from unity, a more significantly overconservative estimate may be expected from the von Neumann procedure.

It is of some importance to notice that this procedure would reduce to the present one, in the case of (315), if the assumption (342) were replaced by the assumption

$$w_{m,n} = \beta^n b^m e^{im\alpha} \qquad (346)$$

by writing $\phi = \alpha - i \log b$, where now α is to take on real values.

To conclude this section, we generalize the preceding analysis to the *nine-point* difference equation

$$(a_3 w_{m+1,n+1} + b_3 w_{m,n+1} + c_3 w_{m-1,n+1}) + (a_2 w_{m+1,n} + b_2 w_{m,n} + c_2 w_{m-1,n}) + (a_1 w_{m+1,n-1} + b_1 w_{m,n-1} + c_1 w_{m-1,n-1}) = 0, \qquad (347)$$

in which the coefficients are constants, subject to the requirement that corresponding a's and c's are in a constant ratio,

$$c_1 = \rho^2 a_1, \quad c_2 = \rho^2 a_2, \quad c_3 = \rho^2 a_3, \tag{348}$$

where ρ is *real* and may be taken to be positive. We suppose that end conditions of the type described by (316) or by the footnote on page 337, *with* b *replaced by* ρ, are imposed for $m = 0$ and $m = M$ and that appropriate initial conditions are imposed for two consecutive values of n.* *Error* distributions propagated from numerical accuracies introduced at a given stage in the calculation must then satisfy corresponding *homogeneous* end conditions for succeeding values of n, as before.

If a product solution of the form

$$w_{m,n} = f_m g_n$$

is assumed, and use is made of (348), equation (347) then can be separated in the form

$$-\frac{f_{m+1} + \rho^2 f_{m-1}}{f_m} = \frac{b_3 g_{n+1} + b_2 g_n + b_1 g_{n-1}}{a_3 g_{n+1} + a_2 g_n + a_1 g_{n-1}} = \lambda,$$

where λ is an arbitrary constant. Thus f_m and g_n must satisfy the equations

$$f_{m+1} + \lambda f_m + \rho^2 f_{m-1} = 0 \tag{349a}$$

and

$$(b_3 - \lambda a_3)g_{n+1} + (b_2 - \lambda a_2)g_n + (b_1 - \lambda a_1)g_{n-1} = 0. \tag{349b}$$

Since (349a) is identified with (318) by writing $a = 0$ and $b = \rho$, it follows that all characteristic values of λ, corresponding to end conditions of the class described above, are of the form

$$\lambda = -2\rho \cos \alpha, \tag{350}$$

where α *is real.* If this relation is introduced into (349b), the result of setting $g_n = \beta^n$ is the equation

$$(b_3 + 2\rho a_3 \cos \alpha)\beta^2 + (b_2 + 2\rho a_2 \cos \alpha)\beta + (b_1 + 2\rho a_1 \cos \alpha) = 0, \tag{351}$$

which generalizes (328). The stability criterion is thus the requirement that the roots of (351) not exceed unity in absolute value for

* If $a_1 = b_1 = c_1 = 0$, only *one* initial condition is to be prescribed.

all real values of α, and that neither $\beta = 1$ nor $\beta = -1$ may be a repeated root. For any *specific* admissible end conditions this requirement is slightly conservative, as in the preceding case, since only a finite number of values of α then need be considered.

In connection with equations (328) and (351), use may be made of the easily established fact that if the coefficients of the equation

$$\beta^2 + A\,\beta + B = 0 \tag{352}$$

are *real*, necessary and sufficient conditions that neither root of that equation shall exceed unity in absolute value are that the inequalities

$$|\,A\,| \leqq B + 1 \leqq 2 \tag{353}$$

be satisfied.

In illustration, we analyze the equation

$$\kappa(w_{m,n+1} - w_{m,n}) = \tfrac{1}{2}[(w_{m+1,n+1} - 2w_{m,n+1} + w_{m-1,n+1}) + (w_{m+1,n} - 2w_{m,n} + w_{m-1,n})], \tag{354}$$

which was proposed by von Neumann (see Reference 11) as an approximation to the heat-flow equation (283) (with unit diffusivity), in place of the simpler approximation (290). As before, we have written

$$\kappa = \frac{h_x{}^2}{h_t}. \tag{355}$$

We may identify (354) with (347) by writing

$$a_1 = b_1 = c_1 = 0, \qquad a_2 = c_2 = a_3 = c_3 = -\tfrac{1}{2},$$

$$b_2 = 1 - \kappa, \qquad b_3 = 1 + \kappa,$$

and (348) is satisfied by taking

$$\rho = 1,$$

after which the nonzero solution of (351) is found to be

$$\beta = \frac{\kappa - 2\sin^2(\alpha/2)}{\kappa + 2\sin^2(\alpha/2)}. \tag{356}$$

Since the condition $|\,\beta\,| \leqq 1$ is satisfied for *any* positive value of κ, and for any real α, it follows that the formulation (354) is stable for

any spacing ratio h_x^2/h_t (when proper end conditions are imposed), whereas formulation (290), subject to (293), is stable only when

$$\kappa \geqq 2 \cos^2 \frac{\pi}{2M}.$$

In practice, it is often found that the time increment must be taken to be inconveniently small in order to insure stability of a procedure based on the result of inserting the actual diffusivity parameter in (290). The corresponding formulation of type (354) has the advantage that the spacing ratios are unrestricted. However, it possesses the disadvantage that the values to be determined at the nth stage of an advancing calculation are not expressed explicitly in terms of values which are *known* at that stage. In order to advance the calculation, it is necessary to solve $M - 1$ simultaneous linear algebraic equations in the $M - 1$ desired following entries.

It may be noticed that the von Neumann procedure is equivalent to the present one when (348) is satisfied by $\rho = 1$. In particular, that procedure leads to the preceding result in the case of equation (354) (see Reference 11). In other cases, this agreement generally is not present. *If* (347) *represents an approximation to a partial differential equation*, as a result of replacing derivatives by divided differences or combinations of divided differences, it is easily shown that ρ will differ from unity at worst by an amount which tends to zero as the net is continually refined, so that the stability prediction afforded by the von Neumann procedure in such cases will increase in accuracy with decreasing net spacing when proper end conditions are imposed. For spacings of practical size, however, the discrepancy when $\rho \neq 1$ may be of some consequence.

In this connection, it may be pointed out that if, in such cases, *either* procedure predicts stability when the relevant *spacing ratio* κ is such that $\kappa \geqq \kappa_0$ [or $\kappa \leqq \kappa_0$] *in the limit when the spacings tend to zero*, then a condition which is generally *sufficient* to insure *ultimate* stability (at *some* stage of the refinement) is that $\kappa > \kappa_0$ [or $\kappa < \kappa_0$]. This result is a consequence of the content of the preceding paragraph, combined with the further observation that the lower [or upper] limit of stability will vary in a continuous way as the net is continually refined, and is of some importance in theoretical considerations.

As has already been pointed out, only scattered information is available as to general *convergence* of solutions of difference-equation problems to solutions of approximated differential-equation problems, with increasing net refinement. However, the existing evidence indicates that, if *stability* is attained at some stage of the refinement, then *convergence* generally follows when the prescribed functions involved in the end conditions and initial conditions are sufficiently well behaved. Whereas it has been shown that *lack* of stability does not inevitably imply lack of convergence, this result is of limited practical significance since instability generally renders a numerical procedure useless unless special methods of controlling propagated errors are employed.

A technique which sometimes can be used to establish stability in a simple and direct way is illustrated by the content of Problem 105, which is easily generalized.

REFERENCES

1. Milne-Thomson, L. M.: *The Calculus of Finite Differences*, Macmillan and Company, Ltd., London, 1933.
2. Milne, W. E.: *Numerical Calculus*, Princeton University Press, Princeton, N. J., 1949.
3. Bennett, A. A., W. E. Milne, H. Bateman, and L. E. Ford: *Numerical Integration of Differential Equations*, Bull. Nat. Res. Council, 1933.
4. Fort, T.: *Finite Differences and Difference Equations in the Real Domain*, Oxford University Press, New York, 1948.
5. Southwell, R. V.: *Relaxation Methods in Engineering Science*, Oxford University Press, New York, 1940.
6. Southwell, R. V.: *Relaxation Methods in Theoretical Physics*, Oxford University Press, New York, 1946.
7. Courant, R., K. Friedrichs, and H. Lewy: "Über die partiellen Differenzengleichungen der mathematischen Physik," *Math. Ann.*, Vol. 100, pp. 32–74 (1928).
8. Liebmann, H.: "Die angenahrte Ermittlung harmonischer Functionen und konformer Abbildungen," *Sitzber. math. naturw. Abt. bayer. Akad. Wiss. München*, p. 385 (1918).
9. Phillips, H. B., and N. Wiener: "Nets and the Dirichlet Problem," *J. Math. Phys.*, Vol. 2, pp. 105–124 (1923).

10. Emmons, H. W.: "The Numerical Solution of Partial Differential Equations," *Quart. Applied Math.*, Vol. 2, No. 3, pp. 173–195 (1940).
11. O'Brien, G. G., M. A. Hyman, and S. Kaplan: "A Study of the Numerical Solution of Partial Differential Equations," *J. Math. Phys.*, Vol. 29, pp. 223–251 (1951).

PROBLEMS

Section 3.1.

1. (a) If y_k satisfies the difference equation

$$y_{k+1} - 2y_k \cos \alpha + y_{k-1} = 0 \qquad (k = 1, 2, \cdots)$$

and the initial conditions $y_0 = 0$, $y_1 = 1$, determine y_2, y_3, and y_4 in terms of the real constant α.

(b) Verify that the expression $y_k = (\sin k\alpha)/(\sin \alpha)$ satisfies the difference equation and the initial conditions, and that it agrees with the results obtained when $k = 2$, 3, and 4.

2. (a) If y_k satisfies the difference equation

$$y_{k+1} - \lambda y_k + y_{k-1} = 0 \qquad (k = 1, 2, 3)$$

and the end conditions $y_0 = 0$, $y_4 = 0$, determine those values of the constant λ for which a nontrivial solution exists. [Determine y_2, y_3, and y_4 successively in terms of y_1 and λ, and determine λ such that $y_4 = 0$ but $y_1 \neq 0$.]

(b) Verify that the permissible values of λ are $\lambda_n = 2 \cos (n\pi/4)$ where $n = 1$, 2, and 3, and that the corresponding solutions $y_{n,k}$ are arbitrary multiples of $\sin (n\pi k/4)$.

3. Let f_k denote the kth term of the sequence 1, 3, 6, 10, 15, 21, ...

(a) By considering differences, show that f_k satisfies the equation $f_{k+1} - 2f_k + f_{k-1} = 1$, with $f_1 = 1$ and $f_2 = 3$.

(b) Verify that the difference equation is satisfied by $f_k = c_1 + c_2 k + \frac{1}{2}k^2$, for any constant values of c_1 and c_2. Evaluate c_1 and c_2, and determine the 100th term of the sequence.

Section 3.2.

4. Reduce the difference equation

$$A_k \Delta^2 y_k + B_k \Delta y_k + C_k y_k = \phi_k$$

to an equation of the form

$$a_k y_{k+2} + b_k y_{k+1} + c_k y_k = \phi_k.$$

[Write $\Delta = E - 1$.]

5. (a) Derive the formula

$$y_{k+1} = y_1 + k\,\Delta y_1 + \frac{k(k-1)}{2\cdot 1}\,\Delta^2 y_1 + \cdots \quad (k = 1, 2, \cdots).$$

[Notice that $E^k = (1 + \Delta)^k$.]

(b) Apply this result in the determination of the kth term of the sequence of Problem 3.

6. (a) Derive the operational formula

$$\sum_{n=1}^{k} y_n = \frac{E^k - 1}{E - 1}\,y_1.$$

(b) By writing $E = 1 + \Delta$, and formally expanding the ratio $(E^k - 1)/(E - 1)$ in ascending powers of Δ, derive the summation formula

$$\sum_{n=1}^{k} y_n = \left[k + \frac{k(k-1)}{2!}\,\Delta + \frac{k(k-1)(k-2)}{3!}\,\Delta^2 + \cdots\right] y_1$$

$(k = 1, 2 \ldots)$. [Notice that the series on the right terminates after $N + 1$ terms if y_k is a polynomial in k of order N.]

7. Use the summation formula of Problem 6(b) to show that

$$\sum_{n=1}^{k} n^3 = k + 7\,\frac{k(k-1)}{2} + 12\,\frac{k(k-1)(k-2)}{6} + 6\,\frac{k(k-1)(k-2)(k-3)}{24}$$

$$= \tfrac{1}{4}k^2(k+1)^2.$$

[Form a table of differences of $f_k = k^3$ near $k = 1$.]

Section 3.3.

8. A continuous uniform beam rests on N equally spaced supports, with separation h, and is unloaded between successive supports (Figure 3.38). Show that the bending moment M_k at the kth support satisfies the difference equation

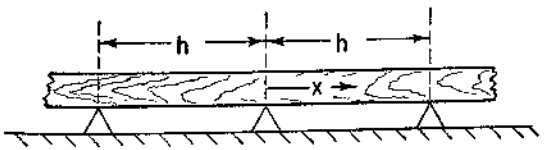


FIGURE 3.38

$$M_{k+1} + 4M_k + M_{k-1} = 0 \qquad (k = 2, 3, \cdots, N - 1),$$

where M_1 and M_N are determined by conditions of loading or support at the ends of the beam. [Denoting by x distance to the right from the kth support, show that $M(x) = M_k + (M_k - M_{k-1})\dfrac{x}{h}$ for $-h \leqq x \leqq 0$, and $M(x) = M_k + (M_{k+1} - M_k)\dfrac{x}{h}$ for $0 \leqq x \leqq h$. Recalling that the deflection $y(x)$ is governed by the equation $EIy'' = M$, where EI is the constant flexural rigidity, show that the requirements that y vanish when $x = 0$ and $\pm h$, and that y' be *continuous* at $x = 0$, lead to the desired relation.]

9. A mechanical system consists of N identical masses attached in series by identical springs to fixed end supports (Figure 3.39). Show that a

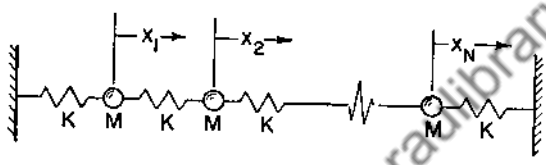


FIGURE 3.39

small displacement x_k of the kth mass satisfies the difference-differential equation

$$M\ddot{x}_k = K(x_{k+1} - 2x_k + x_{k-1}) \qquad (k = 1, 2, \cdots, N),$$

and the end conditions $x_0 = x_{N+1} = 0$, when no external forces are acting, where K is the spring constant of each spring. Show also that the assumption $x_k = A_k \cos(\omega t + \beta)$, where ω is the frequency of a natural mode of vibration and A_k is the amplitude of the oscillation of the kth mass, leads to the difference equation

$$A_{k+1} - 2A_k + A_{k-1} + \frac{M\omega^2}{K} A_k = 0 \qquad (k = 1, 2, \cdots, N),$$

with $A_0 = A_{N+1} = 0$.

10. The Bessel function $J_k(x)$ can be defined by the integral

$$J_k(x) = \frac{1}{\pi}\int_0^\pi \cos(x \sin\theta - k\theta)\, d\theta,$$

when k is zero or a positive integer. Determine A, B, and C in such a way that $A\,J_{k+1} + B\,J_k + C\,J_{k-1}$ is represented by an integral which can be evaluated by elementary methods (where A, B, and C may depend upon x and k), and hence show that $J_k(x)$ satisfies the difference equation (or "recurrence formula")

$$\frac{x}{2}J_{k+1}(x) - k\,J_k(x) + \frac{x}{2}J_{k-1}(x) = 0$$

when k is a positive integer. [With $u_k \equiv x \sin\theta - k\theta$, determine A, B, and C such that $A \cos u_{k+1} + B \cos u_k + C \cos u_{k-1}$ reduces to $\cos u_k \ (du_k/d\theta)$.]

11. The Tschebycheff polynomials $T_k(x)$ are defined by the expression

$$T_k(x) = \frac{1}{2^{k-1}} \cos (k \cos^{-1} x),$$

when k is a positive integer or zero and $|x| \leqq 1$. By considering the expressions for T_{k+1}, T_k, and T_{k-1}, obtain the recurrence formula

$$T_{k+1} - x\,T_k + \tfrac{1}{4}T_{k-1} = 0 \qquad (k = 1, 2, \cdots),$$

where $T_0 = 2$ and $T_1 = x$. Also, use this result to write out T_2, T_3, T_4 in explicit polynomial form.

12. Let a sequence of functions $r_k(x)$ be defined as follows: The zeroth function is defined to be $r_0(x) = x$, the first function to be $r_1(x) = a_1/(b_1 + x)$, and the kth function is obtained from the preceding function $r_{k-1}(x)$ by replacing x by $a_k/(b_k + x)$, where $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are constants.

(a) Show that there follows

$$r_0(x) = x, \quad r_1(x) = \frac{a_1}{b_1 + x}, \quad r_2(x) = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + x}}, \quad \cdots,$$

and, in general,

$$r_k(x) = \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdot \cdot \cdot + \cfrac{a_k}{b_k + x}}}}.$$

The expression $r_k \equiv r_k(0)$, *obtained by setting* $x = 0$ *in* $r_k(x)$, *is called a continued fraction of* k *stages.*

(b) Noticing that the result of clearing fractions in the expression for $r_k(x)$ is necessarily the ratio of two linear functions of x, of the form

$$r_k(x) = \frac{A_k + C_k x}{B_k + D_k x} \qquad (k = 0, 1, 2, \cdots),$$

deduce that there must follow

$$\frac{A_k + C_k x}{B_k + D_k x} \equiv \frac{(b_k A_{k-1} + a_k C_{k-1}) + A_{k-1} x}{(b_k B_{k-1} + a_k D_{k-1}) + B_{k-1} x} \qquad (k = 1, 2, \cdots),$$

for all values of x for which $r_k(x)$ is defined, and that also

$$A_1 = \kappa a_1, \quad C_1 = 0, \quad B_1 = \kappa b_1, \quad D_1 = \kappa,$$

where κ is an arbitrary nonzero constant of proportionality. [The right-hand member of the identity is the result of replacing k by $k - 1$ and x by $a_k/(b_k + x)$ in the left-hand member.]

(c) Show that the satisfaction of the identity of part (b) implies the relations

$$C_k = \mu_k A_{k-1}, \qquad D_k = \mu_k B_{k-1},$$

$$A_k = \mu_k(b_k A_{k-1} + a_k C_{k-1}), \qquad B_k = \mu_k(b_k B_{k-1} + a_k D_{k-1}),$$

for $k = 1, 2, \ldots$, where μ_k differs from zero, but is otherwise arbitrary.

(d) Deduce that there follows

$$r_k(x) = \frac{A_k + \mu_k A_{k-1} x}{B_k + \mu_k B_{k-1} x} \qquad (k = 1, 2, \cdots),$$

where A_k and B_k satisfy the difference equations

$$\left.\begin{aligned} A_k &= \mu_k b_k A_{k-1} + \mu_k \mu_{k-1} a_k A_{k-2}, \\ B_k &= \mu_k b_k B_{k-1} + \mu_k \mu_{k-1} a_k B_{k-2} \end{aligned}\right\} \quad (k = 2, 3, \cdots)$$

and the initial conditions

$$A_0 = 0, \quad A_1 = \kappa a_1, \quad \mu_1 B_0 = \kappa, \quad B_1 = \kappa b_1,$$

and where $\kappa, \mu_1, \mu_2, \ldots$ are arbitrary nonzero constants.

(e) By setting $x = 0$, and taking $\kappa = \mu_1 = \mu_2 \ldots = 1$ (for convenience), deduce that the continued fraction r_k of k stages can be expressed in the form

$$r_k \equiv \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cdot \\ \qquad \cdot \\ \qquad \cdot + \cfrac{a_k}{b_k}}} = \frac{A_k}{B_k} \qquad (k = 0, 1, 2, \cdots),$$

where the numerator and denominator of the cleared fraction satisfy the linear difference equations

$$A_k = b_k A_{k-1} + a_k A_{k-2}, \qquad B_k = b_k B_{k-1} + a_k B_{k-2} \qquad (k = 2, 3, \cdots)$$

and the initial conditions

$$A_0 = 0, \quad A_1 = a_1, \quad B_0 = 1, \quad B_1 = b_1.$$

(f) From the result of part (d), show that the value of $r_k \equiv r_k(0)$ is unchanged if each b_i is replaced by $\mu_i b_i$ and each a_i is at the same time replaced by $\mu_i \mu_{i-1} a_i$, where $\mu_0 = 1$, and where $\mu_1, \mu_2, \ldots$ are arbitrary

nonzero constants. [Show that this substitution (known as an *equivalence transformation*) reduces the difference equations and initial conditions of part (e) to those of part (d) with $\kappa = \mu_1$.]

Section 3.4.

13. Find the general solution of each of the following difference equations:

(a) $2y_{k+3} - 7y_{k+2} + 5y_{k+1} + 2y_k = 0.$

(b) $y_{k+3} - 5y_{k+2} + 8y_{k+1} - 4y_k = 0.$

(c) $y_{k+4} + y_k = 0.$

(d) $y_{k+2} + 2y_k + y_{k-2} = 0.$

14. Starting with the difference equation of Problem 11, derive the form given in that Problem for the Tschebycheff polynomial of degree k when $|x| \leqq 1$. If the polynominal is *defined* by the difference equation when $x \geqq 1$, obtain the alternate forms

$$T_k = \frac{1}{2^k}[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k]$$

$$= \frac{1}{2^{k-1}}[\cosh\ (k\ \cosh^{-1} x)]$$

in that case.

15. Suppose that the continuous beam of Problem 8 can be considered as being of infinite extent to the right (so that the number N of supports is infinite), and that the left-hand end of the beam overhangs the first support by a distance h, at which end a concentrated transverse force P is acting. Show that the moment at the kth support is then given by

$$M_k = (-1)^k(2 - \sqrt{3})^{k-1} Ph = (-1)^k Ph\, e^{-\alpha(k-1)} \quad (k = 1, 2, \cdots),$$

where $\alpha = \cosh^{-1} 2 \doteq 1.316$, if no other forces are present. [Here one must have $M_1 = -Ph$ and $\lim_{k \to \infty} M_k = 0$.]

16. Consider the linear difference equation

$$y(x + nh) + A_1 y(x + nh - h) + \cdots + A_{n-1} y(x + h) + A_n y(x) = 0,$$

where h and the A's are real constants.

(a) If x takes on only the values $x_0 + kh$, where x_0 is fixed and k is integral, show that the results of Section 3.4 lead to the solution

$$y(x) = c_1\beta_1^{x/h} + c_2\beta_2^{x/h} + \cdots + c_n\beta_n^{x/h},$$

where the c's are arbitrary constants, and where the β's are roots of the characteristic equation

$$\beta^n + A_1\beta^{n-1} + \cdots + A_{n-1}\beta + A_n = 0,$$

in the case when the n roots are real, distinct, and positive.

(b) Verify that this expression satisfies the given equation identically, regardless of whether the argument x is considered to take on only discrete values or to vary continuously.

(c) Verify that, when x varies continuously, the result of replacing each arbitrary c by any periodic function $\omega(x)$, *which is of period* h, also satisfies the difference equation, so that the expression

$$y(x) = \omega_1(x)\beta_1^{x/h} + \omega_2(x)\beta_2^{x/h} + \cdots + \omega_n(x)\beta_n^{x/h}$$

is a solution for any choice of the n functions $\omega_i(x)$ of period h. [It can be shown that this solution is the *most general* one.]

(d) In the case when β_1 is real and negative, say $\beta_1 = -\rho_1$, show that the real or imaginary part of the expression

$$\omega(x)(\rho_1 e^{i\pi})^{x/h} = \omega(x)\rho_1^{x/h}e^{\pi i x/h}$$

is a solution, so that the solution corresponding to the root $\beta_1 = -\rho_1$ can be taken in either of the real forms

$$y(x) = \omega_1(x)\rho_1^{x/h}\cos\frac{\pi x}{h} \qquad \text{or} \qquad y(x) = \omega_2(x)\rho_1^{x/h}\sin\frac{\pi x}{h}.$$

Notice also that the second form is identified with the first by writing $\omega_1(x) = \omega_2(x)\tan(\pi x/h)$, since the function $\tan(\pi x/h)$ is itself of period h. [Except in this case, the results of Section 3.4 are directly generalized to the continuous case by replacing y_{k+r} by $y(x + rh)$, k by x/h, and the arbitrary c_i by arbitrary functions $\omega_i(x)$ of period h.]

17. Apply the results of Problem 16 in obtaining the solutions listed in the following cases, with the convention that $\omega_i(x)$ is an arbitrary function of period h:

(a) $y(x + h) - 2y(x) + y(x - h) = 0$;

$$y(x) = \omega_1(x) + \frac{x}{h}\omega_2(x).$$

(b) $y(x + h) + 2y(x) + y(x - h) = 0$;

$$y(x) = \cos\frac{\pi x}{h}\left[\omega_1(x) + \frac{x}{h}\omega_2(x)\right].$$

(c) $y(x + h) - 2y(x) + 2y(x - h) = 0$;

$$y(x) = 2^{x/2h}\left[\omega_1(x)\cos\frac{\pi x}{4h} + \omega_2(x)\sin\frac{\pi x}{4h}\right].$$

In each case, also verify the correctness of the result by direct substitution.

Section 3.5.

18. Find the general solution of the equation

$$y_{k+1} - 2y_k + y_{k-1} = \phi_k$$

in each of the following cases:

(a) $\phi_k = a^k \quad (a \neq 1)$. (b) $\phi_k = e^{bk} \quad (b \neq 0)$.

(c) $\phi_k = \sin ck$. (d) $\phi_k = 1$.

(e) $\phi_k = k$. (f) $\phi_k = k\, e^{bk} \quad (b \neq 0)$.

19. (a) If L_{11}, L_{12}, L_{21}, and L_{22} are linear difference operators with constant coefficients, show that all solutions of the *simultaneous equations*

$$L_{11}u_k + L_{12}v_k = f_k, \qquad L_{21}u_k + L_{22}v_k = g_k$$

are also solutions of the uncoupled equations

$$L\,u_k = L_{22}f_k - L_{12}g_k, \qquad L\,v_k = L_{11}g_k - L_{21}f_k,$$

where L is the operator $L_{11}L_{22} - L_{12}L_{21}$. (Since the converse is not generally true, conditions on the arbitrary constants in the general solutions of the latter equations must be determined by substitution into the original equations.)

(b) Obtain the general solution of the equations

$$u_{k+1} + u_k - 2v_{k+1} + v_k = 0,$$

$$u_{k+1} - u_k - v_{k+1} + v_k = 2 \cdot 3^k$$

in the form

$$u_k = c_1 + c_2 2^k + 5 \cdot 3^k, \quad v_k = 2c_1 + c_2 2^k + 4 \cdot 3^k,$$

by the method of part (a) and by at least one other method.

20. Obtain the general solution of the equation

$$y_{k+1} - 2y_k + y_{k-1} = f_k \quad (k = 1, 2, \cdots)$$

in the form

$$y_k = \sum_{n=1}^{k} (k - n)f_n + c_1 + c_2 k.$$

Also, specialize when $f_k = \delta_{kr}$.

21. It is required to determine the coefficients f_k in the Maclaurin expansion

$$\frac{F(t)}{A - 2B\,t + C\,t^2} = \sum_{k=0}^{\infty} f_k\, t^k,$$

where A, B, and C are constants, under the assumption that the expansion of $F(t)$,

$$F(t) = \sum_{k=0}^{\infty} F_k t^k,$$

is known. By equating coefficients of like powers of t in the relation

$$\sum_{k=0}^{\infty} F_k t^k = (A - 2Bt + Ct^2) \sum_{k=0}^{\infty} f_k t^k,$$

show that f_k satisfies the difference equation

$$A f_k - 2B f_{k-1} + C f_{k-2} = F_k \qquad (k = 2, 3, \cdots)$$

and the initial conditions

$$A f_0 = F_0, \qquad A f_1 - 2B f_0 = F_1.$$

[If $\phi(t) = \sum_{0}^{\infty} f_k t^k$, the function $\phi(t)$ is called the *generating function* of f_k.]

22. (a) Use the result of Problem 21 to obtain the expansion

$$\frac{1}{1 - 2t\cos\theta + t^2} = \sum_{k=0}^{\infty} t^k \frac{\sin(k+1)\theta}{\sin\theta} \qquad (\theta \neq n\pi, \ |t| < 1).$$

(b) By setting t successively equal to $1/n$ and $-1/n$, where $n > 1$, deduce the relations,

$$\sum_{k=1}^{\infty} \frac{\sin k\theta}{n^k} = \frac{n \sin\theta}{n^2 - 2n\cos\theta + 1} \qquad (n > 1),$$

$$\sum_{k=1}^{\infty} (-1)^k \frac{\sin k\theta}{n^k} = \frac{-n \sin\theta}{n^2 + 2n\cos\theta + 1} \qquad (n > 1).$$

23. A function $\phi(t)$ is the generating function of f_k [that is, $\phi(t)$ is defined by the series $\phi(t) = \sum_{0}^{\infty} f_k t^k$] where f_k is known to satisfy the difference equation

$$f_{k+2} - 2f_{k+1} + f_k = 1 \qquad (k = 2, 3, \cdots)$$

and the conditions $f_0 = 1$, $f_1 = 0$. Use the result of Problem 21 to show that

$$\phi(t) = \frac{1 - 2t + t^2(1 + t + t^2 + \cdots)}{1 - 2t + t^2} = \frac{1 - 3t + 3t^2}{(1 - t)^3} \qquad (|t| < 1).$$

Section 3.6.

24. Two players take part in a game of coin tossing, the first starting with m coins and the second with n coins. It is agreed that play is finished if either player wins all. What is the probability that the first player will win? [Let p_k represent that probability when he has k coins and show that $p_k = (p_{k+1} + p_{k-1})/2$, where $p_0 = 0$ and $p_{m+n} = 1$. Then determine p_k and, finally, p_m.]

25. A flywheel of moment of inertia I_f is attached rigidly to a fixed support, by a shaft of length $(N + 1)h$ on which are mounted N identical disks, each of moment of inertia I (Figure 3.40). The portion of shaft

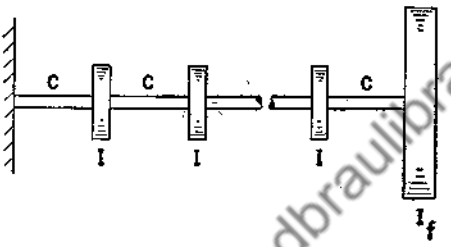


FIGURE 3.40

joining successive disks exerts a restraining torque numerically equal to an elastic constant c times the relative rotation of those disks.

(a) Show that the rotation θ_k of the kth disk is governed by the equation

$$I\frac{d^2\theta_k}{dt^2} = c(\theta_{k+1} - 2\theta_k + \theta_{k-1}) + T_k \quad (k = 1, 2, \cdots, N),$$

where T_k is the external torque applied to that disk, and that the end conditions are of the form

$$\theta_0 = 0, \qquad I_f\frac{d^2\theta_{N+1}}{dt^2} = -c(\theta_{N+1} - \theta_N).$$

(b) In the case of free torsional oscillations of the system [$T_k = 0$, $\theta_k = A_k \cos(\omega t + \beta)$], show that the amplitude A_k satisfies the equation

$$A_{k+1} - 2A_k + A_{k-1} + \frac{I\omega^2}{c}A_k = 0 \quad (k = 1, 2, \cdots, N)$$

and the end conditions $A_0 = 0$, $A_{N+1} = C\,A_N$, where $C = c/(c - I_f\omega^2)$ is the so-called *dynamic elastic constant* of the support.

(c) Show that the natural frequencies are of the form

$$\omega_n = 2\sqrt{\frac{c}{I}}\,\sin\frac{\alpha_n}{2},$$

where α_n is the nth solution of the transcendental equation

$$\sin (N + 1)\alpha = C \sin N\alpha,$$

in which

$$C = \frac{1}{1 - 4\dfrac{I_f}{I} \sin^2 \dfrac{\alpha}{2}}.$$

26. (a) Obtain the natural frequencies and modes of the mechanical system of Figure 3.39 (Problem 9), by using the results of equations (111a–d).

(b) If the problem of part (a) is modified in such a way that the end $k = N + 1$ is unrestrained, show that the natural frequencies are given by

$$\omega_n = 2\sqrt{\frac{K}{M}} \sin \left(\frac{2n - 1}{2N + 1} \frac{\pi}{2}\right) \qquad (n = 1, 2, \cdots, N).$$

[The condition at the free end is $x_{N+1} = x_N$.]

27. A schematic representation of a *string insulator* is given in Figure 3.41, one end being grounded and the other end being attached to a line

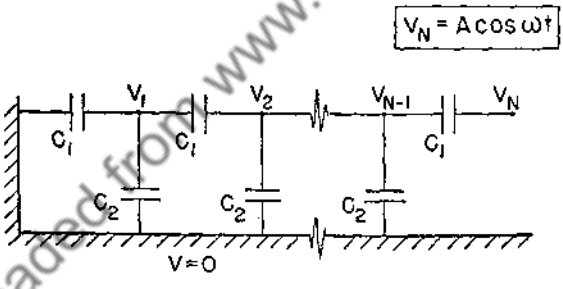


FIGURE 3.41

conductor which carries alternating current of frequency ω. The line carries $N - 1$ identical insulators, the capacity between successive conducting segments being denoted by C_1, and the capacity relative to the ground by C_2. Show that the voltage V_k of the kth conducting segment is given by

$$V_k = A \frac{\sinh \alpha k}{\sinh \alpha N} \cos \omega t,$$

where $\cosh \alpha = 1 + C_2/2C_1$. [Notice that, if I_k is the current flowing in the kth stage, there follows $I_k = -C_1(\dot{V}_k - \dot{V}_{k-1})$ and $I_{k+1} - I_k = -C_2\dot{V}_k$.]

28. A string of length $(N + 1)h$ carries N identical equally spaced masses which are connected with a fixed support by equal springs (Figure 3.42). If the string is stretched under a large uniform tension T, and the

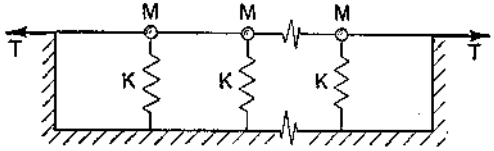


FIGURE 3.42

ends are fixed, show that the natural frequencies of small transverse oscillations are given by the expression

$$\omega_n = 2\sqrt{\frac{T}{Mh}}\sqrt{\sin^2\frac{n\pi}{2(N+1)} + \frac{hK}{4T}} \quad (n = 1, 2, \cdots, N),$$

where K is the spring constant and h the spacing.

29. The so-called *Fibonacci numbers* comprise the sequence 0, 1, 1, 2, 3, 5, 8, . . . , such that each number is the sum of the two preceding numbers.

(a) Show that the kth number n_k is given by

$$n_k = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k\right] \quad (k = 0, 1, 2, \cdots).$$

(b) Show that the ratio n_k/n_{k+1} of the kth number to the following number tends to the limit $2/(1 + \sqrt{5}) = (\sqrt{5} - 1)/2$ as $k \to \infty$. [This number is often known as the "golden mean" and is said, for example, to be the ratio of the sides of that rectangle of most pleasing proportions.]

30. Assume that rabbits reproduce at a rate such that one pair is born each month from each pair of adults not less than two months old. If one pair is present initially, and if none die, show that the total number in successive months is given by the Fibonacci sequence 1, 2, 3, 5, 8, . . . considered in Problem 29, that 377 pairs will be present after a year, and that the ratio of the number in a given month to that in the following month tends toward the "golden mean" with increasing time.

31. If the ratio n_k/n_{k+1} of successive Fibonacci numbers (Problem 29) is denoted by r_k, show that r_k satisfies the nonlinear difference equation

$$r_k(r_{k-1} + 1) = 1 \quad \text{or} \quad r_k = \frac{1}{1 + r_{k-1}} \qquad (k = 1, 2, \cdots),$$

with the initial condition $r_0 = 0$. Hence show that r_k can be expressed as the continued fraction

$$r_k = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdot \\ \qquad \cdot \\ \qquad\quad \cdot + \cfrac{1}{1}}}}}$$

which is terminated at the kth stage (after the kth division). Also, use the result of Problem 29(b) to deduce the expansion

$$\frac{\sqrt{5} - 1}{2} = \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdot \\ \qquad \cdot \\ \qquad\quad \cdot}}}}$$

of the "golden mean," where the divisions are continued indefinitely. [Notice that the ratio n_k/n_{k+1} is hence the kth "approximant" (or "convergent") of the continued-fraction expansion of the "golden mean."]

32. (a) If r_k satisfies the nonlinear difference equation

$$r_k(1 + a\, r_{k-1}) = 1,$$

with $r_0 = 0$, show that r_k can be expressed as the continued fraction

$$r_k = \cfrac{1}{1 + \cfrac{a}{1 + \cfrac{a}{1 + \cfrac{a}{1 + \cdot \\ \qquad \cdot \\ \qquad\quad \cdot + \cfrac{a}{1}}}}}$$

which terminates at the kth stage.

(b) By making the substitution $r_k = n_k/n_{k+1}$, reduce the difference equation of part (a) to the form $n_{k+1} - n_k - a\, n_{k-1} = 0$, with $n_0 = 0$. Obtain an expression for n_k, assuming that a is real, and considering separately the cases $a > -\frac{1}{4}$, $a = -\frac{1}{4}$, and $a = -\frac{1}{4} - c$, where $c > 0$. (Notice that the solution involves an arbitrary multiplicative constant.) Hence obtain an explicit expression for r_k in each case.

(c) By considering the behavior of r_k as $k \to \infty$, deduce that the continued fraction of part (a) converges to the limit

$$\frac{2}{\sqrt{1+4a}+1}$$

when $a \geqq -\frac{1}{4}$ and fails to converge when $a < -\frac{1}{4}$, as the number of divisions is continued indefinitely.

33. Deal as in Problem 32 with the difference equation

$$r_k(b + r_{k-1}) = 1,$$

with $r_0 = 0$ and $b \neq 0$, showing that the infinite continued fraction

$$\cfrac{1}{b + \cfrac{1}{b + \cfrac{1}{b + \cfrac{1}{b + \ddots}}}}$$

converges to $\frac{1}{2}(\sqrt{b^2+4} - b)$ when $b > 0$, and to $-\frac{1}{2}(\sqrt{b^2+4} + b)$ when $b < 0$. In particular, deduce the expansion

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots}}} = \sqrt{2}.$$

34. Let $w_k(t)$ represent the value of w at position $x = x_0 + kh$ along the x-axis at time t. Verify that the real and imaginary parts of the expression

$$w_k(t) = A\, e^{i\omega(t-ak)}$$

represent "traveling waves" which move in the positive x-direction with velocity h/α, without damping, and whose amplitude oscillates in time with circular frequency ω.

35. With the terminology of Problem 34, show that, if $w_k(t)$ satisfies an equation of the form

$$\delta^2 w_k \equiv w_{k+1} - 2w_k + w_{k-1} = \mu \frac{d^2 w_k}{dt^2},$$

where μ is a positive constant, then traveling waves may be propagated without attenuation (damping) only if their frequency is such that

$$\omega < \frac{2}{\sqrt{\mu}}.$$

[A system (mechanical, electrical, acoustical, or otherwise) which is governed by such an equation is an example of what is known as a "low-pass filter." The value $\omega_c = 2/\sqrt{\mu}$ is called the *cutoff frequency;* waves of higher frequency are damped out as they progress along the x-axis.]

36. If the governing equation of Problem 35 is modified by the addition of a term $\gamma\, w_k$ on the right,

$$\delta^2 w_k = \mu \frac{d^2 w_k}{dt^2} + \gamma\, w_k,$$

where γ is a positive constant, show that traveling waves may be propagated without attenuation only if $\left|1 + \frac{\gamma}{2} - \frac{\mu\,\omega^2}{2}\right| < 1$ or

$$\sqrt{\frac{\gamma}{\mu}} < \omega < \sqrt{\frac{\gamma + 4}{\mu}}.$$

[Such a system is an example of a "band-pass filter." Only frequencies between $\omega_1 = \sqrt{\gamma/\mu}$ and $\omega_2 = \sqrt{(\gamma + 4)/\mu}$ are not damped out as k increases.]

37. If the governing equation of Problem 35 is replaced by the equation

$$\delta^2\left(\frac{d^2 w_k}{dt^2}\right) = \kappa\, w_k,$$

where κ is a positive constant, show that the condition for absence of attenuation becomes

$$\omega > \frac{\sqrt{\kappa}}{2}.$$

[Such a system exemplifies a "high-pass filter."]

38. Show that the mechanical system of Figure 3.42 (Problem 28) has the properties of a band-pass filter, with cutoff frequencies $\omega_1 = \sqrt{K/M}$ and $\omega_2 = \sqrt{(Kh + 4T)/Mh}$. [This means, for example, that the effect of forced vibrations of one support, with frequencies outside this band, will be damped out with distance from that support.]

39. Show that the networks of Figures 3.43(a), (b), and (c) represent respectively low-pass, high-pass, and band-pass filters. [If $I_k = \dot{Q}_k$ is the current in the kth loop, in each case, show that there follows

$$-\frac{1}{C}\,\delta^2 Q_k + L\,\ddot{Q}_k = 0 \text{ in (a)},$$

$$-L\,\delta^2 \ddot{Q}_k + \frac{1}{C}\,Q_k = 0 \text{ in (b)},$$

and

$$-\frac{1}{C_2}\,\delta^2 Q_k + L\,\ddot{Q}_k + \frac{1}{C_1}\,Q_k = 0 \text{ in (c).}]$$

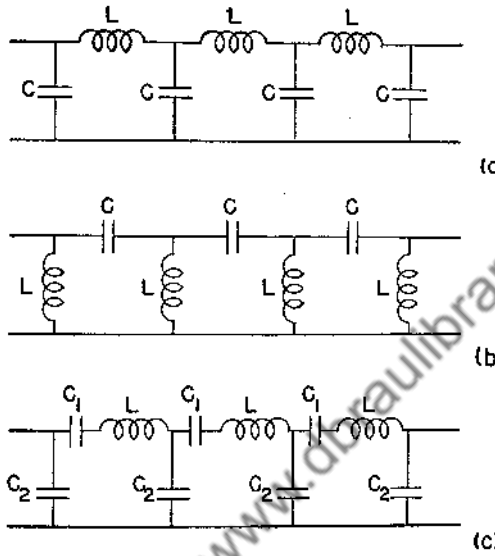


FIGURE 3.43

Sections 3.7, 3.8.

40. Show that $\sum_{k=1}^{K} \sin k\alpha = 0$ if $\alpha = 2n\pi/K$ or $2n\pi/(K+1)$, when n is an integer. Also, verify this result directly in the case when $K = 3$ and $n = 1$.

41. Show that

$$\sum_{k=1}^{K} \sin^2 \frac{n\pi k}{K+1} = \frac{K+1}{2} \qquad (n \neq 0)$$

when n is an integer. By replacing $\sin^2 u$ by $1 - \cos^2 u$, deduce also that

$$\sum_{k=0}^{K} \cos^2 \frac{n\pi k}{K+1} = \begin{cases} \dfrac{K+1}{2}, & n \neq 0, \\ K+1, & n = 0, \end{cases}$$

when n is an integer.

42. Show that

$$\sum_{k=0}^{K} \cos k\alpha = \frac{\cos K\alpha \sin \frac{K+1}{2}\alpha}{\sin \frac{1}{2}\alpha} \quad (\alpha \neq 2r\pi),$$

and deduce that

$$\sum_{k=0}^{K} \cos \frac{(2n+1)\pi k}{2K} = 0 \quad \text{and} \quad \sum_{k=0}^{K} \cos \frac{2n\pi k}{K+1} = 0$$

when n is an integer, and when $n/(K+1)$ is nonintegral in the second case.

43. Let $C_n^k \equiv \binom{k}{n}$ denote the binomial coefficient

$$\frac{k(k-1)\cdots(k-n+1)}{n(n-1)\cdots 1}.$$

(a) Show that $C_n^k = \dfrac{k^{(n)}}{n!}$.

(b) If n is fixed, show that $\triangle C_n^k = \dfrac{k^{(n-1)}}{(n-1)!} = C_{n-1}^k$, and deduce that

$$C_n^{k+1} = C_n^k + C_{n-1}^k.$$

44. Show that the formula derived in Problem 5 can be written in the form

$$y_{k+1} = \sum_{n=0}^{k} C_n^k \triangle^n y_1 = \sum_{n=0}^{k} \frac{\triangle^n y_1}{n!} k^{(n)}.$$

[Notice the analogy with the Maclaurin power-series expansion.]

45. Express each of the following sums in closed form:

(a) $1 \cdot 2 \cdot 3 + 2 \cdot 3 \cdot 4 + \cdots + n(n+1)(n+2)$.

(b) $\dfrac{1}{1 \cdot 2 \cdot 3} + \dfrac{1}{2 \cdot 3 \cdot 4} + \cdots + \dfrac{1}{n(n+1)(n+2)}$.

(c) $\dfrac{1}{1 \cdot 3} + \dfrac{1}{3 \cdot 5} + \cdots + \dfrac{1}{(2n+1)(2n+3)}$.

(d) $1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + \cdots + n(n+2)$.

46. The functions $\Psi(x)$, $\Psi'(x)$, $\Psi''(x)$, and so forth, where primes denote differentiation with respect to x, are tabulated functions. [The Psi function, defined by equation (154), is often also called the *digamma func-*

tion, and its successive derivatives the trigamma function, the tetragamma function, and so forth.]

(a) Show that $\Psi(x) = \Psi(x-1) + 1/x$.

(b) Show that the rth derivative of the Psi function has the following property:

$$\sum_{k=1}^{n} \frac{1}{(k+x)^{r+1}} = \frac{(-1)^r}{r!}\left[\Psi^{(r)}(n+x) - \Psi^{(r)}(x)\right] \qquad (r = 0, 1, \cdots).$$

(c) Show that

$$\Psi(n+x) - \Psi(n) - \Psi(x) = \gamma + \sum_{k=1}^{n}\left(\frac{1}{k+x} - \frac{1}{k}\right),$$

when n is a positive integer.

47. Use the result of Problem 46(b) to obtain the more general summation formula

$$\sum_{k=1}^{n} \frac{1}{(ak+b)^{r+1}} = \frac{(-1)^r}{a^{r+1}\,r!}\left[\Psi^{(r)}\left(n+\frac{b}{a}\right) - \Psi^{(r)}\left(\frac{b}{a}\right)\right].$$

48. It can be shown that $\Psi(n+x) - \Psi(n)$ tends to zero for fixed x as $n \to \infty$, and that all derivatives of $\Psi(n+x)$ also tend to zero as $n \to \infty$. Assuming these facts, obtain the following relations from the results of Problem 46:

(a) $$\Psi(x) = -\gamma + \sum_{k=1}^{\infty}\left(\frac{1}{k} - \frac{1}{k+x}\right).$$

(b) $$\Psi^{(r)}(x) = (-1)^{r+1}\,r! \sum_{k=1}^{\infty} \frac{1}{(k+x)^{r+1}} \qquad (r = 1, 2, \cdots).$$

49. Use the result of Problem 48(a) to show that

$$\sum_{k=1}^{\infty}\left[\frac{A_1}{k+\alpha_1} + \cdots + \frac{A_n}{k+\alpha_n}\right]$$

$$= \sum_{k=1}^{\infty}\left[A_1\left(\frac{1}{k+\alpha_1} - \frac{1}{k}\right) + \cdots + A_n\left(\frac{1}{k+\alpha_n} - \frac{1}{k}\right)\right]$$

$$= -[A_1\Psi(\alpha_1) + \cdots + A_n\Psi(\alpha_n)],$$

if $A_1 + \cdots + A_n = 0$. Show also that this last condition is necessary in order that the given series converge.

50. By expanding each summand in partial fractions, and using the results of Problems 48 and 49, obtain the following results:

$$\text{(a)} \quad \sum_{k=1}^{\infty} \frac{1}{(k+a)(k+b)} = \frac{1}{a-b}\,[\Psi(a) - \Psi(b)] \qquad (a \neq b).$$

$$\text{(b)} \quad \sum_{k=1}^{\infty} \frac{1}{(k+a)^2(k+b)} = \frac{1}{(a-b)^2}\,[\Psi(a) - \Psi(b)] - \frac{1}{a-b}\,\Psi'(a) \quad (a \neq b).$$

Also, verify that the result of part (a) agrees with the limiting form of equation (153b) when $a = 3$ and $b = 1$. [Use equation (157).]

Section 3.9.

51. (a) Show that the characteristic functions of the problem

$$y_{k+1} - 2y_k + y_{k-1} + \lambda\, y_k = 0 \quad (k = 1, 2, \cdots, N),$$

$$y_0 = 0, \qquad y_{N+1} - \mu\, y_N = 0,$$

where μ is a constant such that $0 \leqq \mu \leqq 1$, are of the form

$$\phi_{n,k} = \sin \alpha_n k$$

where α_n is the nth solution of the transcendental equation

$$\sin (N+1)\alpha = \mu \sin N\alpha,$$

and that the corresponding characteristic values of λ are given by

$$\lambda_n = 4 \sin^2 \frac{\alpha_n}{2}.$$

(b) In the special case $\mu = 1$, in which the second end condition becomes $\Delta y_N = 0$, show that there follows

$$\phi_{n,k} = \sin\left(\frac{2n-1}{2N+1}\,\pi k\right) \quad \text{and} \quad \lambda_n = 4 \sin^2\left(\frac{2n-1}{2N+1}\,\frac{\pi}{2}\right),$$

where $n = 1, 2, \ldots, N$. [Compare Problem 26(b).]

(c) If $\mu > (N+1)/N$, show that one of the characteristic numbers, say λ_1, is negative and is then given by

$$\lambda_1 = -4 \sinh^2 \tfrac{1}{2}\gamma \quad \text{where} \quad \sinh (N+1)\gamma = \mu \sinh N\gamma,$$

corresponding to the characteristic function $\phi_{1,k} = \sinh \gamma k$. [Notice that here $\alpha_1 = i\gamma$ in the notation of part (a).]

52. Suppose that the end $x = (N+1)h$ of the vibrating loaded string considered in Section 3.6 is completely restrained from motion in the

x-direction, and is partially restrained from transverse motion by a spring which exerts a restoring force numerically equal to $K\,y_{N+1}$, where K is the spring constant. If the end $x = 0$ is fixed at the origin, show that the linearized formulation is that of Problem 51 if we take $\lambda = Mh\omega^2/T$ and $\mu = T/(T + Kh)$, where T is the tension (assumed to be constant). [Notice that $\mu = 0$ corresponds to fixity, whereas $\mu = 1$ corresponds to absence of transverse restraint.]

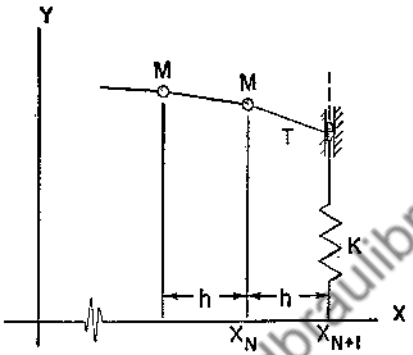


FIGURE 3.44

53. It can be shown (see Problem 55) that exactly N of the roots α_n defined in Problem 51 lead to distinct characteristic numbers. If f_k is defined for $k = 1, 2, \ldots, N$, show that the coefficients in the expansion

$$f_k = \sum_{n=1}^{N} A_n \sin \alpha_n k \qquad (k = 1, 2, \cdots, N)$$

are then determined by the equations

$$A_n \left[\frac{N}{2} - \frac{\sin 2(N+1)\alpha_n}{4\mu \sin \alpha_n} \right] = \sum_{k=1}^{N} f_k \sin \alpha_n k \quad (n = 1, 2, \cdots, N).$$

[Make use of equations (167) and (137a).]

Section 3.10.

54. Show that the characteristic numbers of the Nth-order matrix

$$\mathbf{a} = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

are also characteristic values of λ for the problem

$$y_{k+1} - 2y_k + y_{k-1} + \lambda y_k = 0 \quad (k = 1, 2, \cdots, N),$$

$$y_0 = 0, \qquad y_{N+1} = 0,$$

and conversely, and hence are of the form

$$\lambda_n = 4 \sin^2 \frac{n\pi}{2(N+1)} \qquad (n = 1, 2, \cdots, N).$$

55. Consider the problem

$$y_{k+1} - 2a\, y_k + b^2 y_{k-1} + \lambda y_k = 0 \quad (k = 1, 2, \cdots, N),$$

$$y_0 = \mu_1 y_1, \qquad y_{N+1} = \mu_2 y_N,$$

where a and b are real constants, and where $b > 0$.

(a) Show that the difference equation is identified with (159) or (168) by setting

$$p_k = b^{-2k}, \quad q_k = b^{-2k}(1 - 2a + b^2), \quad r_k = b^{-2k},$$

and hence, in particular, deduce that the matrix $\mathbf{r}$ of Section 3.10 is positive definite, and that the N characteristic values of λ are *real*.

(b) By writing

$$\lambda = 2(a - b \cos \alpha),$$

obtain the general solution of the difference equation in the form

$$y_k = b^k(c_1 \sin \alpha k + c_2 \cos \alpha k),$$

and show that permissible values of α are then determined by the equation

$$\sin (N+1)\alpha - (\bar{\mu}_1 + \bar{\mu}_2) \sin N\alpha + \bar{\mu}_1\bar{\mu}_2 \sin (N-1)\alpha = 0,$$

where $\bar{\mu}_1 = b\,\mu_1$ and $\bar{\mu}_2 = b^{-1}\mu_2$.

(c) Show that the reality of the characteristic values of λ assures the reality of $\cos \alpha$, so that either α is real *or* $\cos \alpha = \cosh \gamma$ where $\alpha = i\gamma$ *or* $\cos \alpha = -\cosh \delta$ where $\alpha = \pi + i\,\delta$, where γ and δ are real and positive.

(d) Under the assumption that a value $\alpha = i\gamma$ is permissible, where $\gamma > 0$, show that γ must satisfy the equation

$$\sinh (N+1)\gamma - (\bar{\mu}_1 + \bar{\mu}_2) \sinh N\gamma + \bar{\mu}_1\bar{\mu}_2 \sinh (N-1)\gamma = 0.$$

Denoting the left-hand member of this equation by F, verify that $\partial F/\partial \bar{\mu}_1$ and $\partial F/\partial \bar{\mu}_2$ are negative for all positive values of γ when $0 \leqq \bar{\mu}_1 \leqq 1$ and $0 \leqq \bar{\mu}_2 \leqq 1$, and that F is positive for all positive values of γ when $\bar{\mu}_1 = \bar{\mu}_2 = 1$. Hence deduce that no value of α of the form $\alpha = i\gamma$, where $\gamma > 0$, can be permissible if $0 \leqq \mu_1 \leqq b^{-1}$ and $0 \leqq \mu_2 \leqq b$.

(e) Show that the assumption $\alpha = \pi + i\,\delta$, where $\delta > 0$, leads to the requirement

$$\sinh (N+1)\delta + (\bar{\mu}_1 + \bar{\mu}_2) \sinh N\delta + \bar{\mu}_1\bar{\mu}_2 \sinh (N-1)\delta = 0,$$

and hence cannot be valid if μ_1 and μ_2 are non-negative.

(f) If $\mu_1 = 0$, show that all permissible values of α are real if and only if $|\bar{\mu}_2| \leqq (N+1)/N$ or $|\mu_2| \leqq b(N+1)/N$, whereas the corresponding condition when $\mu_2 = 0$ is $|\mu_1| \leqq b^{-1}(N+1)/N$.

[Show that $\{\sinh (N+1)\gamma\}/\{\sinh N\gamma\}$ takes on all positive values greater than $(N+1)/N$, and only those values.]

Section 3.11.

56. Solve the modification of the problem considered in Section 3.11 in which the end $x = (N+1)h$ is not restrained from transverse motion. [See Problem 52.]

57. A conducting rod of cross-sectional area A, thermal conductivity K, and length $L = (N+1)h$, connects N mass points of equal mass M and specific heat s. The masses are at a constant separation h, and the rod extends a distance h beyond each of the extreme masses (Figure 3.45).

FIGURE 3.45

(a) If Q_k represents the rate of flow of heat into the kth mass, and T_k is the temperature of the kth mass, show that Q_k satisfies the two conditions

$$Q_k = s\,M\,\frac{dT_k}{dt}, \qquad Q_k = \frac{K\,A}{h}\,[(T_{k+1} - T_k) - (T_k - T_{k-1})].$$

Hence deduce that T_k satisfies the difference equation

$$\frac{dT_k}{dt} = \frac{K\,A}{h\,s\,M}\,(T_{k+1} - 2T_k + T_{k-1}) \quad (k = 1, 2, \cdots, N),$$

together with appropriate end conditions and initial conditions.

(b) Show also that as the distribution of masses tends to become continuous, in such a way that $h \to 0$, $(N+1)h$ is constantly equal to L, and $M/Ah \to \rho$, where ρ is a limiting linear mass density, the governing equation tends toward the heat-flow equation

$$\frac{\partial T}{\partial t} = \alpha^2 \frac{\partial^2 T}{\partial x^2},$$

where $\alpha^2 = K/\rho s$ is the thermal diffusivity. [Notice that, in the discrete case, the temperature will vary linearly between successive mass points.]

58. Suppose that, at the time $t = 0$, the temperatures of the mass points of Problem 57 are prescribed in such a way that

$$T_k(0) = \phi_k \qquad (k = 1, 2, \cdots, N),$$

and that, at all following times ($t > 0$), the ends of the conducting rod are maintained at temperature zero, so that

$$T_0(t) = 0, \quad T_{N+1}(t) = 0 \quad (t > 0).$$

Determine the temperature at each mass point at time t by the following procedure:

(a) Assume a particular "product solution" of the form

$$T_k(t) = f_k\, U(t),$$

where f_k is independent of time and $U(t)$ is independent of position, and show that, with the notation $\gamma^2 = (K\,A)/(h\,s\,M)$, there must follow

$$\frac{f_{k+1} - 2f_k + f_{k-1}}{f_k} = \frac{1}{\gamma^2 U}\frac{dU}{dt} = -\mu^2,$$

where μ^2 is an arbitrary constant, and where f_k satisfies the end conditions

$$f_0 = f_{N+1} = 0.$$

(b) Obtain permissible product solutions in the form

$$T_{k,n} = C_n \sin\frac{n\pi k}{N+1}\, e^{-\gamma^2\mu_n^2 t} \quad (n = 1, 2, \cdots, N),$$

where

$$\mu_n = 2\sin\frac{n\pi}{2(N+1)}.$$

(c) By superimposing such solutions, and satisfying the initial condition $T_k(0) = \phi_k$, obtain the desired solution in the form

$$T_k(t) = \sum_{n=1}^{N} C_n \sin\frac{n\pi k}{N+1}\, e^{-\gamma^2\mu_n^2 t} \quad (k = 0, 1, \cdots, N+1),$$

where

$$C_n = \frac{2}{N+1}\sum_{k=1}^{N} \phi_k \sin\frac{n\pi k}{N+1}.$$

59. Specialize the solution of Problem 58(c) in the following cases:

(a) Take $\phi_k = \sin\dfrac{r\pi k}{N+1}$, where r is an integer.

(b) Take $\phi_k = 0$ when $k \neq r$, and $\phi_r \neq 0$.

(c) Take $\phi_k = 1$

60. Suppose that the spacing h tends to zero (and $N \to \infty$) in Problem 58(c), in such a way that $(N+1)h$ remains equal to L, and that M/Ah tends to a constant ρ.

(a) Show that $\gamma^2\mu_n^2 = 4\dfrac{KA}{hs\bar{M}}\sin^2\dfrac{n\pi h}{2L}$ tends to $\dfrac{n^2\pi^2}{L^2}\dfrac{K}{s\rho} \equiv \dfrac{n^2\pi^2}{L^2}\alpha^2$, where α^2 is the thermal diffusivity.

(b) Show formally that the solution takes the form

$$T(x,t) = \sum_{n=1}^{\infty} C_n \sin\frac{n\pi x}{L}\, e^{-n^2\pi^2\alpha^2 t/L^2}$$

where

$$C_n = \frac{2}{L}\lim_{N\to\infty}\sum_{k=1}^{N}\phi(x_k)\sin\frac{n\pi x_k}{L}\Delta x = \frac{2}{L}\int_0^L \phi(x)\sin\frac{n\pi x}{L}\,dx.$$

Section 3.12.

61. Obtain the solution of the equation $y_{k+1} - e^{rk}y_k = 0$, in the form

$$y_k = c\, e^{(k^2-k)r/2}.$$

62. If y_k satisfies the equation $y_{k+1} - a_k y_k = 0$, where $a_k > 0$, show that $y_k = e^{u_k}$, where u_k is the general solution of the equation $u_{k+1} - u_k = a_k$. Apply this procedure to the solution of Problem 61.

63. Show that the equation

$$y_{k+1} - a_k y_k = b_k$$

is equivalent to the equation

$$\Delta(p_k y_k) = p_{k+1}b_k,$$

if p_k is defined by the relation

$$p_k = \frac{1}{\Pi^k a_{n-1}}.$$

Hence, with the notation

$$q_k = \Pi^k a_{n-1},$$

deduce that the general solution of the given equation is of the form

$$y_k = q_k\left(\Sigma^k\frac{b_{n-1}}{q_n} + C\right).$$

64. By making the use of the fact that $y_k = 1$ satisfies the equation

$$(k+2)y_{k+2} - (k+3)y_{k+1} + y_k = 0,$$

obtain the general solution in the form

$$y_k = C_1 \sum_{n=0}^{k} \frac{1}{n!} + C_2 \qquad (k \geq 0).$$

65. Verify that the substitution $y_k = (u_k/u_{k+1}) - B$ reduces the *nonlinear* equation

$$y_k y_{k-1} + A\, y_k + B\, y_{k-1} = C$$

to the *linear* equation

$$(AB + C)u_{k+1} - (A - B)u_k - u_{k-1} = 0.$$

[Special cases in which $B = 0$ occur in Problems 32 and 33.]

Section 3.13.

66. (a) If $y(x)$ satisfies the differential equation $y'' + x\,y = 0$, and if $y(0) = 0$, obtain the approximating difference equation

$$y_{k+1} = (2 - h^3 k)y_k - y_{k-1} \quad \text{where} \quad y_0 = 0,$$

by writing $x_k = kh$, and $y_k = y(x_k)$, and replacing y_k'' by $(\delta^2 y_k)/h^2$.

(b) Taking $h = \frac{1}{5}$, express $y_2, \ldots, y_5$ as numerical multiples of y_1, retaining slide-rule accuracy.

(c) From these results, obtain approximate values of the solution of the initial-value problem for which $y(0) = 0$ and $y'(0) = 1$.

(d) From the results of part (b), obtain approximate values of the solution of the boundary-value problem for which $y(0) = 0$ and $y(1) = 1$.

67. Repeat the calculations of Problem 66, with a halved spacing $h = \frac{1}{10}$, and compare the results of the two calculations.

68. Obtain approximations to the smallest characteristic values of λ for the problem $y'' + \lambda\, x\, y = 0$, where $y(0) = y(1) = 0$, taking sucessively $N = 1$ and $N = 2$ interior division points. [Write the approximating difference equation in the form $y_{k+1} - (2 - \bar{\lambda}\, k)y_k + y_{k-1} = 0$, where $\bar{\lambda} = h^3\lambda$, and where $y_0 = y_{N+1} = 0$.]

69. Obtain approximate values of the solution of the problem

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2y = 0, \quad y(0) = y'(0) = 1,$$

at the points $x = 0.1$ and 0.2, by replacing d^2y/dx^2 by $(\delta^2 y_k)/h^2$ and replacing dy/dx successively by $(y_k - y_{k-1})/h$, $(y_{k+1} - y_k)/h$, and $(y_{k+1} - y_{k-1})/2h$, with $h = 0.1$. In each case use the initial conditions $y_0 = 1$ and $y_1 - y_0 = h$. Compare the three values at $x = 0.2$ with the true value $y(0.2) = e^{0.2} \doteq 1.2214$.

Section 3.14.

70. A uniform rod of length $L = 1$ ft and diffusivity $\alpha^2 = 0.02$ sq ft per hr is initially at a uniform temperature 200°. The end $x = 1$ is then maintained at 200°, whereas the temperature of the end $x = 0$ is then reduced at a constant rate in such a way that it becomes 100° after five hours, after which that end temperature is maintained. Using four interior division points along the rod, determine approximately the temperature variation over the first ten hours.

71. Using a halved spacing along the rod in Problem 70, determine the approximate temperature variation over the first hour, and compare the distribution after one hour with the corresponding result of Problem 70. [Notice that dividing h_x by two corresponds to dividing h_t by four].

72. Let Problem 70 be modified in such a way that at the end $x = 0$ heat escapes at a rate proportional to the difference between the temperature T_0 of that end and the temperature T_a of the surrounding medium.

(a) If the constant of proportionality is denoted by μ, show that the condition at that end may be approximated by the requirement

$$T_0\left(1 + \frac{\mu\, h_x}{k\, A}\right) = T_1 + \frac{\mu\, h_x}{k\, A}\, T_a,$$

where k is the thermal conductivity and A is the cross-sectional area of the rod.

(b) Suppose, for simplicity in computation, that $kA/\mu = \frac{1}{3}$ ft and $T_a = 100°$. Again using four interior division points, determine the approximate temperature variation over the first ten hours.

Section 3.15.

73. A uniform plate, in the form of an isosceles right triangle whose legs are of length 1 ft, is initially at a uniform temperature of 100°. At the instant $t = 0$, the temperature along one leg of the triangular boundary is abruptly reduced to temperature 0° and so maintained, while the remainder of the boundary is maintained at 100°. Take $h_x = h_y = \frac{1}{4}$ ft, and suppose, for convenience, that $\alpha^2 = \frac{1}{64}$ sq ft per hr so that $h_t = 1$ hr. Determine approximate temperatures at the three interior net points at the end of each of the first five following hours.

74. Obtain a corresponding approximate solution to the modification of Problem 73 in which one leg of the boundary is abruptly reduced to 0° and so maintained, the other leg is maintained at 100°, and the remainder of the boundary is insulated when $t > 0$.

Section 3.16.

75. The temperature along the boundary of a square plate $ABCD$ is maintained in such a way that it varies linearly from 0° to 200° along

AB, is constantly 200° along BC, varies linearly from 200° to 100° along CD, and from 100° to 0° along DA. Obtain the approximate steady-state temperature distribution by replacing the plate by a network of 25 interconnected point masses, estimating temperatures at the nine interior net points, and proceeding by the iterative method of Section 3.16.

76. Obtain a corresponding approximate solution to the modification of Problem 75 in which the edge BC is insulated.

Section 3.17.

77. Apply the relaxation method to the treatment of Problem 75.

78. Apply the relaxation method to Problem 76.

79. A blindfolded prisoner is placed in a square maze consisting of N equally spaced interior passageways extending in the x-direction, crossed at right angles by N similarly spaced interior passageways in the y-direction (Figure 3.46). Along three of the boundaries of the maze a deep moat is present, whereas the fourth boundary $y = 0$ represents access to freedom. Let the passages in the y-direction be denoted by $i = 1, 2, \ldots, N$, and those in the x-direction by $j = 1, 2, \ldots, N$.

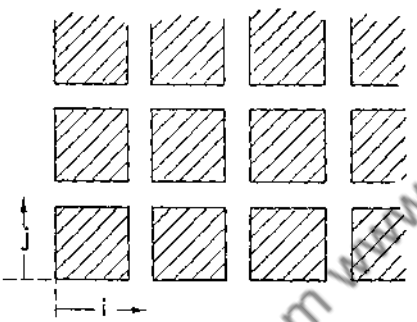


FIGURE 3.46

(a) If the probability of eventual escape when the prisoner is at the junction of the ith and jth corridors is denoted by p_{ij}, show that there must follow

$$p_{ij} = \tfrac{1}{4}(p_{i+1,j} + p_{i-1,j} + p_{i,j+1} + p_{i,j-1}),$$

where i and j vary from 1 to N, and that the boundary conditions

$$p_{0j} = 0, \quad p_{N+1,j} = 0, \quad p_{i,N+1} = 0, \quad p_{i0} = 1$$

must be satisfied. [Notice that the difference equation is completely analogous to equation (235).]

(b) Determine the probability of eventual escape at each junction of a maze for which $N = 3$, by relaxation methods or otherwise, obtaining each probability correct to two decimal places. [For convenience in calculation, multiply all probabilities by 1000.]

Section 3.18.

80. The temperature along the boundary of the plate $ABCDE$ of Figure 3.47, where CD is a quadrant of a circle and $\overline{AB} = \overline{EA} = 2\,\overline{ED}$,

is maintained in such a way that it varies linearly from 100° to 200° along AB, is constantly 200° along BC, varies linearly along the arc CD from 200° to 100°, and is constantly 100° along DE and EA. Determine approximate steady-state temperatures at interior points of a square net with spacing $\overline{AB}/4$.

81. Modify Problem 80 in such a way that the edges AB and EA are insulated (or are lines of symmetry).

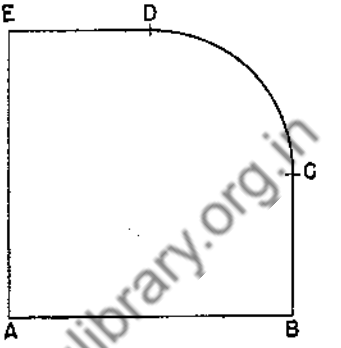


FIGURE 3.47

Section 3.19.

82. A function $\phi(x, y)$ satisfies Poisson's equation

$$\nabla^2\phi + \beta = 0,$$

where β is a positive constant, over a square of length a, and vanishes along the boundary of this square.

(a) By replacing the square by a network with spacing $h_x = h_y = a/(N + 1)$, and writing

$$\phi_k = \frac{\beta a^2}{100(N+1)^2} u_k,$$

show that u_k is then dimensionless, and that the residual relevant to the relaxation procedure takes the form

$$R_0 = u_1 + u_2 + u_3 + u_4 - 4u_0 + 100,$$

with the notation of Figure 3.17.

(b) Obtain an approximate solution to the problem of part (a) with $N = 3$.

(c) Use the results of part (b) to obtain an approximation to the integral of ϕ over the square.

83. A thin square plate of uniform thickness is bounded by the edges $x = 0$, $x = a$, $y = 0$, and $y = a$. In the absence of external loading, a small deflection $w(x, y)$ satisfies the equation $\nabla^4 w = 0$ (see Section 2.16). Suppose that the plate is clamped and undeflected along the three edges $x = 0$, $x = a$, and $y = a$, so that the conditions $w = 0$ and $\partial w/\partial n = 0$ are satisfied along those edges, and that the edge $y = 0$ is clamped in a deflected parabolic form, in such a way that $w = 4w_{\max}\dfrac{x}{a}\left(1 - \dfrac{x}{a}\right)$ and $\partial w/\partial y = 0$ along that edge. By replacing the plate by a network with spacing $h_x = h_y = a/4$, obtain approximate deflections at the nine interior net points. [Write $w/w_{\max} = u/100$.]

84. The interior of a long cylindrical furnace of inner radius a and outer radius b is maintained at constant temperature 200° [Figure 3.48(a)]. Half of the outer boundary ($\bar{C}BC$) is insulated, whereas heat escapes from the remainder ($CD\bar{C}$) at a rate proportional to the difference between the boundary temperature T_b, at a point, and the temperature T_a of the surrounding air, so that a condition of the form

$$\frac{\partial T}{\partial r} = -c(T - T_a)$$

must be satisfied along the boundary $CD\bar{C}$.

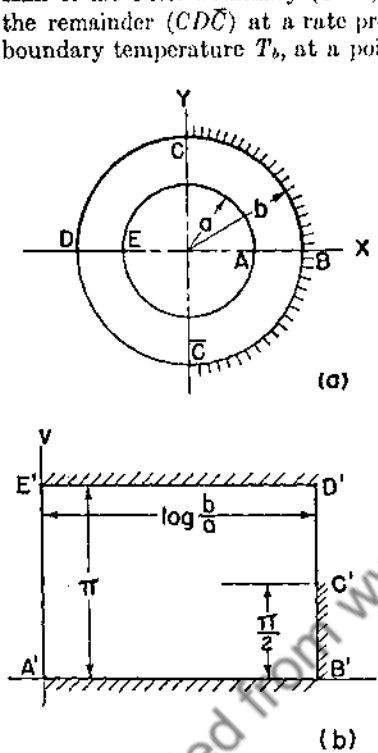


FIGURE 3.48

(a) By using the transformation of equation (263), show that the semisection $ABCDEA$ is mapped into the rectangle $A'B'C'D'E'A'$ of the uv-plane in Figure 3.48(b). Show also that the transformed problem consists in determining the solution of the equation

$$\frac{\partial^2 T'}{\partial u^2} + \frac{\partial^2 T'}{\partial v^2} = 0$$

for which $\partial T/\partial u = 0$ along $A'B'C'$ and along $D'E'$, $T = 200°$ along $E'A'$, and $\partial T/\partial u = -bc\,(T - T_a)$ along $C'D'$.

(b) If the rectangle in the uv-plane is replaced by a square net with spacing h, show that an appropriate condition along $C'D'$ is of the form

$$(1 + b\,h\,c)T_b = T_i + b\,h\,c\,T_a,$$

where T_b is the temperature at a boundary point, T_i that at the adjacent interior net point, and T_a that of the surrounding air.

(c) For convenience in calculation, suppose that $b/a = e^{\pi/2}$ and that $b\,c = 4/\pi$, so that $b\,h\,c = 2/(N + 1)$, where N represents the number of interior division points in the u-direction. Assuming also that $T_a = 100°$, obtain the approximate steady-state temperatures at the net points, taking $N = 1$. Arbitrarily consider the point C' to be part of the boundary $B'C'$, and the point D' to be part of $C'D'$.

(d) Indicate on a diagram the temperatures so obtained at corresponding points of the original region of Figure 3.48(a). Also, sketch a

few corresponding equithermal lines and lines of heat flow. (Notice that equithermal lines must intersect insulated boundaries at right angles.)

Section 3.20.

85. (a) Show that the solution of the equation $y' + y = 1$, for which $y(0) = 0$, is given by $y = 1 - e^{-x}$.

(b) By replacing dy/dx by the approximation $(y_{k+1} - y_k)/h$, and so obtaining the difference equation $y_{k+1} - (1 - h)y_k = h$, with $y_0 = 0$, obtain the approximate solution

$$y_k = 1 - (1 - h)^k \quad \text{or} \quad y(x_k) = 1 - (1 - h)^{x_k/h}.$$

Show that this expression converges to the exact solution as $h \to 0$. [Recall that $\lim_{\epsilon \to 0} (1 + \epsilon)^{1/\epsilon} = e$.]

(c) Show that the approximation $dy/dx \approx (y_k - y_{k-1})/h$ leads to the approximate solution $y(x_k) = 1 - (1 + h)^{-x_k/h}$, and that this expression also converges to the exact solution as $h \to 0$.

(d) Show that the approximation of dy/dx by $(y_{k+1} - y_{k-1})/2h$ (which also becomes exact as $h \to 0$, and which is in general more nearly accurate than either of the preceding approximations) leads to the difference equation $y_{k+1} + 2hy_k - y_{k-1} = 2h$, with general solution

$$y_k = 1 + c_1(\sqrt{1 + h^2} - h)^k + c_2 \cos \pi k(\sqrt{1 + h^2} + h)^k.$$

By setting $k = x_k/h$, show that the coefficient of c_2 does not tend to a limit as $h \to 0$ for fixed x_k, whereas the coefficient of c_1 tends to e^{-x_k}. In addition to prescribing the value $y_0 = 0$, one must prescribe a fictitious value, say, to y_{-1}. Show that unless the value $y_{-1} = 1 - h - \sqrt{1 + h^2}$ happens to be chosen, convergence to the exact solution (or to *any* function of x) cannot follow as $h \to 0$.

86. If $\beta(h)$ is a differentiable function of h for small values of h and at $h = 0$, and if $\beta(0) = 1$, show that

$$\lim_{h \to 0} [\beta(h)]^{1/h} = e^{\beta'(0)}.$$

[Write $u(h) = \beta^{1/h}$ and evaluate $\lim_{h \to 0} [\log u(h)]$ by using L'Hospital's rule.]

87. Given the differential equation $y'' + y' - 2y = 0$, show that the general solution of the difference equation obtained by replacing y'' by $(y_{k+1} - 2y_k + y_{k-1})/h^2$, and y' by either $(y_{k+1} - y_k)/h$, $(y_k - y_{k-1})/h$, or $(y_{k+1} - y_{k-1})/2h$, converges to the general solution of the differential equation as $h \to 0$. [In the first case, show that the general solution is of the form $y(x_k) = c_1\beta_1^{x_k/h} + c_2\beta_2^{x_k/h}$, where

$$\beta_{1,2}(h) = \frac{(2 + h + 2h^2) \pm h\sqrt{9 + 4h + 4h^2}}{2 + 2h},$$

and use the result of Problem 86, showing that $\beta_1'(0) = 1$ and $\beta_2'(0) = -2$.]

Section 3.21.

88. Verify that, if the conditions (273a,b) are replaced by the more general conditions $\phi(x, 0) = F(x)$ and $\phi_t(x, 0) = G(x)$, where F and G are twice differentiable functions of x, the relevant solution of (272) is of the form

$$\phi(x, t) = \frac{1}{2}[F(x - Vt) + F(x + Vt)] - \frac{1}{2V}[H(x - Vt) - H(x + Vt)]$$

where $H(x)$ is a function such that $H'(x) = G(x)$. Thus establish the fact that the region PAB of Figure 3.34 is the region of determination in this more general case.

89. Suppose that the function $F(x)$ of equation (273a) differs from zero only for $a \leqq x \leqq b$. Show that the solution (274), for $t > 0$, then differs from zero only in the two strips bounded by the lines $x - Vt = a$ and $x - Vt = b$, and by the lines $x + Vt = a$ and $x + Vt = b$, respectively. [Notice that any irregularities at points along the initial line $y = 0$ are therefore propagated along the characteristics which pass through those points.]

90. Let the independent variables x and y in the equation

$$a\,\phi_{xx} + 2b\,\phi_{xy} + c\,\phi_{yy} + d\,\phi_x + e\,\phi_y + f\,\phi = g$$

be replaced by new independent variables u and v, which are prescribed functions of x and y. By making the calculations $\phi_x = u_x\phi_u + v_x\phi_v$, $\phi_{xx} = u_x^2\phi_{uu} + 2u_xv_x\phi_{uv} + v_x^2\phi_{vv} + u_{xx}\phi_u + v_{xx}\phi_v$, and so forth, show that, if the original equation is written in the abbreviated form

$$L(\phi) + f\,\phi = g,$$

the transformed equation can be written in the form

$$B(u, u)\phi_{uu} + 2B(u, v)\phi_{uv} + B(v, v)\phi_{vv} + L(u)\phi_u + L(v)\phi_v + f\,\phi = g,$$

with the additional abbreviation

$$B(\alpha, \beta) \equiv a\,\alpha_x\beta_x + b(\alpha_x\beta_y + \alpha_y\beta_x) + c\,\alpha_y\beta_y.$$

91. Let $u = P(x, y)$ and $v = Q(x, y)$, where $P(x, y) = C_1$ and $Q(x, y) = C_2$ are independent integrals of the characteristic equation (281):

$$a(dy)^2 - 2b\,dx\,dy + c(dx)^2 = 0.$$

With the notation of Problem 90, show that then $B(u, u) = B(v, v) = 0$, so that the transformed equation takes the form of equation (282),

$$\phi_{uv} + A\,\phi_u + B\,\phi_v + C\,\phi = D,$$

with

$$A = \frac{L(u)}{2B(u, v)}, \quad B = \frac{L(v)}{2B(u, v)}, \quad C = \frac{f}{2B(u, v)}, \quad D = \frac{g}{2B(u, v)},$$

where the coefficients are to be expressed in terms of u and v.

[Notice that, if the equation $P(x, y) = C_1$ is considered as defining y in terms of x, there follows $P_x\,dx + P_y\,dy = 0$, and hence $dy/dx = -P_x/P_y$. Then show that the requirement that the function y so defined satisfy the characteristic equation leads to the result $B(P, P) = 0$.]

92. (a) Show that the characteristics of the equation

$$y\,\phi_{xx} - 2x\,\phi_{xy} - y\,\phi_{yy} = g(x, y)$$

are solutions of the equation

$$\frac{dx}{dy} = \frac{x}{y} \pm \sqrt{\frac{x^2}{y^2} + 1},$$

and (solving this equation by taking x/y as a new variable) obtain the equations of the characteristics in the form

$$\sqrt{x^2 + y^2} + x = c_1, \quad \sqrt{x^2 + y^2} - x = c_2.$$

(b) Noticing that here c_1 and c_2 cannot be negative, show that the characteristics are all members of the family $y^2 = k^2 - 2kx$ of confocal parabolas, where the members of the first set open to the left, and those of the second set open to the right. Sketch the two sets in the upper half-plane $y > 0$.

(c) If ϕ and $\partial\phi/\partial y$ are prescribed along the x-axis, and the solution of the given partial differential equation is required in the upper half-plane, represent in a sketch the region of determination for the point (3, 4).

93. If, in Problem 92, one writes

$$u = \sqrt{x^2 + y^2} + x, \quad v = \sqrt{x^2 + y^2} - x,$$

restricting attention to the upper half-plane $y > 0$, show that there follows also

$$x = \tfrac{1}{2}(u - v), \quad y = \sqrt{u\,v}.$$

Verify that the upper half of the xy-plane then corresponds to the first quadrant of the uv-plane, with the y-axis corresponding to the line $v = u$. Show also that the strip bounded by the x- and y-axes and the line $x = a$ corresponds to the diagonal strip bounded by the lines $u = v$, $u - v = 2a$, and the u-axis.

94. With the terminology of Problems 90 to 93, verify directly that $B(u, u) = B(v, v) = 0$ in the special case under consideration, and show that

$$B(u, v) = -2y = -2\sqrt{u\,v}, \quad L(u) = L(v) = \frac{y}{\sqrt{x^2 + y^2}} = \frac{2\sqrt{u\,v}}{u + v}.$$

Hence deduce that the differential equation of Problem 92 takes the form

$$\phi_{uv} - \frac{1}{2(u + v)}(\phi_u + \phi_v) + \frac{g}{4\sqrt{u\,v}} = 0$$

in terms of the characteristic variables u and v of the Problem 93. [Notice that the characteristics of the modified equation are then lines u = constant and v = constant, so that finite-difference methods (with net lines parallel to the u- and v-axes) are appropriate.]

95. Obtain an explicit solution to the finite-difference approximation of the problem in which $T(x, y)$ satisfies Laplace's equation $T_{xx} + T_{yy} = 0$ in the semi-infinite strip $0 < x < \pi$, $y > 0$, vanishes along the edges $x = 0$ and $x = \pi$, and takes on the value $T(x, 0) = \sin rx$ (where r is an integer) at points along the edge $y = 0$, by the following steps:

(a) Obtain the approximating difference equation

$$\kappa^2(T_{m+1,n} - 2T_{mn} + T_{m-1,n}) + (T_{m,n+1} - 2T_{mn} + T_{m,n-1}) = 0,$$

where $\kappa = h_y/h_x$, and where $T_{mn} \equiv T(m\,h_x, n\,h_y)$, subject to the conditions

$$T_{0n} = 0, \quad T_{Mn} = 0, \quad T_{m0} = \sin \frac{mr\pi}{M}.$$

(b) Assume a product solution $T_{mn} = f_m g_n$, and obtain the conditions

$$f_{m+1} - (2 - \lambda)f_m + f_{m-1} = 0, \quad f_0 = f_M = 0,$$

$$g_{n+1} - (2 + \kappa^2\lambda)g_n + g_{n-1} = 0, \quad \lim_{n \to \infty} g_n \text{ finite},$$

together with the condition relevant to $n = 0$.

(c) Show that there must follow $\lambda = 4 \sin^2 (r\pi/2M)$, and determine T_{mn} in the form

$$T_{mn} = e^{-\alpha n} \sin \frac{mr\pi}{M},$$

where α is a constant defined by the equation

$$\cosh \alpha = 1 + 2\kappa^2 \sin^2 \frac{r\pi}{2M}.$$

96. (a) Show that the solution of Problem 95 can be written in the form

$$T(x_m, y_n) = e^{-\beta y_n} \sin rx_m,$$

where

$$\beta \equiv \frac{\alpha}{h_y} = \frac{\cosh^{-1} [1 + 2\kappa^2 \sin^2 (rh_x/2)]}{\kappa\, h_x}.$$

(b) Show that, as the spacings tend to zero in such a way that their ratio κ retains *any* fixed value, the constant β tends to r, so that the solution tends to the expression

$$T(x, y) = e^{-ry} \sin rx,$$

and verify that this limiting solution is indeed the solution of the exact problem. [Show first that the function $f(u) \equiv \cosh^{-1}(1 + u^2)$ is given by $\sqrt{2}\left(u - \frac{u^3}{12} + \cdots\right)$ for small positive values of u, by considering the series expansion of $f'(u)$.]

Section 3.22.

97. In the problem governed by equations (283 to 285), let $\partial T/\partial t$ be replaced by a divided first forward difference, while $\partial^2 T/\partial x^2$ is not approximated.

(a) Show that the appropriate solution of the resultant difference-differential equation,

$$T(x, t_{k+1}) - T(x, t_k) = h_t \frac{\partial^2 T(x, t_k)}{\partial x^2},$$

is of the form

$$T(x, t_k) = (1 - r^2 h_t)^{t_k/h_t} \sin rx.$$

(b) Show that this solution converges to $e^{-r^2 t} \sin rx$ as $h_t \to 0$, for any fixed r, but that the solution oscillates unboundedly as $t \to \infty$ unless $h_t < 2/r^2$.

98. In the problem governed by equations (283 to 285), let $\partial^2 T/\partial x^2$ be replaced by a divided second central difference, while $\partial T/\partial t$ is not approximated.

(a) Show that the appropriate solution of the resultant difference-differential equation,

$$T(x_{k+1}, t) - 2T(x_k, t) + T(x_{k-1}, t) = h_x{}^2 \frac{\partial T(x_k, t)}{\partial t},$$

is of the form

$$T(x_k, t) = e^{-\left(\frac{4}{h_x{}^2} \sin^2 \frac{1}{2} r h_x\right) t} \sin rx_k.$$

(b) Show that this solution converges to $e^{-r^2 t} \sin rx$ as $h_x \to 0$ for any fixed r. (Notice also that the solution does not possess oscillations in time for any value of h_x.)

Section 3.23.

99. Show that the difference equation (275), replacing the wave equation $V^2 \phi_{xx} = \phi_{tt}$, is stable only when $h_x \geqq V h_t$. (Notice that, according to Section 3.21, this requirement is also the condition for *convergence*.)

100. Show that the result of approximating Laplace's equation by the difference equation

$$\frac{w_{m+1,n} - 2w_{m,n} + w_{m-1,n}}{h_x{}^2} + \frac{w_{m,n+1} - 2w_{m,n} + w_{m,n-1}}{h_y{}^2} = 0$$

would be unstable for *any* ratio h_x/h_y if it were treated as an initial-value problem.

101. Show that the result of replacing $(T_{m,n+1} - T_{m,n})/h_t$ by $(T_{m,n+1} - T_{m,n-1})/2h_t$ in equation (290) is unstable for any ratio h_x^2/h_t.

102. Show that the result of replacing $(w_{m+1,n} - w_{m,n})/h_x$ by $(w_{m+1,n} - w_{m-1,n})/2h_x$ in (334) is stable if

$$\frac{h_x^2}{h_t} \geqq 1 + \sqrt{1 - \tfrac{1}{4}h_x^2}.$$

103. Show that the result of replacing the heat-flow equation $w_{xx} = w_t$ by the difference equation

$$w_{m+1,n+1} - 2w_{m,n+1} + w_{m-1,n+1} = \kappa(w_{m,n+1} - w_{m,n}),$$

where $\kappa = h_x^2/h_t$, is stable for any positive value of κ.

104. Show that the result of replacing the wave equation $V^2\phi_{xx} = \phi_{tt}$ by the equation

$$\kappa^2(\phi_{m,n+1} - 2\phi_{m,n} + \phi_{m,n-1}) = \tfrac{1}{2}[(\phi_{m+1,n+1} - 2\phi_{m,n+1} + \phi_{m-1,n+1}) + (\phi_{m+1,n-1} - 2\phi_{m,n-1} + \phi_{m-1,n-1})],$$

where $\kappa = h_x/V h_t$, is stable for any positive value of κ.

105. Prove directly that *sufficient* conditions for stability of the formulation

$$w_{m,n+1} = c_1 w_{m+1,n} + c_2 w_{m,n} + c_3 w_{m-1,n} \qquad (1 \leqq m \leqq M - 1,\ n \geqq 1),$$

$$w_{0,n} = \mu_1 w_{1,n}, \qquad w_{M,n} = \mu_2 w_{M-1,n} \qquad (n \geqq 1),$$

$$w_{m,0} = f_m \qquad (0 \leqq m \leqq M)$$

are that the relations

$$c_1 \geqq 0, \quad c_2 \geqq 0, \quad c_3 \geqq 0, \quad c_1 + c_2 + c_3 \leqq 1,$$

$$0 \leqq \mu_1 \leqq 1, \quad 0 \leqq \mu_2 \leqq 1$$

be satisfied. [Suppose that f_m represents an error distribution. Let K denote the maximum value of $|f_m|$ for $0 \leqq m \leqq M$, and show by induction that then $|w_{m,n}| \leqq K$ (for all $0 \leqq m \leqq M$) when $n \geqq 0$.] Also, illustrate this result in the case of equation (290).

CHAPTER FOUR

Integral Equations

4.1. Introduction. An *integral equation* is an equation in which a function to be determined appears under an integral sign. We consider here only *linear* equations, that is, equations in which no nonlinear functions of the unknown function are involved.

Linear integral equations of most frequent occurrence in practice are conventionally divided into two classifications. First, an equation of the form

$$\alpha(x)f(x) = F(x) + \lambda \int_a^b K(x, \xi)f(\xi)\, d\xi, \tag{1}$$

where α, F, and K are given functions and λ, a, and b are constant, is known as a *Fredholm equation*. The function $f(x)$ is to be determined. The given function $K(x, \xi)$, which depends upon the current variable x as well as the auxiliary variable ξ, is known as the *kernel* of the integral equation. If the upper limit of the integral is not a constant, but is identified instead with the current variable, the equation takes the form

$$\alpha(x)f(x) = F(x) + \lambda \int_a^x K(x, \xi)f(\xi)\, d\xi, \tag{2}$$

and is known as a *Volterra equation*.

It is clear that the constant λ could be incorporated into the kernel $K(x, \xi)$ in both (1) and (2). However, in many applications this constant represents a significant parameter which may take on various values in a particular discussion. Also, it will be seen that the introduction of this parameter is advantageous in theoretical treatments.

When $\alpha \not\equiv 0$, the above equations involve the unknown function f both inside and outside the integral. In the special case when $\alpha \equiv 0$, the unknown function appears only under the integral sign, and the equation is known as an *integral equation of the first kind*, while in the case when $\alpha \equiv 1$ the equation is said to be of the *second kind*.

In the more general case when α is not a constant, but is a prescribed function of x, the equation is sometimes called an integral equation of the *third kind*. However, by suitably redefining the unknown function and/or the kernel, it is always possible to rewrite such an equation in the form of an equation of the second kind. In particular, when the function $\alpha(x)$ is positive throughout the interval (a, b), equation (1) can be rewritten in an equivalent symmetric form

$$\sqrt{\alpha(x)}\, f(x) = \frac{F(x)}{\sqrt{\alpha(x)}} + \lambda \int_a^b \frac{K(x, \xi)}{\sqrt{\alpha(x)\alpha(\xi)}} \sqrt{\alpha(\xi)}\, f(\xi)\, d\xi, \quad (3)$$

and hence, in this form, can be considered an integral equation of the *second* kind in the unknown function $\sqrt{\alpha(x)}\, f(x)$, with a modified kernel. Whereas other similar rearrangements are clearly possible, it frequently happens that $K(x, \xi)$ is a symmetric function of x and ξ; the modified kernel in (3) then preserves this symmetry. As will be seen, symmetric *kernels* are of the same importance in the theory of linear *integral* equations as are symmetric *matrices* in the theory of sets of linear *algebraic* equations (Chapter 1).

In the preceding equations the unknown function depends only upon one independent variable. If f depends upon two current variables x and y, the corresponding two-dimensional Fredholm equation is of the form

$$\alpha(x, y) f(x, y) = F(x, y) + \lambda \iint_R K(x, y; \xi, \eta) f(\xi, \eta)\, d\xi\, d\eta. \quad (4)$$

In general, an integral equation comprises the complete formulation of the problem, in the sense that additional conditions need not and cannot be specified. That is, auxiliary conditions are, in a sense, already written into the equation.

Certain integral equations can be deduced from or reduced to differential equations. In order to accomplish the reduction, it is

frequently necessary to make use of the known formula,

$$\frac{d}{dx}\int_{A(x)}^{B(x)} F(x, \xi)\, d\xi = \int_A^B \frac{\partial F(x, \xi)}{\partial x}\, d\xi + F[x, B(x)]\frac{dB}{dx} - F[x, A(x)]\frac{dA}{dx}, \tag{5}$$

for differentiation of an integral involving a parameter.*

As a useful application of this formula, we consider the differentiation of the function $I_n(x)$ defined by the equation

$$I_n(x) = \int_a^x (x - \xi)^{n-1} f(\xi)\, d\xi, \tag{6}$$

where n is a positive integer and a is a constant. With

$$F(x, \xi) = (x - \xi)^{n-1} f(\xi),$$

equation (5) gives the derivative of (6) in the form

$$\frac{dI_n}{dx} = (n - 1)\int_a^x (x - \xi)^{n-2} f(\xi)\, d\xi + [(x - \xi)^{n-1} f(\xi)]_{\xi=x}.$$

Hence, if $n > 1$, there follows

$$\frac{dI_n}{dx} = (n - 1) I_{n-1} \qquad (n > 1), \tag{7}$$

while if $n = 1$, we have

$$\frac{dI_1}{dx} = f(x). \tag{8}$$

Repeated use of (7) leads to the general relation

$$\frac{d^k I_n}{dx^k} = (n - 1)(n - 2) \cdots (n - k) I_{n-k} \qquad (n > k). \tag{9}$$

In particular, we obtain the result

$$\frac{d^{n-1} I_n}{dx^{n-1}} = (n - 1)!\, I_1(x), \tag{9a}$$

and hence by using (8), there follows

$$\frac{d^n I_n}{dx^n} = (n - 1)!\, f(x). \tag{9b}$$

* The formula of equation (5) is valid if both F and $\partial F/\partial x$ are *continuous* functions of both x and ξ.

If we notice that $I_n(a) = 0$ when $n \geqq 1$, it follows from (9) and (9a) that $I_n(x)$ and its first $(n-1)$ derivatives all vanish when $x = a$.

Thus we may conclude that $I_n(x)/(n-1)!$ is equivalent to the result of integrating $f(x)$ n times from a to x; that is, we have the result

$$\overbrace{\int_a^x \cdots \int_a^x}^{n \text{ times}} f(x) \overbrace{dx \cdots dx}^{n \text{ times}} = \frac{1}{(n-1)!} \int_a^x (x-\xi)^{n-1} f(\xi)\, d\xi. \tag{10}$$

This result will be useful in the work which follows.

4.2. Relations between differential and integral equations. We consider first the initial-value problem consisting of the linear second-order differential equation

$$\frac{d^2y}{dx^2} + A\frac{dy}{dx} + By = f(x), \tag{11}$$

where A and B may be functions of x, together with the prescribed initial conditions

$$y(a) = y_0, \qquad y'(a) = y_0'. \tag{12}$$

If we solve (11) for d^2y/dx^2, integrate the result with respect to x over the interval (a, x), and use (12), there follows

$$\frac{dy}{dx} - y_0' = -\int_a^x A\frac{dy}{dx}\,dx - \int_a^x B\,y\,dx + \int_a^x f\,dx$$

or, after integrating the first term on the right by parts,

$$\frac{dy}{dx} = -A\,y - \int_a^x (B - A')y\,dx + \int_a^x f\,dx + A(a)y_0 + y_0'.$$

A second integration then gives the relation

$$y - y_0 = -\int_a^x A(x)y(x)\,dx - \int_a^x\int_a^x [B(x) - A'(x)]y(x)\,dx\,dx + \int_a^x\int_a^x f(x)\,dx\,dx + [A(a)y_0 + y_0'](x-a).$$

If use is made of equation (10), this equation can be put in the form

$$y(x) = -\int_a^x \{A(\xi) + (x - \xi)[B(\xi) - A'(\xi)]\}y(\xi)\,d\xi + \int_a^x (x - \xi)f(\xi)\,d\xi + [A(a)y_0 + y_0'](x - a) + y_0$$

or, equivalently,

$$y(x) = \int_a^x K(x, \xi)y(\xi)\,d\xi + F(x), \tag{13}$$

where we have written

$$K(x, \xi) = (\xi - x)[B(\xi) - A'(\xi)] - A(\xi) \tag{14a}$$

and

$$F(x) = \int_a^x (x - \xi)f(\xi)\,d\xi + [A(a)y_0 + y_0'](x - a) + y_0. \tag{14b}$$

This equation is seen to be a *Volterra equation of the second kind.* We may notice that the kernel K is a *linear* function of the current variable x. It must be assumed, of course, that the coefficients A and B and the function $f(x)$ are such that the indicated integrals exist.

In illustration, the problem

$$\left.\begin{aligned} \frac{d^2y}{dx^2} + \lambda y &= f(x), \\ y(0) = 1, \qquad y'(0) &= 0 \end{aligned}\right\} \tag{15}$$

is transformed in this way to the integral equation

$$y(x) = \lambda \int_0^x (\xi - x)y(\xi)\,d\xi + 1 - \int_0^x (\xi - x)f(\xi)\,d\xi. \tag{16}$$

Conversely, the use of (5) permits the reduction of (13) to (11) by two differentiations. The initial conditions (12) are recovered by setting $x = a$ in (13) and in the result of the first differentiation. Thus, differentiation of (16) gives

$$\frac{dy}{dx} = -\lambda \int_0^x y(\xi)\,d\xi + \int_0^x f(\xi)\,d\xi, \tag{17}$$

and a second differentiation leads to the original differential equation. Since the integrals vanish when the upper and lower limits

coincide, equations (16) and (17) supply the initial values $y(0) = 1$ and $y'(0) = 0$.

To illustrate the corresponding procedure in the case of *boundary-value* problems, we consider first a simple example. Starting with the problem

$$\left.\begin{aligned} \frac{d^2y}{dx^2} + \lambda\, y &= 0, \\ y(0) = 0, \qquad y(a) &= 0 \end{aligned}\right\}, \tag{18}$$

we obtain after a first integration over $(0, x)$ the relation

$$\frac{dy}{dx} = -\lambda \int_0^x y(x)\, dx + C, \tag{19}$$

where C represents the unknown value of $y'(0)$. A second integration over $(0, x)$ then leads to the relation

$$y(x) = -\lambda \int_0^x (x - \xi) y(\xi)\, d\xi + C\, x. \tag{20}$$

While the condition $y(0) = 0$ has been incorporated into this relation, it remains to determine C so that the second end condition $y(a) = 0$ is satisfied. When this condition is imposed on (20) there follows

$$\lambda \int_0^a (a - \xi) y(\xi)\, d\xi = C\, a. \tag{21}$$

If the value of C so determined is introduced into (20), this relation takes the form

$$y(x) = -\lambda \int_0^x (x - \xi) y(\xi)\, d\xi + \lambda \frac{x}{a} \int_0^a (a - \xi) y(\xi)\, d\xi$$

or

$$y(x) = \lambda \int_0^x \frac{\xi}{a} (a - x) y(\xi)\, d\xi + \lambda \int_x^a \frac{x}{a} (a - \xi) y(\xi)\, d\xi. \tag{22}$$

With the abbreviation

$$K(x, \xi) = \begin{cases} \dfrac{\xi}{a}(a - x) & \text{when} \quad \xi < x, \\ \dfrac{x}{a}(a - \xi) & \text{when} \quad \xi > x, \end{cases} \tag{23}$$

equation (22) becomes

$$y(x) = \lambda \int_0^a K(x, \xi) y(\xi)\, d\xi. \tag{24}$$

Thus, the integral equation corresponding to the *boundary-value* problem (18) is a *Fredholm equation of the second kind.*

To recover (18) from (24), we differentiate the equal members of (22) twice, making use of (5), as follows:

$$\frac{dy}{dx} = \frac{\lambda}{a}\left[- \int_0^x \xi\, y(\xi)\, d\xi + x(a - x)y(x) + \int_x^a (a - \xi)y(\xi)\, d\xi - x(a - x)y(x) \right]$$

$$= \frac{\lambda}{a}\left[- \int_0^x \xi\, y(\xi)\, d\xi + \int_x^a (a - \xi)y(\xi)\, d\xi \right]$$

and

$$\frac{d^2y}{dx^2} = \frac{\lambda}{a}[-x\, y(x) - (a - x)y(x)] = -\lambda\, y(x),$$

in accordance with (18). The boundary conditions $y(0) = y(a) = 0$ follow directly from (22) by setting $x = 0$ and $x = a$.

We may notice that the kernel (23) has different analytic expressions in the two regions $\xi < x$ and $\xi > x$, but that the expressions are equivalent when $\xi = x$. Thus, if we think of K as *a function of* x, for a fixed value of ξ, then K is *continuous* at $x = \xi$. However, the derivative $\partial K/\partial x$ is given by $1 - \xi/a$ when $x < \xi$ and by $-\xi/a$ when $x > \xi$. Thus $\partial K/\partial x$ is discontinuous at $x = \xi$, and it has a finite jump of magnitude -1 as x *increases* through ξ. Further, we notice that in each region K is a linear function of x, that is, it satisfies the differential equation $\partial^2 K/\partial x^2 = 0$, and K vanishes at the end points $x = 0$ and $x = a$. Finally, it is seen that $K(x, \xi)$ is unchanged if x and ξ are interchanged; that is, $K(x, \xi) = K(\xi, x)$. Kernels having this last property are said to be *symmetric.*

If analogous methods are used in the case of the more general homogeneous second-order equation

$$\frac{d^2y}{dx^2} + A\frac{dy}{dx} + B\, y = 0,$$

with homogeneous end conditions, a kernel is obtained which is *discontinuous* at $x = \xi$ unless the coefficient A is zero (see Problem 8). However, a kernel which is continuous can, in general, be obtained by a different procedure which is outlined in the following section.

4.3. The Green's function. We consider first the problem consisting of the differential equation

$$L\,y + \Phi(x) = 0, \tag{25}$$

where L is the differential operator

$$L = \frac{d}{dx}\left(p\,\frac{d}{dx}\right) + q = p\,\frac{d^2}{dx^2} + \frac{dp}{dx}\,\frac{d}{dx} + q, \tag{25a}$$

together with homogeneous boundary conditions, each of the form $\alpha\,y + \beta\,\dfrac{dy}{dx} = 0$ for some constant values of α and β, which are imposed at the end points of an interval $a \leqq x \leqq b$.

In order to obtain a convenient form of the solution of this problem, we first attempt the determination of a function G which, for a given number ξ, is given by $G_1(x)$ when $x < \xi$ and by $G_2(x)$ when $x > \xi$, and which has the four following properties:

1. The functions G_1 and G_2 satisfy the equation $L\,G = 0$ in their intervals of definition; that is, $L\,G_1 = 0$ when $x < \xi$, and $L\,G_2 = 0$ when $x > \xi$.

2. The function G satisfies the homogeneous conditions prescribed at the end points $x = a$ and $x = b$; that is, G_1 satisfies the condition prescribed at $x = a$, and G_2 that corresponding to $x = b$.

3. The function G is continuous at $x = \xi$; that is, $G_1(\xi) = G_2(\xi)$.

4. The derivative of G has a discontinuity of magnitude $-1/[p(\xi)]$ at the point $x = \xi$; that is, $G_2'(\xi) - G_1'(\xi) = -1/[p(\xi)]$.

We then show that if this function G, in which ξ will appear as a parameter, exists, then the solution of the original problem is of the form

$$y(x) = \int_a^b \Phi(\xi)G(x,\,\xi)\,d\xi.$$

For this purpose, let $y = u(x)$ be a solution of $L\,y = 0$ which satisfies the prescribed homogeneous condition at $x = a$, and let $y = v(x)$ be a solution which satisfies the condition at $x = b$. Then the same is true of $c_1u(x)$ and $c_2v(x)$, where c_1 and c_2 are arbitrary

constants. Thus, conditions 1 and 2 are satisfied if we write $G_1 = c_1 u(x)$ and $G_2 = c_2 v(x)$, so that

$$G = \begin{cases} c_1 u(x) & \text{when} \quad x < \xi, \\ c_2 v(x) & \text{when} \quad x > \xi. \end{cases} \tag{26}$$

Conditions 3 and 4 then determine c_1 and c_2 in terms of the value of ξ. For 3 requires that

$$c_2 v(\xi) - c_1 u(\xi) = 0, \tag{27a}$$

while 4 gives the requirement

$$c_2 v'(\xi) - c_1 u'(\xi) = -\frac{1}{p(\xi)}. \tag{27b}$$

Equations (27a,b) possess a unique solution if the determinant

$$W[u(\xi), v(\xi)] \equiv \begin{vmatrix} u(\xi) & v(\xi) \\ u'(\xi) & v'(\xi) \end{vmatrix} = u(\xi)v'(\xi) - v(\xi)u'(\xi) \tag{28}$$

does not vanish. This quantity is the *Wronskian determinant* of the solutions u and v of the equation $L\,y = 0$, and it cannot vanish unless the functions u and v are linearly dependent. According to *Abel's formula,** this expression has the value $A/p(\xi)$, where A is a certain constant independent of ξ; that is, we have

$$u(\xi)v'(\xi) - v(\xi)u'(\xi) = \frac{A}{p(\xi)}. \tag{29}$$

With this relation, the solution of (27a,b) becomes

$$c_1 = -\frac{v(\xi)}{A}, \qquad c_2 = -\frac{u(\xi)}{A},$$

and hence (26) takes the form

$$G(x, \xi) = \begin{cases} -\dfrac{1}{A}\,u(x)v(\xi) & \text{when} \quad x < \xi, \\ -\dfrac{1}{A}\,u(\xi)v(x) & \text{when} \quad x > \xi, \end{cases} \tag{30}$$

* Abel's formula may be derived as follows: The requirements that $u(x)$ and $v(x)$ satisfy (25) are $(p\,u')' + q\,u = 0$ and $(p\,v')' + q\,v = 0$. By multiplying the second equation by u and the first by v, and subtracting the results, there follows $u(p\,v')' - v(p\,u')' \equiv [p(u\,v' - v\,u')]' = 0$. Hence we have $p(u\,v' - v\,u') = A$, where A is a constant, in accordance with (29).

where A is a *constant*, independent of x and ξ, which is determined by (29).

This determination fails if and only if A vanishes, so that u and v are linearly dependent, and hence are each multiples of a certain function $U(x)$. In this case, the function $U(x)$ satisfies the equation $L\,y = 0$ and *both* end conditions. Thus, for example, since the function $U(x) = 1$ solves the problem $d^2y/dx^2 = 0$, $y'(0) = y'(1) = 0$, the Green's function does not exist for the expression $L\,y \equiv d^2y/dx^2$, relevant to the end conditions $y'(0) = y'(1) = 0$. A generalized definition of G which is appropriate to such exceptional situations is given in Problem 16.

We now show that, with the definition of equation (30), the relation

$$y(x) = \int_a^b G(x, \xi)\Phi(\xi)\,d\xi \tag{31}$$

implies the differential equation

$$L\,y + \Phi(x) = 0, \tag{32}$$

together with the prescribed boundary conditions. For this purpose, we write (31) in the explicit form

$$y(x) = -\frac{1}{A}\left[\int_a^x v(x)u(\xi)\Phi(\xi)\,d\xi + \int_x^b u(x)v(\xi)\Phi(\xi)\,d\xi\right]. \tag{33}$$

Two differentiations, making use of (5), then lead to the relations

$$y'(x) = -\frac{1}{A}\left[\int_a^x v'(x)u(\xi)\Phi(\xi)\,d\xi + \int_x^b u'(x)v(\xi)\Phi(\xi)\,d\xi\right] \tag{34}$$

and

$$y''(x) = -\frac{1}{A}\left[\int_a^x v''(x)u(\xi)\Phi(\xi)\,d\xi + \int_x^b u''(x)v(\xi)\Phi(\xi)\,d\xi\right] - \frac{1}{A}[v'(x)u(x) - u'(x)v(x)]\Phi(x). \tag{35}$$

If we form the combination

$$L\,y \equiv p(x)y''(x) + p'(x)y'(x) + q(x)y(x)$$

from these results, and make use of (29), there follows

$$L\,y(x) = -\frac{1}{A}\left\{\int_a^x [L\,v(x)]u(\xi)\Phi(\xi)\,d\xi + \int_x^b [L\,u(x)]v(\xi)\Phi(\xi)\,d\xi\right\} - \frac{1}{A}\left[p(x)\cdot\frac{A}{p(x)}\cdot\Phi(x)\right].$$

But since $u(x)$ and $v(x)$ satisfy $L\,y = 0$, the two integrands vanish identically, and this relation becomes merely

$$L\,y(x) = -\Phi(x),$$

so that (31) implies (32). That is, the function y defined by (31) satisfies the differential equation (32). Also, since (33) and (34) give

$$y(a) = -\frac{1}{A}u(a)\int_a^b v(\xi)\Phi(\xi)\,d\xi,$$

$$y'(a) = -\frac{1}{A}u'(a)\int_a^b v(\xi)\Phi(\xi)\,d\xi,$$

it follows that the function $y(x)$ defined by (31) satisfies the same homogeneous conditions at $x = a$ as the function $u(x)$. But these conditions were specified as those which are imposed on the solution of (32). A similar statement applies to the satisfaction of the condition prescribed at $x = b$.

If now we replace $\Phi(x)$ by $\lambda\,r(x)y(x)$ it follows that satisfaction of the integral equation

$$y(x) = \lambda\int_a^b G(x,\,\xi)r(\xi)y(\xi)\,d\xi \tag{36}$$

implies satisfaction of the differential equation

$$L\,y(x) + \lambda\,r(x)y(x) = 0, \tag{37}$$

together with the relevant homogeneous boundary conditions. The converse statement can also be shown to be true (see Problem 14), so that the two formulations are entirely equivalent.

More generally, the presence of a prescribed function $f(x)$ in the right-hand member of (37) would correspond to the addition of the term $-\int_a^b G(x,\,\xi)f(\xi)\,d\xi$ to the right-hand member of (36).

We may notice that the kernel $K(x,\,\xi)$ of (36) is actually the product $G(x,\,\xi)r(\xi)$. While the definition (30) shows that $G(x,\,\xi)$ is

symmetric, the product $K(x, \xi)$ is *not* symmetric unless $r(x)$ is a constant. However, if we write

$$\sqrt{r(x)}\, y(x) = Y(x), \tag{38}$$

under the assumption that $r(x)$ is nonnegative over (a, b), as is usually the case in practice, equation (36) can be written in the form

$$Y(x) = \lambda \int_a^b \bar{K}(x, \xi) Y(\xi)\, d\xi, \tag{39}$$

where $\bar{K}$ is defined by the relation

$$\bar{K}(x, \xi) = \sqrt{r(x)r(\xi)}\, G(x, \xi), \tag{40}$$

and hence possesses the same symmetry as G. The importance of symmetry will be seen in later considerations.

The function $G(x, \xi)$ defined by (30), or by the properties 1 to 4 (page 388), is known as the *Green's function* associated with the differential expression $L\,y$ and the associated boundary conditions. In most physical problems, it is subject to a simple physical interpretation, as is illustrated in Section 4.5.

As an application of these results, we consider the problem

$$\left.\begin{aligned} x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + (\lambda x^2 - 1)y = 0, \\ y(0) = 0, \qquad y(1) = 0 \end{aligned}\right\}. \tag{41}$$

The differential equation is first put into the form of (25),

$$\frac{d}{dx}\left(x\frac{dy}{dx}\right) + \left(-\frac{1}{x} + \lambda x\right) y = 0,$$

from which there follows

$$L\,y = \frac{d}{dx}\left(x\frac{dy}{dx}\right) - \frac{y}{x}, \quad p = x, \quad q = -\frac{1}{x}, \quad r = x. \tag{42}$$

The general solution of the equation $L\,y = 0$ is found to be

$$y = c_1 x + c_2 x^{-1}.$$

As a solution for which $y(0) = 0$ we may take $y = u(x)$, where

$$u(x) = x, \tag{43}$$

and as a solution for which $y(1) = 0$ we may take $y = v(x)$, where

$$v(x) = \frac{1}{x} - x. \tag{44}$$

The Wronskian of u and v is then given by

$$u(x)v'(x) - v(x)u'(x) = -\frac{2}{x} \equiv \frac{-2}{p(x)},$$

and hence, with the notation of (29), we have

$$A = -2. \tag{45}$$

Thus (30) becomes

$$G(x, \xi) = \begin{cases} \dfrac{x}{2\xi}(1 - \xi^2) & \text{when } x \leqslant \xi, \\ \dfrac{\xi}{2x}(1 - x^2) & \text{when } x > \xi. \end{cases} \tag{46}$$

It follows from (36) that the problem (41) then corresponds to the integral equation

$$y(x) = \lambda \int_0^1 G(x, \xi)\xi\, y(\xi)\, d\xi. \tag{47}$$

It is easily seen that the Bessel equation (41) has no solution other than the trivial solution $y \equiv 0$, satisfying the prescribed end conditions, unless λ satisfies the characteristic equation

$$J_1(\sqrt{\lambda_n}) = 0, \tag{48}$$

in which case the solution is

$$y = c\, J_1(\sqrt{\lambda_n}\, x). \tag{49}$$

where c is arbitrary. The same statement must then apply to the integral equation (47), with G given by (46).

Similarly, from (18) it follows that the integral equation (24), with K given by (23), has no nontrivial solution unless $\lambda = n^2\pi^2/a^2$, where n is an integer, in which case $y = c \sin(n\pi x/a)$ is a solution for any arbitrary value of the constant c.

A completely analogous procedure can be used in transforming a boundary-value problem consisting of a homogeneous linear differential equation of order n, and relevant homogeneous boundary conditions, to a Fredholm integral equation. The Green's function

corresponding to $L\,y$ in the interval (a, b) then is to possess the following properties:

1. G satisfies the equation $L\,G = 0$ when $x < \xi$ and when $x > \xi$.
2. G satisfies the prescribed homogeneous boundary conditions.
3. G and its first $(n - 2)$ x-derivatives are continuous at $x = \xi$.
4. The $(n - 1)$th x-derivative of G has a jump of magnitude $-1/[s(\xi)]$ as x increases through ξ, where $s(x)$ is the coefficient of d^n/dx^n in L.

With the function G so defined, the relevant solution of the equation $L\,y + \Phi(x) = 0$ is given by

$$y(x) = \int_a^b G(x, \xi)\Phi(\xi)\,d\xi,$$

and also the problem consisting of the equation $L\,y + \lambda\,r\,y = f$ and the prescribed boundary conditions is equivalent to the Fredholm equation

$$y(x) = \lambda \int_a^b G(x, \xi)r(\xi)y(\xi)\,d\xi - \int_a^b G(x, \xi)f(\xi)\,d\xi.$$

In particular, for those *fourth-order* operators which are expressed in the form

$$L = \frac{d^2}{dx^2}\left[s(x)\frac{d^2}{dx^2}\right] + \frac{d}{dx}\left[p(x)\frac{d}{dx}\right] + q(x), \tag{50}$$

it will be found that the Green's function $G(x, \xi)$ is *symmetric*. Most of the linear fourth-order operators occurring in practice can be expressed in this form.

4.4. Alternative definition of the Green's function. A useful interpretation of the above definition of the Green's function may be obtained as follows. We again consider the problem consisting of the linear differential equation

$$L\,y + \Phi(x) = 0, \tag{51}$$

and suitably prescribed homogeneous boundary conditions at the ends of the interval (a, b). Suppose first that $\Phi(x)$ is replaced by a function $\Phi_\epsilon(x)$ which is zero in (a, b) except over a small interval $(\xi - \epsilon, \xi + \epsilon)$ about a point ξ, and is given by $1/(2\epsilon)$ over that interval, so that

$$\int_{\xi-\epsilon}^{\xi+\epsilon} \Phi_\epsilon(x)\,dx = 1. \tag{52}$$

If the equal members of the equation $L\,y + \Phi_\epsilon(x) = 0$ are integrated over $(\xi - \epsilon,\ \xi + \epsilon)$, it follows that the solution of that equation must be such that

$$\int_{\xi-\epsilon}^{\xi+\epsilon} L\,y\,dx = -1. \tag{53}$$

For explicitness, suppose that

$$L = \frac{d}{dx}\left(p\,\frac{d}{dx}\right) + q, \tag{54}$$

where $p(x)$ and $q(x)$ are continuous in the interval $(a,\ b)$. In this case, (53) takes the form

$$p\,\frac{dy}{dx}\Bigg|_{\xi-\epsilon}^{\xi+\epsilon} + \int_{\xi-\epsilon}^{\xi+\epsilon} q\,y\,dx = -1. \tag{55}$$

We are concerned with the limiting form of this relation as $\epsilon \to 0$.

If we require that y satisfy the equation $L\,y + \Phi_\epsilon(x) = 0$ throughout the interval $(a,\ b)$, the derivative of $p\,dy/dx$ must *exist* at all points of that interval and hence, in particular, the quantities $p\,dy/dx$ and $q\,y$ must remain *continuous* at the point $x = \xi$ as $\epsilon \to 0$. Thus the left-hand members of (55) then must tend to zero as $\epsilon \to 0$, so that (55) cannot be satisfied in the limit.

However, if we relax the requirement to the extent that, while y is still to be continuous throughout $(a,\ b)$, a discontinuity in dy/dx is permitted at the point $x = \xi$, it is seen that the limiting condition is satisfied if dy/dx has a jump of magnitude $-1/[p(\xi)]$ at $x = \xi$. If we require further that the differential equation $L\,y = 0$ be satisfied on both sides of this point, and that the boundary conditions be satisfied, we have exactly the conditions which define the Green's function of the preceding section. The same conclusion is readily obtained in the more general case of a linear operator L of order n, if we require that all derivatives of order less than $(n - 1)$ be continuous at $x = \xi$.

It is convenient (even though lacking in mathematical elegance) to say that, as $\epsilon \to 0$, the function $\Phi_\epsilon(x)$ "tends to the unit singularity function, with singularity at $x = \xi$." This latter "function" is then considered to be zero throughout $(a,\ b)$ except at the point $x = \xi$, and is imagined to be infinite at that point in such a way that the integral of the function across its singularity is unity.

This function is often known as the "unit impulse function" or as the "delta function."

The convention is extended to two- or three-dimensional space in an obvious way. Thus, in three dimensions, we start with a function which vanishes except inside a small *sphere* S_ϵ of radius ϵ surrounding a certain point Q, and which is so defined inside that sphere that its integral over the volume of the sphere is unity for all values of ϵ. We then solve a problem, the formulation of which involves that function, and consider the limit of the *solution* as the sphere S_ϵ enclosing the point Q shrinks to a point. It is then convenient to say that the limit of the solution (if it exists) is the "solution" corresponding to a "unit singularity function, with singularity at Q."

If we agree to the meaning of this convention, we may say that the Green's function, associated with a linear differential operator L and given boundary conditions, is the "solution" of the equation $L\,y + \delta_Q = 0$, subject to the same boundary conditions, where δ_Q is the unit singularity function (or "delta function"), with singularity at a point Q. The Green's function thus involves the coordinates of Q, as well as the current variables representing position in the space considered.

When only one independent variable is involved, the Green's function can be obtained by the procedure outlined in the preceding section, and we have seen that if it is of the form $G(x, \xi)$ the solution of the equation $L\,y + \Phi(x) = 0$, subject to the relevant homogeneous boundary conditions, is merely

$$y(x) = \int_a^b G(x, \xi)\Phi(\xi)\,d\xi.$$

In order to indicate the plausibility of the truth of an analogous statement in the more general case, we consider the determination of a function $w(x, y, z)$ which satisfies a linear partial differential equation of the form

$$L\,w + F(x, y, z) = 0 \tag{56}$$

inside a three-dimensional region R, together with appropriate homogeneous boundary conditions along the boundary of R. Let $G_\epsilon(x, y, z; \xi, \eta, \zeta)$ be a function which, for any relevant fixed values of ξ, η, and ζ, satisfies the equation

$$L\,G_\epsilon + \Phi_\epsilon = 0$$

and the same boundary conditions, where Φ_ϵ, considered as a function of (x, y, z), vanishes outside a sphere S_ϵ with center at the point $Q(\xi, \eta, \zeta)$ and radius ϵ, and has the property that

$$\iiint_{S_\epsilon} \Phi_\epsilon \, dx \, dy \, dz = 1.$$

Then also Φ_ϵ, considered as a function of (ξ, η, ζ), vanishes outside a sphere S'_ϵ with center at the point $P(x, y, z)$ and radius ϵ, and has the property that

$$\iiint_{S'_\epsilon} \Phi_\epsilon \, d\xi \, d\eta \, d\zeta = 1.$$

If we then define the function

$$w_\epsilon(x, y, z) = \iiint_R G_\epsilon(x, y, z; \xi, \eta, \zeta) F(\xi, \eta, \zeta) \, d\xi \, d\eta \, d\zeta,$$

and calculate $L\, w_\epsilon$ by *formally* differentiating under the integral sign, there follows

$$\begin{aligned} L\, w_\epsilon &= -\iiint_R F(\xi, \eta, \zeta) \Phi_\epsilon \, d\xi \, d\eta \, d\zeta \\ &= -\iiint_{S'_\epsilon} F(\xi, \eta, \zeta) \Phi_\epsilon \, d\xi \, d\eta \, d\zeta, \end{aligned}$$

where again S'_ϵ is a sphere of radius ϵ with center at the point $P(x, y, z)$. If the function F is continuous at P, its values in S'_ϵ will approximate $F(x, y, z)$ for small values of ϵ, so that it may be expected that the approximation

$$L\, w_\epsilon \approx -F(x, y, z) \iiint_{S'_\epsilon} \Phi_\epsilon \, d\xi \, d\eta \, d\zeta = -F(x, y, z)$$

will, in general, tend to an equality as ϵ tends to zero. Hence, if we denote the limits of G_ϵ and w_ϵ by G and w, respectively, (and if $L\, w_\epsilon$ tends to $L\, w$) these formal arguments indicate that the function

$$w(x, y, z) = \iiint_R G(x, y, z; \xi, \eta, \zeta) F(\xi, \eta, \zeta) \, d\xi \, d\eta \, d\zeta \qquad (57)$$

satisfies the differential equation (56). The rigorous establishment of this fact [and of the fact that (57) also satisfies the same homogeneous conditions as does the Green's function G] is complicated by the fact that G generally becomes infinite when the points $P(x, y, z)$ and $Q(\xi, \eta, \zeta)$ coincide, but is possible in most practical cases.

As this statement implies, the function G will not satisfy the differential equation *at* the point Q where the unit singularity is located. In the special case of the partial differential operator

$$L = \frac{\partial}{\partial x}\left(p\,\frac{\partial}{\partial x}\right) + \frac{\partial}{\partial y}\left(p\,\frac{\partial}{\partial y}\right) + \frac{\partial}{\partial z}\left(p\frac{\partial}{\partial z}\right) + q, \tag{58}$$

associated with a three-dimensional region, it is found that G must behave near the point $Q(\xi, \eta, \zeta)$ in such a way that the integral of the *normal derivative* of G over the *surface* of the sphere S_ϵ tends to $-1/p[(\xi, \eta, \zeta)]$ as the radius ϵ tends to zero:

$$\lim_{\epsilon \to 0} \oiint_{S_\epsilon} \frac{\partial G}{\partial n}\, dS = -\frac{1}{p(\xi, \eta, \zeta)}. \tag{59}$$

Here $\partial G/\partial n$ represents the derivative of G in the direction of the *outward* normal at points of the spherical boundary.

This result can be obtained by noticing first that the equation $L\,G_\epsilon + \Phi_\epsilon = 0$ can be written, in terms of the vector differential operator ∇, in the form

$$\nabla \cdot (p\,\nabla G_\epsilon) + q\,G_\epsilon = -\Phi_\epsilon.$$

If the equal members of this equation are integrated over the volume V_ϵ bounded by the sphere S_ϵ with center at $Q(\xi, \eta, \zeta)$, and the condition

$$\iiint_{V_\epsilon} \Phi_\epsilon\, dV = 1$$

is imposed, it follows that G_ϵ must satisfy the condition

$$\iiint_{V_\epsilon} \nabla \cdot (p\,\nabla G_\epsilon)\, dV + \iiint_{V_\epsilon} q\,G_\epsilon\, dV = -1.$$

The first volume integral on the left can be transformed to a surface integral, by use of the *divergence theorem*, so that this condition takes the form

$$\iint_{S_\epsilon} p\,\frac{\partial G_\epsilon}{\partial n}\, dS + \iiint_{V_\epsilon} q\,G_\epsilon\, dV = -1.$$

It is clear that this condition cannot be satisfied in the limit as $\epsilon \to 0$ if G_ϵ and its first partial derivatives are required to remain finite at the point Q in the limit. For small values of ϵ, the first term on the left is approximated by $4\pi\epsilon^2$ times the mean value of

$p\,\partial G_\epsilon/\partial n$ on S_ϵ, while the second term is approximated by $4\pi\epsilon^3/3$ times the mean value of $q\,G_\epsilon$ in V_ϵ. If we represent radial distance from the point Q by the variable r, so that $\partial G_\epsilon/\partial n = \partial G_\epsilon/\partial r$ on S_ϵ, the first term thus approaches a finite nonzero limit as ϵ tends to zero if and only if $\partial G_\epsilon/\partial r$ becomes infinite like $1/r^2$ on S_ϵ as $\epsilon \to 0$. In this case, G_ϵ becomes infinite like $1/r$ and the second term is thus small of order ϵ^2 when ϵ is small. Hence, as $\epsilon \to 0$, we must require that the first term tend to the value -1. For small values of ϵ, the function p may be evaluated at Q and the condition (59) follows.

In the case of the special operator

$$L = \frac{\partial}{\partial x}\left(p\,\frac{\partial}{\partial x}\right) + \frac{\partial}{\partial y}\left(p\,\frac{\partial}{\partial y}\right) + q, \tag{60}$$

associated with a *two*-dimensional problem, we consider a *circle* C_ϵ, of radius ϵ, surrounding the point $Q(\xi, \eta)$. The condition corresponding to (59) then requires that the integral of $\partial G/\partial n$ around the perimeter of C_ϵ tend to $-1/[p(\xi, \eta)]$ as ϵ tends to zero:

$$\lim_{\epsilon \to 0} \oint_{C_\epsilon} \frac{\partial G}{\partial n}\,ds = -\frac{1}{p(\xi, \eta)}. \tag{61}$$

The condition (59) or (61), together with the requirements that G satisfy the equation $L\,G = 0$ except at the point Q, as well as the prescribed boundary conditions, serves (in general) to determine the Green's function relevant to the operator (58) or (60).

In the two-dimensional case, it is convenient to represent by r the distance from the point $Q(\xi, \eta)$ to the point $P(x, y)$,

$$r = \sqrt{(x - \xi)^2 + (y - \eta)^2}. \tag{62}$$

On the circle C_ϵ surrounding Q, we may then write $ds = r\,d\theta$, where θ represents angular position and $r = \epsilon$. Equation (61) can then be written in the form

$$\lim_{r \to 0} \int_0^{2\pi} \frac{\partial G}{\partial r}\,r\,d\theta = -\frac{1}{p(\xi, \eta)}. \tag{61'}$$

This condition can be satisfied only if $r\,\partial G/\partial r$ tends to the value $-1/[2\pi\,p(\xi, \eta)]$ as r tends to zero. Thus the function G must behave like $-(\log r)/[2\pi\,p(\xi, \eta)]$ in the neighborhood of the point $Q(\xi, \eta)$.

In particular, when $p = 1$, the Green's function relevant to the operator (60) must be of the form

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log \sqrt{(x-\xi)^2 + (y-\eta)^2} + g(x, y; \xi, \eta), \quad (63)$$

where g satisfies the equation $L\,g = \nabla^2 g + q\,g = (q\ 2\pi) \log r$ in the prescribed region, and is so determined that the right-hand member of (63) satisfies the prescribed boundary conditions. When also $q = 0$, the equation $L\,w + \Phi = 0$ becomes *Poisson's equation*,

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \Phi = 0. \quad (64)$$

If the Green's function (63) is known for a region R with specified homogeneous boundary conditions, then the solution of (64) in R which satisfies those conditions is given by the integral

$$w = \iint_R G(x, y; \xi, \eta)\Phi(\xi, \eta)\, d\xi\, d\eta. \quad (64a)$$

Similar considerations lead to the fact that in the three-dimensional case of (58) the Green's function must behave like $r^{-1}/[4\pi\, p(\xi, \eta, \zeta)]$ near the point $Q(\xi, \eta, \zeta)$, where here r represents the distance

$$r = \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}. \quad (65)$$

In particular, when $p = 1$, the Green's function relevant to (58) must be of the form

$$G(x, y, z; \xi, \eta, \zeta) = \frac{1}{4\pi} \frac{1}{\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}} + g(x, y, z; \xi, \eta, \zeta), \quad (66)$$

where g satisfies $\nabla^2 g + q\,g = -q/4\pi r$ everywhere in the given region and $g + 1/4\pi r$ satisfies the prescribed boundary conditions.

We may notice that whereas the Green's function relevant to the one-dimensional operator (54) merely possesses a discontinuous derivative at Q, that function relevant to (60) becomes logarithmically infinite at Q, while that corresponding to (58) becomes infinite like $1/r$, where r represents distance from Q.

In the case of the operator (54), which is clearly a one-dimensional specialization of (60), the circle C_ϵ is replaced by a *line*

segment extending from the point $\xi - \epsilon$ to the point $\xi + \epsilon$. The *outward* derivative of G at the point $x = \xi + \epsilon$ is given by the value of dG/dx at that point, whereas the outward derivative at $x = \xi - \epsilon$ is given by the *negative* of dG/dx at that point. The requirement that the *sum* of these outward derivatives tend to $-1/[p(\xi)]$ as ϵ tends to zero,

$$\lim_{\epsilon \to 0} \left[\left. \frac{dG}{dx} \right|_{\xi+\epsilon} - \left. \frac{dG}{dx} \right|_{\xi-\epsilon} \right] = -\frac{1}{p(\xi)},$$

is seen to be analogous to (59) and (61), and identical with condition 4 of page 388.

We have seen that the Green's function can be defined, first, as the limit of the solution of a certain problem in which a prescribed function tends to a unit singularity function and, second, as a function satisfying a differential equation (with boundary conditions) except at a certain point, and having a certain prescribed behavior near that point. The two definitions are equivalent, the second usually being more convenient than the first in actual applications. In the following section a third alternative interpretation of the Green's function is presented.

4.5. Linear equations in cause and effect. The influence function. Linear integral equations arise most frequently in physical problems as a result of the possibility of superimposing the effects due to several causes. To indicate the general reasoning involved, we suppose that x and ξ are variables, each of which may take on all values in a certain common interval or region R. We may, for example, think of x and ξ as each representing position (in space of one, two, or three dimensions) or time. We suppose further that a distribution of causes is active over the region R, and that we are interested in studying the resultant distribution of effects in R.

If the effect at x due to a *unit* cause concentrated at ξ is denoted by the function $G(x, \xi)$, then the differential effect at x due to a uniform distribution of causes of intensity $c(\xi)$ over an elementary region $(\xi, \xi + d\xi)$ is given by $c(\xi)G(x, \xi)\, d\xi$. Hence the effect $e(x)$ at x, due to a distribution of causes $c(\xi)$ over the entire region R is given by the integral

$$e(x) = \int_R G(x, \xi)c(\xi)\, d\xi \tag{67}$$

if superposition is valid, that is, if the effect due to the sum of two separate causes is (exactly or approximately) the sum of the effects due to each of the causes.

The function $G(x, \xi)$, which represents *the effect at x due to a unit concentrated cause at ξ*, is often known as the *influence function* of the problem. As may be expected, this function is either identical with or proportional to the Green's function defined in the preceding sections, when that definition is applicable.

If the distribution of *causes* is prescribed, and if the influence function is known, (67) permits the determination of the effect by direct integration. However, if it is required to determine a distribution of causes which will produce a known or desired *effect* distribution, (67) represents a *Fredholm integral equation of the first kind* for the determination of c. The kernel is then identified with the influence function of the problem.

If, instead, the physical problem prescribes neither the cause nor the effect separately, but requires merely that they satisfy a certain linear relation of the form

$$c(x) = \phi(x) + \lambda e(x), \tag{68}$$

where ϕ is a given function or zero and λ is a constant, then the effect e can be eliminated between (67) and (68), to give the relation

$$e(x) = \phi(x) + \lambda \int_R G(x, \xi) e(\xi)\, d\xi. \tag{69}$$

This relation is a Fredholm integral equation of the *second* kind, for the determination of the cause distribution. Alternatively, if the cause c is eliminated between (67) and (68), the equation

$$e(x) = \int_R G(x, \xi)\phi(\xi)\, d\xi + \lambda \int_R G(x, \xi) e(\xi)\, d\xi \tag{70}$$

serves to determine the effect distribution. Both cause and effect are determined by solving either (69) or (70), and using (68).

As an explicit example of such derivations, we consider the study of small deflections of a string fixed at the points $x = 0$ and $x = a$, under a loading distribution of intensity $p(x)$. We suppose that the string is initially so tightly stretched that nonuniformity of the tension, due to small deflections, can be neglected. If a *unit*

concentrated load is applied in the y-direction at an arbitrary point ξ (Figure 4.1), the string will then be deflected into two linear parts with a corner at the point $x = \xi$. If we denote the (approximately)

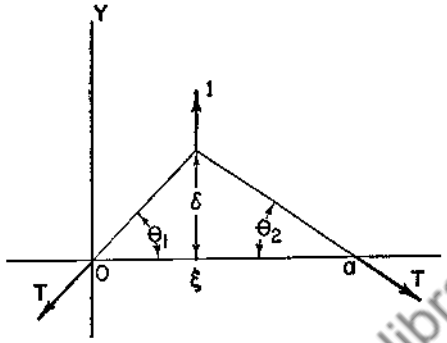


FIGURE 4.1

uniform tension by T, the requirement of force equilibrium in the y-direction leads to the condition

$$T \sin \theta_1 + T \sin \theta_2 = 1, \tag{71}$$

with the notation of Figure 4.1. For small deflections (and slopes) we have the approximations

$$\left.\begin{aligned} \sin \theta_1 \approx \tan \theta_1 &= \frac{\delta}{\xi}, \\ \sin \theta_2 \approx \tan \theta_2 &= \frac{\delta}{a - \xi} \end{aligned}\right\}, \tag{72}$$

where δ is the maximum deflection of the string, at the loaded point ξ. The introduction of these approximations into (71) leads to the relation

$$T\left(\frac{\delta}{\xi} + \frac{\delta}{a - \xi}\right) = 1,$$

and hence determines the deflection δ in the form

$$\delta = \frac{1}{T\,a}\,\xi(a - \xi). \tag{73}$$

The equation of the corresponding deflection curve is then readily obtained in the form

$$y = \begin{cases} \delta \dfrac{x}{\xi} & \text{when} \quad x < \xi, \\ \delta \dfrac{a - x}{a - \xi} & \text{when} \quad x > \xi, \end{cases} \tag{74}$$

where δ is given by (73), so that the *influence function* (for small deflections) is given by the expression

$$G(x, \xi) = \begin{cases} \dfrac{x}{T a} (a - \xi) & \text{when} \quad x < \xi, \\ \dfrac{\xi}{T a} (a - x) & \text{when} \quad x > \xi. \end{cases} \tag{75}$$

Hence, by superposition, the deflection $y(x)$ due to a loading distribution $p(x)$ is given by

$$y(x) = \int_0^a G(x, \xi) p(\xi)\, d\xi. \tag{76}$$

If the *deflection* is prescribed, this relation constitutes an integral equation of the first kind for the determination of the necessary loading distribution.

Suppose next that the string is rotating uniformly about the x-axis, with angular velocity ω, and that in addition a continuous distribution of loading $f(x)$ is imposed in the direction outward from the axis of revolution. If the linear mass density of the string is denoted by $\rho(x)$, the total effective load intensity can be written in the form

$$p(x) = \omega^2 \rho(x) y(x) + f(x), \tag{77}$$

so that (76) takes the form

$$y(x) = \omega^2 \int_0^a G(x, \xi) \rho(\xi) y(\xi)\, d\xi + \int_0^a G(x, \xi) f(\xi)\, d\xi. \tag{78}$$

It may be noticed that the influence function (75) differs from the kernel (23) of Section 4.2 only in a multiplicative factor $1/T$. By performing two differentiations, as in the reduction of (24) to (18), it is easily shown that (78) is equivalent to the more familiar formulation in terms of a differential equation with boundary conditions,

$$\left.\begin{aligned} T \frac{d^2 y}{dx^2} + \rho\, \omega^2 y + f = 0, \\ y(0) = 0, \qquad y(a) = 0 \end{aligned}\right\}. \tag{79}$$

If we write the differential equation in the form $L\,y + \Phi = 0$, where $L = T\,d^2/dx^2$ and $\Phi = \rho\,\omega^2 y + f$, we may recall that the *Green's function* of the problem would then be that function which satisfies $T\,d^2y/dx^2 = 0$ except at $x = \xi$, which vanishes when $x = 0$ and $x = a$, which is continuous at $x = \xi$, and for which the vertical force resultant $T\,dy/dx$ *decreases abruptly by unity* at $x = \xi$. These are precisely the conditions which determined the *influence function*.

It is of interest to notice that if a concentrated mass m_0 were, in addition, attached to the rotating string at the point $x = x_0$, the integral equation (78) would be modified by merely adding the deflection $m_0\omega^2 y(x_0)\,G\,(x, x_0)$ to the right-hand member. The corresponding modification in the differential-equation formulation would be somewhat more complicated.

In many physical problems, the Green's function is obtained empirically, and is specified only by a table of numerical values. Thus, for example, in studying small deflections of a beam of irregular cross section, subject to certain physical end restraints, a number of points $x_1, x_2, \ldots, x_n$ may be first selected along the span of the beam. By applying loads successively at the points $\xi = x_j$ and (in each case) measuring the deflections at each of the points $x = x_i$, a table of values of the influence function $G(x, \xi)$ can be obtained, giving deflections at points $x = x_i$ due to *unit* loads applied at points $\xi = x_j$. If the beam extends from $x = 0$ to $x = a$, we thus obtain n^2 entries, specifying $G(x, \xi) = G(x_i, x_j)$ at symmetrically placed points of the square $(0 \leqq x \leqq a,\ 0 \leqq \xi \leqq a)$ in a fictitious $x\xi$-plane.

It is known that the deflection at a point x, due to a unit load at a point ξ, is equal to the deflection at ξ, due to a unit load at x. The truth of this *reciprocity* relation reduces the number of necessary measurements by a factor of nearly two, and shows that the relevant Green's function is *symmetric;* that is, $G(x, \xi) = G(\xi, x)$. Accordingly, the matrix of entries $G(x_i, x_j)$ is symmetrical with respect to its principal diagonal.

The determination of small deflections of the same beam when it is rotating and subject to a radial force distribution can then be based on the solution of an integral equation of the same form as (78). Numerical methods which are appropriate to the solution of such an equation, whether $G(x, \xi)$ is given analytically or by a table of values, are described in later sections of this chapter.

It is important to notice that the influence function itself incorporates the end conditions appropriate to the problem. Thus, as was mentioned previously, the integral equation serves to specify the problem completely.

Similar formulations of two- and three-dimensional problems are clearly possible. However, for the purpose of simplicity, attention will be restricted in most of what follows to problems involving only one independent variable. We consider next certain analytical procedures which are available for the *exact* solution of certain linear integral equations, after which we describe numerical methods of obtaining *approximate* solutions.

4.6. Fredholm equations with separable kernels. We shall speak of a kernel $K(x, \xi)$ as *separable* if it can be expressed as the sum of a finite number of terms, each of which is the product of a function of x alone and a function of ξ alone. Such a kernel is thus expressible in the form

$$K(x, \xi) = \sum_{n=1}^{N} f_n(x) g_n(\xi). \tag{80}$$

There is no loss in generality if we assume that the N functions $f_n(x)$ are *linearly independent* in the relevant interval.

Any *polynomial* in x and ξ is of this type. Further, we may notice, for example, that the kernel $\sin(x + \xi)$ is separable in this sense, since we can write

$$\sin(x + \xi) = \sin x \cos \xi + \cos x \sin \xi.$$

Integral equations with separable kernels do not occur frequently in practice. However, they are easily treated and, furthermore, the results of their consideration lead to a better understanding of integral equations of more general type. Also, it is often possible to apply the methods to be developed in this section to the *approximate* solution of Fredholm equations in which the kernel can be satisfactorily *approximated* by a polynomial in x and ξ, or by a separable kernel of more general form.

A Fredholm equation of the second kind, with (80) as its kernel, can be written in the form

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi + F(x) \tag{81}$$

or

$$y(x) = \lambda \int_a^b \left[\sum_{n=1}^{N} f_n(x) g_n(\xi) \right] y(\xi)\, d\xi + F(x)$$

or

$$y(x) = \lambda \sum_{n=1}^{N} f_n(x) \left[\int_a^b g_n(\xi) y(\xi)\, d\xi \right] + F(x). \tag{82}$$

It is clear that the coefficients of $f_1(x)$, $f_2(x)$, . . . , $f_N(x)$ in (82) are *constants*, although their values are unknown. If we introduce the abbreviation

$$c_n = \int_a^b g_n(x) y(x)\, dx \qquad (n = 1, 2, \cdots, N), \tag{83}$$

equation (82) takes the form

$$y(x) = F(x) + \lambda \sum_{n=1}^{N} c_n f_n(x). \tag{84}$$

This is thus the form of the required *solution* of the integral equation (81), and it remains only to determine the N constants c_1, c_2, . . . , c_N.

For this purpose, we may obtain N equations involving only these N constants and calculable quantities by multiplying both members of (84) successively by $g_1(x)$, $g_2(x)$, . . . , $g_N(x)$, and integrating the results over the interval (a, b). This procedure introduces calculable integrals which may be denoted as follows:

$$\alpha_{mn} = \int_a^b g_m(x) f_n(x)\, dx, \tag{85a}$$

$$\beta_m = \int_a^b g_m(x) F(x)\, dx. \tag{85b}$$

With these abbreviations, the N equations so obtained can be written in the form

$$\left.\begin{array}{l} (1 - \lambda\, \alpha_{11})c_1 - \lambda\, \alpha_{12}c_2 - \lambda\, \alpha_{13}c_3 + \cdots - \lambda\, \alpha_{1N}c_N = \beta_1, \\ -\lambda\, \alpha_{21}c_1 + (1 - \lambda\, \alpha_{22})c_2 - \lambda\, \alpha_{23}c_3 + \cdots - \lambda\, \alpha_{2N}c_N = \beta_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ -\lambda\, \alpha_{N1}c_1 - \lambda\, \alpha_{N2}c_2 - \lambda\, \alpha_{N3}c_3 + \cdots + (1 - \lambda\, \alpha_{NN})c_N = \beta_N \end{array}\right\}. \tag{86}$$

This set of equations possesses a *unique* solution for the c's if and only if the determinant Δ of the coefficients of the c's does not vanish. In the matrix notation of Chapter 1, the set may be written in the abbreviated form

$$(\mathbf{I} - \lambda\, \boldsymbol{\alpha})\mathbf{c} = \boldsymbol{\beta}, \tag{87}$$

where $\mathbf{I}$ is the unit matrix of order N and $\boldsymbol{\alpha}$ is the matrix $[\alpha_{ij}]$. Thus the results of Chapter 1, relevant to sets of linear equations, are immediately applicable to the present discussion of solutions of the integral equation (81), with a separable kernel.

If the function $F(x)$ is identically zero in (81), the integral equation is said to be *homogeneous*, and is obviously satisfied by the *trivial solution* $y(x) \equiv 0$, corresponding to the trivial solution $c_1 = c_2 = \ldots = c_N = 0$ of (86) when the right-hand members vanish. Unless the determinant $\Delta \equiv |\, \mathbf{I} - \lambda\, \boldsymbol{\alpha}\, |$ vanishes, this is the *only* solution. However, if $\Delta = 0$, at least one of the c's can be assigned arbitrarily, and the remaining c's can be determined accordingly. Thus, in such cases, infinitely many solutions of the integral equation (81) exist.

Those values of λ for which $\Delta(\lambda) = 0$ are known as the *characteristic values* (or *eigenvalues*), and any nontrivial solution of the homogeneous integral equation (with a convenient choice of the arbitrary constant or constants) is then called a corresponding *characteristic function* (eigenfunction) of the integral equation. If k of the constants $c_1, c_2, \ldots, c_N$ can be assigned arbitrarily for a given characteristic value of λ, then k linearly independent corresponding characteristic functions are obtained.

If the function $F(x)$ is not identically zero, but is *orthogonal* to all the functions $g_1(x), g_2(x), \ldots, g_N(x)$, equation (85b) shows that the right-hand members of (86) again vanish. The preceding discussion again applies to this case, except for the fact that here the solution (84) of the integral equation involves also the function $F(x)$. The trivial values $c_1 = c_2 = \ldots = c_N = 0$ thus lead to the solution $y = F(x)$. Solutions corresponding to characteristic values of λ are now expressed as the sum of $F(x)$ and arbitrary multiples of characteristic functions.

Finally, if at least one right-hand member of (86) does not vanish, a unique nontrivial solution of (86) exists, leading to a unique nontrivial solution of the integral equation (81), *if the*

determinant $\Delta(\lambda)$ *does not vanish.* However, in this case, *if* $\Delta(\lambda)$ *does vanish* equations (86) are either *incompatible,* and no solution exists, or they are *redundant,* and infinitely many solutions exist.

4.7. Illustrative example. The several possible cases just discussed may be illustrated by a consideration of the integral equation

$$y(x) = \lambda \int_0^1 (1 - 3x\xi)y(\xi)\, d\xi + F(x). \tag{88}$$

This equation can be rewritten in the form

$$y(x) = \lambda(c_1 - 3c_2x) + F(x), \tag{89}$$

where

$$c_1 = \int_0^1 y(\xi)\, d\xi, \qquad c_2 = \int_0^1 \xi\, y(\xi)\, d\xi. \tag{90}$$

To determine c_1 and c_2, we multiply both sides of (89) successively by 1 and x and integrate the results over (0, 1), to obtain the equations

$$c_1 = \lambda\left(c_1 - \frac{3}{2}c_2\right) + \int_0^1 F(x)\, dx,$$

$$c_2 = \lambda\left(\frac{1}{2}c_1 - c_2\right) + \int_0^1 x\, F(x)\, dx,$$

or

$$\left.\begin{aligned} (1 - \lambda)c_1 + \frac{3}{2}\lambda\, c_2 &= \int_0^1 F(x)\, dx, \\ -\frac{1}{2}\lambda\, c_1 + (1 + \lambda)c_2 &= \int_0^1 x\, F(x)\, dx \end{aligned}\right\}. \tag{91a,b}$$

The determinant of coefficients is given by

$$\Delta(\lambda) = \begin{vmatrix} 1 - \lambda & \frac{3}{2}\lambda \\ -\frac{1}{2}\lambda & 1 + \lambda \end{vmatrix} = \tfrac{1}{4}(4 - \lambda^2). \tag{92}$$

It follows that a *unique* solution exists if and only if

$$\lambda \neq \pm 2, \tag{93}$$

and is obtained by solving (91a,b) for c_1 and c_2, and introducing the results into (89). In particular, if $F(x) = 0$ and $\lambda \neq \pm 2$, the only solution is the trivial one, $y(x) = 0$. The numbers $\lambda = \pm 2$ are the characteristic numbers of the problem.

If $\lambda = +2$, equations (91a,b) take the form

$$\left.\begin{aligned} -c_1 + 3c_2 &= \int_0^1 F(x)\,dx, \\ -c_1 + 3c_2 &= \int_0^1 x\,F(x)\,dx \end{aligned}\right\}, \tag{94a,b}$$

while if $\lambda = -2$ equations (91a,b) become

$$\left.\begin{aligned} c_1 - c_2 &= \frac{1}{3}\int_0^1 F(x)\,dx, \\ c_1 - c_2 &= \int_0^1 x\,F(x)\,dx \end{aligned}\right\}. \tag{95a,b}$$

Equations (94a,b) are incompatible unless the prescribed function $F(x)$ satisfies the condition

$$\int_0^1 F(x)\,dx = \int_0^1 x\,F(x)\,dx \quad \text{or} \quad \int_0^1 (1-x)F(x)\,dx = 0, \tag{96}$$

whereas (95a,b) are incompatible unless

$$\frac{1}{3}\int_0^1 F(x)\,dx = \int_0^1 x\,F(x)\,dx \quad \text{or} \quad \int_0^1 (1-3x)F(x)\,dx = 0, \tag{97}$$

in which cases the corresponding equation pairs (94) or (95) are redundant.

We consider first the case when

$$F(x) = 0, \tag{98}$$

so that (88) is *homogeneous*. Then, if $\lambda \neq \pm 2$, the only solution is the trivial one $y(x) = 0$, as was mentioned above. If $\lambda = 2$ (and $F = 0$) equations (94) are redundant, and either equation gives the single condition $c_1 = 3c_2$. Thus (89) then gives the solution

$$y(x) = A(1 - x) \quad \text{when} \quad \lambda = 2, \tag{99}$$

where $A = 6c_2$ is an *arbitrary* constant. Thus the function $1 - x$ (or any convenient multiple of that function) is the *characteristic function* corresponding to the characteristic number $\lambda = +2$.

In a similar way, we find the solution

$$y(x) = B(1 - 3x) \quad \text{when} \quad \lambda = -2, \tag{100}$$

where $B = -2c_1 = -2c_2$ is an arbitrary constant, so that $1 - 3x$ is the characteristic function corresponding to $\lambda = -2$.

Equation (89) shows that any solution of (88) is expressible in the form

$$y(x) = F(x) + C_1(1 - x) + C_2(1 - 3x), \tag{101}$$

where we have written $C_1 = 3\lambda(c_1 - c_2)/2$ and $C_2 = \lambda(3c_2 - c_1)/2$. Thus it follows that *any solution of* (88) *can be expressed as the sum of* $F(x)$ *and some linear combination of the characteristic functions.* The fact that this statement can be applied to a wide class of integral equations is of basic importance, as will be seen.

In the *nonhomogeneous* case, $F(x) \not\equiv 0$, a *unique solution* exists if $\lambda \neq \pm 2$. If $\lambda = 2$, equation (96) shows that *no solution exists unless* $F(x)$ *is orthogonal to* $1 - x$ *over the relevant interval* $(0, 1)$, that is, *unless* $F(x)$ *is orthogonal to the characteristic function corresponding to* $\lambda = 2$.* If F satisfies this restriction equations (94a,b) are again equivalent. Hence, if we use (94a), we may obtain $c_1 = 3c_2 - \int_0^1 F(x)\,dx$, so that (89) gives the solution as follows:

$$\lambda = 2: \quad y(x) = F(x) - 2\int_0^1 F(x)\,dx + A(1 - x),$$

when $$\int_0^1 (1 - x)F(x)\,dx = 0. \tag{102}$$

Here $A = 6c_2$ is again an arbitrary constant. Thus, in this case, *infinitely many* solutions exist, *differing from each other by a multiple of the relevant characteristic function.*

Similarly, if $\lambda = -2$ there is no solution unless $F(x)$ is orthogonal to $1 - 3x$ over $(0, 1)$, in which case infinitely many solutions exist as follows:

$$\lambda = -2: \quad y(x) = F(x) - \frac{2}{3}\int_0^1 F(x)\,dx + B(1 - 3x),$$

when $$\int_0^1 (1 - 3x)F(x)\,dx = 0. \tag{103}$$

Here $B = -2c_2$ is an arbitrary constant.

4.8. Hilbert-Schmidt theory. In those cases when the kernel $K(x, \xi)$ of a homogeneous Fredholm equation is not of the form (80), in particular, if $K(x, \xi)$ is given by different analytic expressions in the intervals for which $x < \xi$ and $x > \xi$, there are

* As will be seen in the following section, this situation is a consequence of the *symmetry* of the kernel $K(x, \xi) = 1 - 3x\xi$ in (88).

generally infinitely many characteristic numbers λ_n ($n = 1, 2, 3, \ldots$), each corresponding to a characteristic function defined within an arbitrary multiplicative constant. In exceptional cases, a given characteristic number λ_k may correspond to two or more independent characteristic functions. In this section we investigate certain properties of these characteristic functions.

Let $y_m(x)$ and $y_n(x)$ be characteristic functions corresponding respectively to two *different* characteristic numbers λ_m and λ_n of the homogeneous Fredholm equation

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{104}$$

and *suppose that the kernel* $K(x, \xi)$ *is symmetric*, so that

$$K(x, \xi) = K(\xi, x). \tag{105}$$

As has been indicated in preceding sections, such kernels are of frequent occurrence in the formulation of physically motivated problems. We may notice that $\lambda = 0$ *cannot* be a characteristic number since it leads necessarily to the trivial solution $y(x) \equiv 0$.

The functions y_m and y_n must accordingly satisfy the equations

$$\left.\begin{aligned} y_m(x) &= \lambda_m \int_a^b K(x, \xi) y_m(\xi)\, d\xi, \\ y_n(x) &= \lambda_n \int_a^b K(x, \xi) y_n(\xi)\, d\xi \end{aligned}\right\}. \tag{106a,b}$$

If we multiply both members of (106a) by $y_n(x)$, and integrate the results with respect to x over (a, b), there then follows

$$\int_a^b y_m(x) y_n(x)\, dx = \lambda_m \int_a^b y_n(x) \left[\int_a^b K(x, \xi) y_m(\xi)\, d\xi \right] dx. \tag{107}$$

If the order of integration is reversed in the right-hand member, equation (107) becomes

$$\int_a^b y_m(x) y_n(x)\, dx = \lambda_m \int_a^b y_m(\xi) \left[\int_a^b K(x, \xi) y_n(x)\, dx \right] d\xi. \tag{108}$$

We now make use of the assumed *symmetry* (105) to rewrite the inner integral on the right in the form

$$\int_a^b K(\xi, x) y_n(x)\, dx.$$

But this integral differs from the coefficient of λ_n in (106b) only in that x and ξ are interchanged, and hence it is equivalent to

$$\frac{1}{\lambda_n} y_n(\xi),$$

so that (108) becomes

$$\int_a^b y_m(x)y_n(x)\,dx = \frac{\lambda_m}{\lambda_n}\int_a^b y_m(\xi)y_n(\xi)\,d\xi. \tag{109}$$

Since the integrals in (109) are equivalent, (109) can be rewritten in the form

$$(\lambda_m - \lambda_n)\int_a^b y_m(x)y_n(x)\,dx = 0. \tag{110}$$

Thus we conclude that *if $y_m(x)$ and $y_n(x)$ are characteristic functions of* (104) *corresponding to distinct characteristic numbers, then $y_m(x)$ and $y_n(x)$ are orthogonal over the interval* (a, b).

If two or more linearly independent characteristic functions correspond to the same characteristic number, then an equal number of orthogonalized linear combinations can be formed by the Schmidt procedure (Section 1.12). In the remainder of this chapter it will be assumed that this has been done, when such exceptional cases arise.

It is important to notice that the preceding results apply only to a *symmetric* kernel.

We show next that *the characteristic numbers of a Fredholm equation with a real symmetric kernel are all real.** This result is established by noticing that if λ_m were a *complex* characteristic number, corresponding to a complex characteristic function $y_m(x)$, then the complex conjugate number $\bar{\lambda}_m$ would necessarily also be a characteristic number, corresponding to the characteristic function $\bar{y}_m(x)$ which is the complex conjugate of $y_m(x)$. Hence, by replacing λ_n by $\bar{\lambda}_m$ and y_n by $\bar{y}_m$ in (110), it would follow that

$$(\lambda_m - \bar{\lambda}_m)\int_a^b y_m(x)\bar{y}_m(x)\,dx = 0.$$

If we write $\lambda_m = \alpha_m + i\,\beta_m$ and $y_m(x) = f_m(x) + i\,g_m(x)$, this relation takes the form

$$2i\,\beta_m \int_a^b (f_m^2 + g_m^2)\,dx = 0.$$

* Proof that such an equation always possesses at least one characteristic number, when K is continuous, is omitted.

But, since $y_m(x) \not\equiv 0$, the integral cannot vanish, and we conclude that the imaginary part of λ_m must vanish, as was to be shown.

A Fredholm equation with a nonsymmetric kernel may possess characteristic numbers which are not real.

In more advanced works* the following basic theorem is established:

Any function $f(x)$ which can be generated from a continuous function $\Phi(x)$ by the operation $\int_a^b K(x, \xi)\Phi(\xi)\, d\xi$, where $K(x, \xi)$ is continuous and symmetric, so that

$$f(x) = \int_a^b K(x, \xi)\Phi(\xi)\, d\xi$$

for *some* continuous function Φ, *can be represented over* (a, b) *by a linear combination of the characteristic functions* $y_1(x), y_2(x), \cdots$, *of the homogeneous Fredholm integral equation* (104) *with* $K(x, \xi)$ *as its kernel.*†

Because of the orthogonality, the coefficients in the representation

$$f(x) = \sum_n A_n y_n(x) \quad (a \leqq x \leqq b) \tag{111a}$$

are then determined by the familiar formula

$$A_n \int_a^b [y_n(x)]^2\, dx = \int_a^b f(x) y_n(x)\, dx \quad (n = 1, 2, \cdots). \tag{111b}$$

In those cases when only a finite number of characteristic functions exist, the functions generated by the operation

$$\int_a^b K(x, \xi)\Phi(\xi)\, d\xi$$

form a very restricted class. For example, if $K(x, \xi) = \sin(x + \xi)$ and $(a, b) = (0, 2\pi)$, there follows

$$\int_0^{2\pi} K(x, \xi)\Phi(\xi)\, d\xi = \int_0^{2\pi} (\sin x \cos \xi + \cos x \sin \xi)\Phi(\xi)\, d\xi$$

$$= \left[\int_0^{2\pi} \Phi(\xi) \cos \xi\, d\xi\right] \sin x + \left[\int_0^{2\pi} \Phi(\xi) \sin \xi\, d\xi\right] \cos x, \tag{112}$$

* See, for example, Reference 2.

† When the set of characteristic functions is infinite, the resultant infinite series converges *absolutely* and *uniformly* in the interval (a, b).

and hence this operation can generate only functions of the form

$$f(x) = C_1 \sin x + C_2 \cos x, \tag{113}$$

regardless of the form of Φ. The characteristic functions of the associated homogeneous Fredholm integral equation

$$y(x) = \lambda \int_0^{2\pi} \sin (x + \xi) y(\xi)\, d\xi \tag{114}$$

are readily found, by the methods of the preceding section, to be arbitrary multiples of the functions $y_1(x) = \sin x + \cos x$ and $y_2(x) = \sin x - \cos x$, corresponding respectively to $\lambda_1 = 1/\pi$ and $\lambda_2 = -1/\pi$. It is obvious that any function of the form (113), generated by $\int_0^{2\pi} \sin (x + \xi)\Phi(\xi)\, d\xi$, can indeed be expressed as a linear combination of $y_1(x)$ and $y_2(x)$.

Even though the number of relevant independent characteristic functions be infinite, it is not necessarily true that *any* continuous function $f(x)$ defined over (a, b) can be represented over that interval by a series of these functions; that is, the set of characteristic functions, even though infinite in number, may not comprise a *complete* set, in the sense defined in Section 1.28.

Suppose now that we have (in some manner) obtained all members of the set of normalized characteristic functions $y_n(x)$, each corresponding to a characteristic number λ_n of the homogeneous equation

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{115}$$

where $K(x, \xi)$ is *symmetric*. We next show that the knowledge of these functions and constants permits a simple determination of the solution of the corresponding *nonhomogeneous* Fredholm equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{116}$$

of the second kind, when a solution exists.

We suppose that the characteristic numbers have been ordered with respect to magnitude, that characteristic numbers corresponding to k independent characteristic functions have been counted k times, and that such subsets of independent characteristic functions (corresponding to multiple characteristic numbers) have been orthogonalized.

In order to simplify the relations which follow, we suppose also that the arbitrary multiplicative constant associated with each characteristic function y_n is so chosen that the function is *normalized* over the relevant interval (a, b). Thus we write

$$\phi_n = C_n y_n, \tag{117}$$

where the normalizing factor C_n is given by

$$C_n = \frac{1}{\sqrt{\int_a^b [y_n(x)]^2\, dx}} \tag{118}$$

so that there follows

$$\int_a^b [\phi_n(x)]^2\, dx = 1. \tag{119}$$

The expansion (111) of a function $f(x)$ in a series of the normalized characteristic functions then takes the simpler form

$$f(x) = \sum_n a_n \phi_n(x) \quad \text{where} \quad a_n = \int_a^b f(x)\phi_n(x)\, dx. \tag{120}$$

Since the series (111) and (120) are identical, it follows that

$$a_n \phi_n = A_n y_n, \qquad A_n = a_n C_n. \tag{121a,b}$$

These relations permit transition from the expressions to be obtained to corresponding expressions involving nonnormalized characteristic functions.

If the equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi \tag{122}$$

possesses a solution $y(x)$, then the function $y(x) - F(x)$ is generated by the operation $\int_a^b K(x, \xi)[\lambda\, y(\xi)]\, d\xi$, and hence it can be represented by a series (or linear combination) of the normalized characteristic functions $\phi_n(x)$ $(n = 1, 2, \ldots)$, of the form

$$y(x) - F(x) = \sum_n a_n \phi_n(x) \qquad (a \leqq x \leqq b), \tag{123}$$

where the coefficients a_n are given, in virtue of (120), by

$$a_n = \int_a^b [y(x) - F(x)]\phi_n(x)\, dx. \tag{124}$$

With the convenient abbreviations

$$c_n = \int_a^b y(x)\phi_n(x)\,dx, \qquad f_n = \int_a^b F(x)\phi_n(x)\,dx, \quad (125a,b)$$

this relation takes the form

$$a_n = c_n - f_n. \tag{126}$$

In order to obtain a second relation which permits the elimination of the unknown integral $c_n = \int_a^b y\,\phi_n\,dx$ from (126), we multiply both members of (122) by $\phi_n(x)$ and integrate the results over (a, b), so that there follows

$$c_n = f_n + \lambda \int_a^b \phi_n(x) \left[\int_a^b K(x, \xi)y(\xi)\,d\xi \right] dx, \tag{127}$$

with the notation of (126). If the order of integration is reversed in the coefficient of λ, and use is made of the assumed *symmetry* in $K(x, \xi)$, that coefficient becomes

$$\int_a^b y(\xi) \left[\int_a^b K(\xi, x)\phi_n(x)\,dx \right] d\xi = \frac{1}{\lambda_n} \int_a^b y(\xi)\phi_n(\xi)\,d\xi = \frac{c_n}{\lambda_n},$$

and hence (127) is equivalent to the relation

$$c_n = f_n + \frac{\lambda}{\lambda_n} c_n. \tag{128}$$

The elimination of c_n between (126) and (128) then gives

$$a_n = \frac{\lambda}{\lambda_n - \lambda} f_n \tag{129}$$

if $\lambda \neq \lambda_n$. Hence the required solution (123) to (122) takes the form

$$y(x) = F(x) + \lambda \sum_n \frac{f_n}{\lambda_n - \lambda} \phi_n(x) \qquad (\lambda \neq \lambda_n) \tag{130}$$

where, as previously defined,

$$f_n = \int_a^b F(x)\phi_n(x)\,dx \qquad (n = 1, 2, \cdots). \tag{131}$$

We may notice that the constants f_n would be the coefficients in the expansion $F(x) = \Sigma f_n\phi_n(x)$ if $F(x)$ were representable by

such an expansion. It is of some importance to notice that in the preceding derivation it was not necessary to assume the validity of this representation.

The expansion (130) exists uniquely if and only if λ does not take on a characteristic value. Because of the presence of the term $f_k\phi_k(x)/(\lambda_k - \lambda)$, we see that if $\lambda = \lambda_k$, where λ_k is the kth characteristic number, the solution (130) is *nonexistent* unless also $f_k = 0$, that is, *unless $F(x)$ is orthogonal to the corresponding characteristic function or functions.* But if $\lambda = \lambda_k$ and $f_k = 0$ equation (128) reduces to the trivial identity when $n = k$, and hence imposes no restriction on c_k. From (126) it then follows that the coefficient of $\phi_k(x)$ in (130), which formally assumes the form $0/0$, is truly *arbitrary*, so that in this case (122) possesses *infinitely many* solutions, differing from each other by arbitrary multiples of $\phi_k(x)$. If λ assumes a characteristic value and $F(x)$ is *not* orthogonal to the corresponding characteristic function or functions, *no solution exists.* These results are illustrated by the example of Section 4.7, which involves the symmetric kernel $K(x, \xi) = 1 - 3x\xi$.

In virtue of (121), the normalization of the characteristic functions is unnecessary, in the sense that (130) can be replaced by the expression

$$y(x) = F(x) + \lambda \sum_n \frac{F_n}{\lambda_n - \lambda} y_n(x) \qquad (\lambda \neq \lambda_n), \tag{130'}$$

where

$$F_n \int_a^b [y_n(x)]^2\, dx = \int_a^b F(x) y_n(x)\, dx \qquad (n = 1, 2, \cdots). \tag{131'}$$

We consider next the Fredholm equation of the *first* kind,

$$F(x) = \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{132}$$

with a *symmetric kernel*, where F is prescribed and y is to be determined. It follows from the basic expansion theorem (page 414) that (132) has *no continuous solution* unless $F(x)$ can be expressed as a linear combination of the characteristic functions corresponding to the associated homogeneous equation of the second kind,

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi. \tag{133}$$

For example, in the special case for which $K(x, \xi) = \sin(x + \xi)$ and $(a, b) = (0, 2\pi)$, equation (132) becomes

$$F(x) = \int_0^{2\pi} y(\xi) \sin(x + \xi)\, d\xi \tag{134a}$$

or

$$F(x) = \left[\int_0^{2\pi} y(\xi) \cos \xi\, d\xi\right] \sin x + \left[\int_0^{2\pi} y(\xi) \sin \xi\, d\xi\right] \cos x. \tag{134b}$$

This relation can be satisfied only if $F(x)$ is prescribed as a linear combination of $\sin x$ and $\cos x$ or, equivalently, as a related linear combination of the characteristic functions y_1 and y_2 of the associated homogeneous equation (114),

$$y_1 = \sin x + \cos x, \quad y_2 = \sin x - \cos x, \tag{135}$$

corresponding to $\lambda_1 = 1/\pi$ and $\lambda_2 = -1/\pi$. If $F(x)$ is prescribed as

$$F(x) = A \sin x + B \cos x, \tag{136}$$

then (134b) is satisfied by any function y for which

$$\int_0^{2\pi} y(\xi) \cos \xi\, d\xi = A, \quad \int_0^{2\pi} y(\xi) \sin \xi\, d\xi = B. \tag{137}$$

One such function is clearly

$$y(x) = \frac{1}{\pi}(A \cos x + B \sin x). \tag{138}$$

However, if we add to (138) *any function* which is orthogonal to both $\sin x$ and $\cos x$, and hence to the characteristic functions y_1 and y_2, over $(0, 2\pi)$, the conditions (137) will still be satisfied, so that the solution is by no means unique. Unless $F(x)$ is prescribed in the form (136), no solution exists.

Suppose now that (132) *does* possess a continuous solution. Then $F(x)$ is generated from $y(x)$ by the operation $\int_a^b K(x, \xi) y(\xi)\, d\xi$, and hence it can be expanded in a series

$$F(x) = \sum_n f_n \phi_n(x) \qquad (a \leqq x \leqq b), \tag{139}$$

where
$$f_n = \int_a^b F(x)\phi_n(x)\,dx, \tag{140}$$

and where ϕ_n is the nth characteristic function of (133). The series may be finite or infinite. But since ϕ_n satisfies the equation

$$\phi_n(x) = \lambda_n \int_a^b K(x, \xi)\phi_n(\xi)\,d\xi, \tag{141}$$

and since (132) must be satisfied, we may replace $F(x)$ and $\phi_n(x)$ by the right-hand members of (132) and (141), so that (139) takes the form

$$\int_a^b K(x, \xi)y(\xi)\,d\xi = \sum_n \lambda_n f_n \int_a^b K(x, \xi)\phi_n(\xi)\,d\xi \tag{142}$$

or

$$\int_a^b K(x, \xi)\left[y(\xi) - \sum_n \lambda_n f_n \phi_n(\xi)\right]d\xi = 0.* \tag{143}$$

This condition is satisfied if and only if $y(x)$ is of the form

$$y(x) = \sum_n \lambda_n f_n \phi_n(x) + \Phi(x), \tag{144}$$

where $\Phi(x)$ is a solution of the equation

$$\int_a^b K(x, \xi)\Phi(\xi)\,d\xi = 0. \tag{145}$$

We conclude that if (132) possesses a continuous solution, then that solution must be of the form (144), where Φ is any continuous function satisfying (145). From the homogeneity of (145) it is clear that either (145) is satisfied only by the trivial function $\Phi(x) \equiv 0$ or it possesses infinitely many solutions.

If we multiply both members of (145) by $\phi_n(x)$, integrate the results over (a, b) and make use of the assumed symmetry in K, we obtain the condition

$$\int_a^b \phi_n(x)\left[\int_a^b K(x, \xi)\Phi(\xi)\,d\xi\right]dx = \int_a^b \Phi(\xi)\left[\int_a^b K(\xi, x)\phi_n(x)\,dx\right]d\xi$$
$$= \frac{1}{\lambda_n}\int_a^b \Phi(\xi)\phi_n(\xi)\,d\xi = 0. \tag{146}$$

* The validity of the interchange of order of integration and summation is a consequence of the uniformity of the convergence of (139) (see footnote on page 414).

Hence it follows that if (145) possesses a nontrivial solution, then that solution must be orthogonal to all the characteristic functions ϕ_n. If this set of functions is *finite*, then infinitely many linearly independent functions satisfying this condition exist. If the functions ϕ_n comprise an infinite *complete* set over (a, b), then *no* continuous nontrivial function can be simultaneously orthogonal to all functions of the set, so that in this case the function Φ in (144) must be identically zero.

In the preceding developments we have made use of a known expansion theorem to show that if a Fredholm equation with a symmetric kernel possesses a solution, then that solution must be of a certain form. In particular, if the nonhomogeneous equation (116), of the second kind, possesses a solution, then that solution is *unique* unless λ assumes a characteristic value, and is given by (130). If the equation (132), of the first kind, possesses a solution, then it is given by (144) and it is or is not uniquely defined, according as (145) does not or does possess nontrivial solutions. *

In physically motivated problems, questions concerning existence and uniqueness of a solution usually can be resolved by physical considerations. Thus, for example, if the kernel in (145) is the Green's function (75) for a loaded string, a nontrivial solution of (145) clearly cannot exist, since it would represent a static loading which leads to no deflection at any point of the string. From the mathematical point of view, however, such questions are of considerable interest. Certain known results are presented in Sections 4.9 and 4.10.

It should be remarked that before the theory of this section can be applied it is necessary to determine the characteristic numbers and functions of the homogeneous equation. Except in special cases this determination must depend upon numerical (or graphical) procedures, certain of which are discussed in Section 4.14.

4.9. Iterative methods for solving equations of the second kind. In certain cases integral equations of the second kind can be solved by a method of successive approximations. In this section we describe the method and investigate its validity.

* It is possible to prove, by direct substitution, that (130) and (144) do indeed satisfy (116) and (132), respectively, *when the infinite series involved converge uniformly.*

Suppose that in a Fredholm equation of the second kind,

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{147}$$

we replace y under the integral sign by an initial approximation $y^{(0)}$. Then (147) determines an approximation $y^{(1)}$ in the form

$$y^{(1)}(x) = F(x) + \lambda \int_a^b K(x, \xi) y^{(0)}(\xi)\, d\xi. \tag{148}$$

By substituting this approximation into the right-hand member of (147), we then obtain the next approximation $y^{(2)}$, and continue the process in such a way that successive approximations are determined by the formula

$$y^{(n)}(x) = F(x) + \lambda \int_a^b K(x, \xi) y^{(n-1)}(\xi)\, d\xi. \tag{149}$$

The same method is clearly applicable also when the upper limit b is replaced by the current variable x, so that the equation is of the *Volterra* type. It remains to determine under what conditions the successive approximations actually tend toward a solution of (147).

In order to examine this procedure more closely, we write out explicitly the results of the indicated substitutions. Thus we first obtain the result of replacing y in the right-hand member of (147) by $y^{(1)}$, as given by (148). In this substitution, we must replace the current variable x in (148) by the dummy variable ξ appearing in (147). To avoid ambiguity, we must then replace ξ in (148) by another dummy variable, say ξ_1, so that (148) becomes

$$y^{(1)}(\xi) = F(\xi) + \lambda \int_a^b K(\xi, \xi_1) y^{(0)}(\xi_1)\, d\xi_1.$$

The result of the substitution then takes the form

$$y^{(2)}(x) = F(x) + \lambda \int_a^b K(x, \xi) \left[F(\xi) + \lambda \int_a^b K(\xi, \xi_1) y^{(0)}(\xi_1)\, d\xi_1 \right] d\xi$$

or

$$y^{(2)}(x) = F(x) + \lambda \int_a^b K(x, \xi) F(\xi)\, d\xi + \lambda^2 \int_a^b K(x, \xi) \int_a^b K(\xi, \xi_1) y^{(0)}(\xi_1)\, d\xi_1\, d\xi. \tag{150}$$

If we now replace x by ξ, ξ by ξ_1, and ξ_1 by ξ_2, in (150), and substitute once more in the right-hand member of (147), there follows

$$y^{(3)}(x) = F(x) + \lambda \int_a^b K(x, \xi)F(\xi)\, d\xi + \lambda^2 \int_a^b K(x, \xi) \int_a^b K(\xi, \xi_1)F(\xi_1)\, d\xi_1\, d\xi + \lambda^3 \int_a^b K(x, \xi) \int_a^b K(\xi, \xi_1) \int_a^b K(\xi_1, \xi_2)y^{(0)}(\xi_2)\, d\xi_2\, d\xi_1\, d\xi. \quad (151)$$

The analysis is abbreviated considerably if we introduce an *integral operator* $\mathcal{K}$, defined by the equation

$$\mathcal{K} f(x) \equiv \int_a^b K(x, \xi)f(\xi)\, d\xi. \quad (152)$$

The integral equation (147) then takes the symbolic form

$$y(x) = F(x) + \lambda \mathcal{K} y(x), \quad (153)$$

while (149) becomes

$$y^{(n)}(x) = F(x) + \lambda \mathcal{K} y^{(n-1)}(x). \quad (154)$$

Further, equations (148), (150), and (151) take the form

$$\left.\begin{aligned} y^{(1)}(x) &= F(x) + \lambda \mathcal{K} y^{(0)}(x), \\ y^{(2)}(x) &= F(x) + \lambda \mathcal{K} F(x) + \lambda^2 \mathcal{K}^2 y^{(0)}(x), \\ y^{(3)}(x) &= F(x) + \lambda \mathcal{K} F(x) + \lambda^2 \mathcal{K}^2 F(x) + \lambda^3 \mathcal{K}^3 y^{(0)}(x) \end{aligned}\right\}. \quad (155)$$

More generally, after the nth substitution we have

$$y^{(n)}(x) = F(x) + \lambda \mathcal{K} F(x) + \lambda^2 \mathcal{K}^2 F(x) + \lambda^3 \mathcal{K}^3 F(x) + \cdots + \lambda^{n-1} \mathcal{K}^{n-1} F(x) + R_n(x), \quad (156)$$

where $R_n(x)$ is defined by

$$R_n(x) = \lambda^n \mathcal{K}^n y^{(0)}(x). \quad (157)$$

Hence, as $n \to \infty$, we are led to the possibility that the desired solution of (147) can be expressed as the infinite series

$$y(x) = F(x) + \sum_{n=1}^{\infty} \lambda^n \mathcal{K}^n F(x), \quad (158)$$

It remains to determine conditions under which the expression $R_n(x)$ tends to zero and under which the formal series (158) actually converges and represents the solution of (147).

Suppose that for all values of x and ξ in the interval (a, b) the kernel $K(x, \xi)$ is smaller in absolute value than a certain fixed constant M:

$$| K(x, \xi) | < M; \tag{159a}$$

that the prescribed function $F(x)$ is also bounded in (a, b):

$$| F(x) | < m; \tag{159b}$$

and that the initial approximation $y^{(0)}(x)$ is likewise bounded in (a, b):

$$| y^{(0)}(x) | < C. \tag{160}$$

These bounds will certainly exist if K, F, and $y^{(0)}$ are *continuous* in the closed interval (a, b).

With the understanding that $b > a$, there then follows

$$| \mathcal{K} y^{(0)}(x) | = | \int_a^b K(x, \xi) y^{(0)}(\xi)\, d\xi | < \int_a^b M C\, d\xi = M(b - a)C.$$

More generally, we find by iteration that

$$| \mathcal{K}^n y^{(0)}(x) | < M^n (b - a)^n C \tag{161}$$

and, similarly,

$$| \mathcal{K}^n F(x) | < M^n (b - a)^n m. \tag{162}$$

Hence, according to (157) and (161), there follows

$$| R_n(x) | < | \lambda |^n M^n (b - a)^n C, \tag{163}$$

and we may deduce that $R_n(x)$ tends to zero with increasing n if

$$| \lambda | < \frac{1}{M(b - a)}. \tag{164}$$

Further, from (162) it follows that the series

$$| F(x) | + \sum_{n=1}^{\infty} | \lambda |^n | \mathcal{K}^n F(x) |$$

is dominated by the constant series

$$m \left[1 + \sum_{n=1}^{\infty} | \lambda |^n M^n (b - a)^n \right].$$

Since this geometric series converges when λ satisfies inequality (164), we may deduce that the series (158) converges absolutely and uniformly* in (a, b) when (164) is satisfied.

It is easily seen, by direct substitution and term-by-term integration, that the series (158) satisfies the integral equation (147) and hence represents the continuous solution of (147) when (164) is satisfied and K is continuous.

The series (158) is a power series in λ. If we recall that the solution to (147) generally fails to exist when λ takes on a characteristic value, we are led to expect that the series solution (158) will cease to converge at least as soon as $|\lambda|$ becomes equal to the absolute value of the smallest characteristic number λ_1. It can be shown that this is the case and, indeed, that *the series* (158) *converges when* $|\lambda| < |\lambda_1|$, *and only then.*

Noticing that the condition (164) may be conservative, in the sense that the series (158) *may* converge even though (164) is *not* satisfied, we are thus led to the useful relation

$$|\lambda_1| \geqq \frac{1}{M(b-a)}, \tag{165}$$

which gives a lower bound for the magnitude of the smallest characteristic number λ_1. A somewhat more involved analysis (see Problem 75) leads to the inequality

$$|\lambda_1| \geqq \frac{1}{\sqrt{\int_a^b \int_a^b [K(x, \xi)]^2 \, dx \, d\xi}} \tag{166}$$

which gives a sharper lower bound.

In the case of the *Volterra* equation,

$$y(x) = F(x) + \lambda \int_a^x K(x, \xi) y(\xi) \, d\xi, \tag{167}$$

with a *variable* upper limit, we define $\mathcal{K}_x$ as the integral operator such that

$$\mathcal{K}_x f(x) \equiv \int_a^x K(x, \xi) f(\xi) \, d\xi, \tag{168}$$

* A series of functions of x which is dominated by a convergent series of positive *constants* independent of x, for all values of x in an interval (a, b), is *uniformly convergent* in (a, b). Such a series of continuous functions represents a continuous function and it can be integrated term by term in (a, b).

and a procedure completely analogous to that used above leads to the formal solution

$$y(x) = F(x) + \sum_{n=1}^{\infty} \lambda^n \mathcal{K}_x{}^n F(x), \tag{169}$$

if the expression

$$R_n(x) = \lambda^n \mathcal{K}_x{}^n y^{(0)}(x) \tag{170}$$

tends to zero as $n \to \infty$.

In this case, if we consider any interval (a, b), where b is *any* number larger than a, and again assume the bounds (159) and (160) over (a, b), we have

$$| \mathcal{K}_x y^{(0)}(x) | = | \int_a^x K(x, \xi) y^{(0)}(\xi)\, d\xi | \leqq MC \int_a^x d\xi = MC(x - a),$$

when $a \leqq x \leqq b$, and hence also

$$| \mathcal{K}_x{}^2 y^{(0)}(x) | \leqq \int_a^x | K(x, \xi) | MC(\xi - a)\, d\xi$$

$$= M^2 C \int_a^x (\xi - a)\, d\xi = \frac{M^2(x - a)^2}{2 \cdot 1} C.$$

Inductive reasoning then leads to the result

$$|R_n(x)| = | \lambda^n \mathcal{K}_x{}^n y^{(0)}(x) | \leqq | \lambda |^n \frac{M^n (x - a)^n}{n!} C$$

$$\leqq | \lambda |^n \frac{M^n (b - a)^n}{n!} C, \tag{171a}$$

for $a \leqq x \leqq b$. In a similar way, it is found that

$$| \lambda^n \mathcal{K}_x{}^n F(x) | \leqq | \lambda |^n \frac{M^n (b - a)^n}{n!} m \tag{171b}$$

for $a \leqq x \leqq b$. But the last member of (171a) tends to zero as $n \to \infty$, and also the series whose nth term is given by the right-hand member of (171b) converges, for *any* finite value of λ. Thus the method of successive substitutions converges to the series (169) and that series converges absolutely and uniformly, for any finite value of λ, in any interval (a, b) for which $b > a$. A similar result follows for any $b < a$. It then follows, by direct substitution, that

the series (169) *converges to the unique continuous solution of the Volterra equation* (167) *for all values of* λ, *in any interval* (a, b) *in which* $K(x, \xi)$ *is continuous.*

We may notice that, as was to be expected, the final solution in each case is independent of the initial approximation $y^{(0)}(x)$. In practice, it is often desirable to merely evaluate the successive terms in the series (158) or (169) by iteration, as is illustrated by an example which follows, rather than to actually pursue the method of successive *substitutions* which motivated (158) and (169). If the latter method is used, the initial approximation

$$y^{(0)}(x) = F(x) \tag{172}$$

is usually a convenient one, unless advance information as to the nature of the required solution is available.

As a very simple illustration of these results, we consider the Fredholm equation

$$y(x) = 1 + \lambda \int_0^1 (1 - 3x\xi)y(\xi)\, d\xi, \tag{173}$$

the solution of which can be obtained readily from the results of Section 4.7 in the form

$$y(x) = \frac{4 + 2\lambda(2 - 3x)}{4 - \lambda^2} \qquad (\lambda \neq \pm 2). \tag{174}$$

The operation $\mathcal{K}f(x)$ is then of the form

$$\mathcal{K}f(x) = \int_0^1 (1 - 3x\xi)f(\xi)d\xi.$$

To obtain the series solution (158), we make the calculations

$$\mathcal{K}F = \int_0^1 (1 - 3x\xi)\, d\xi = 1 - \frac{3}{2}x,$$

$$\mathcal{K}^2F = \int_0^1 (1 - 3x\xi)\left(1 - \frac{3}{2}\xi\right) d\xi = \frac{1}{4},$$

$$\mathcal{K}^3F = \int_0^1 (1 - 3x\xi)\,\frac{1}{4}\, d\xi = \frac{1}{4}\left(1 - \frac{3}{2}x\right),$$

and so forth. From the form of these results the form of the general result is obvious, and (158) becomes

$$y(x) = 1 + \lambda\left(1 - \frac{3}{2}x\right) + \frac{\lambda^2}{4} + \frac{\lambda^3}{4}\left(1 - \frac{3}{2}x\right) + \frac{\lambda^4}{16} + \frac{\lambda^5}{16}\left(1 - \frac{3}{2}x\right) + \cdots. \quad (175a)$$

This result can be expressed in the form

$$y(x) = \left(1 + \frac{\lambda^2}{4} + \frac{\lambda^4}{16} + \cdots\right)\left[1 + \lambda\left(1 - \frac{3}{2}x\right)\right]. \quad (175b)$$

The power series in (175b) is a geometric series, convergent when $|\lambda| < 2$, with the sum $1/(1 - \lambda^2/4)$. Hence (175) *is the power series expansion, valid when and only when* $|\lambda| < 2$, *of the solution*

$$y(x) = \frac{1 + \lambda(1 - \frac{3}{2}x)}{1 - \frac{1}{4}\lambda^2}, \quad (176)$$

which is identical with (174), *and which itself is valid for all values of* λ *except* $\lambda = \pm 2$.

We may notice that consequently the method of successive substitutions *would not converge*, for example, if it were applied to the integral equation

$$y(x) = 1 + 4\int_0^1 (1 - 3x\xi)y(\xi)\,d\xi,$$

while it *would* converge if applied to the same equation with the factor 4 replaced by any number smaller than two in absolute value.

Obviously the method of Section 4.6 is generally to be preferred when the kernel $K(x, \xi)$ is separable, as in the preceding case. It is important to notice that the methods of Section 4.6 express the solution as the ratio of a function of x and λ to a polynomial $\Delta(\lambda)$, valid for all values of λ except the characteristic values for which $\Delta(\lambda) = 0$. The method of successive substitutions, in such cases, leads to the power series expansion of this *ratio*, valid only when $|\lambda|$ is smaller than the magnitude of the smallest characteristic number.

In the more important cases when the kernel is not separable, there exists a method, due to *Fredholm*, which generalizes the procedure of Section 4.6. This method is discussed in Section 4.11.

In the following section the preceding developments are considered from a slightly different viewpoint.

4.10. The Neumann series. With the notation of equation (152),

$$\mathcal{K} f(x) = \int_a^b K(x, \xi) f(\xi)\, d\xi, \tag{177}$$

there follows also

$$\begin{aligned} \mathcal{K}^2 f(x) &= \int_a^b K(x, \xi_1) \mathcal{K} f(\xi_1)\, d\xi_1 \\ &= \int_a^b K(x, \xi_1) \left[\int_a^b K(\xi_1, \xi) f(\xi)\, d\xi \right] d\xi_1 \\ &= \int_a^b \left[\int_a^b K(x, \xi_1) K(\xi_1, \xi)\, d\xi_1 \right] f(\xi)\, d\xi. \end{aligned} \tag{178}$$

If we define the *iterated kernel* $K_2(x, \xi)$ by the relation

$$K_2(x, \xi) = \int_a^b K(x, \xi_1) K(\xi_1, \xi)\, d\xi_1, \tag{179}$$

equation (178) takes the form

$$\mathcal{K}^2 f(x) = \int_a^b K_2(x, \xi) f(\xi)\, d\xi. \tag{180}$$

By repeating this process, it is easily seen that one can write

$$\mathcal{K}^n f(x) = \int_a^b K_n(x, \xi) f(\xi)\, d\xi, \tag{181}$$

where $K_n(x, \xi)$ is the nth iterated kernel, defined by the recurrence formula

$$K_n(x, \xi) = \int_a^b K(x, \xi_1) K_{n-1}(\xi_1, \xi)\, d\xi_1, \tag{182a}$$

for $n = 2, 3, 4, \ldots$, and where we write

$$K_1(x, \xi) \equiv K(x, \xi). \tag{182b}$$

It is not difficult to establish the consequent validity of the relation

$$K_{p+q}(x, \xi) = \int_a^b K_p(x, \xi_1) K_q(\xi_1, \xi)\, d\xi_1, \tag{183}$$

for any positive integers p and q. Further, if $K(x, \xi)$ is bounded in (a, b), in such a way that

$$| K(x, \xi) | < M \tag{184}$$

in (a, b), then it follows easily that also

$$| K_n(x, \xi) | < M^n (b - a)^{n-1} \tag{185}$$

for values of x and ξ in (a, b).

With the notation of (181), the series (158), representing the solution of the equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi \tag{186}$$

for sufficiently small values of $|\lambda|$, takes the form

$$\begin{aligned} y(x) &= F(x) + \sum_{n=1}^{\infty} \lambda^n \int_a^b K_n(x, \xi) F(\xi)\, d\xi \\ &= F(x) + \lambda \int_a^b \Big[\sum_{n=0}^{\infty} \lambda^n K_{n+1}(x, \xi) \Big] F(\xi)\, d\xi, \end{aligned} \tag{187}$$

assuming the legitimacy of interchange of summation and integration. If we introduce the abbreviation

$$\begin{aligned} \Gamma(x, \xi; \lambda) &= \sum_{n=0}^{\infty} \lambda^n K_{n+1}(x, \xi) \\ &= K(x, \xi) + \lambda K_2(x, \xi) + \lambda^2 K_3(x, \xi) + \cdots, \end{aligned} \tag{188}$$

equation (187) takes the form

$$y(x) = F(x) + \lambda \int_a^b \Gamma(x, \xi; \lambda) F(\xi)\, d\xi. \tag{189}$$

The function $\Gamma(x, \xi; \lambda)$ is known as the *reciprocal* or *resolvent kernel* associated with the kernel $K(x, \xi)$ in the interval (a, b). Further, the series (188) [or, in some references, the series (187)] is known as the *Neumann series*. If use is made of (185), it is found that this series converges (absolutely and uniformly) when $|\lambda| < 1/M(b - a)$. A more precise analysis shows indeed that the series converges when $|\lambda| < |\lambda_1|$, where λ_1 is the smallest characteristic number, as was the case in the analogous expansions of the preceding section. In fact, equation (189) is merely an abbreviation for (187), which is equivalent to (158) in virtue of (181).

In practice, unless the solution of (186) is required for several choices of $F(x)$, it may be more convenient to obtain the series (187) or (158) by the iterative methods of the preceding section, than actually to evaluate (188) and insert the result into (189); the net result is, of course, the same in both cases. However, the Neumann series and the resolvent kernel are of importance in theoretical

developments. An interesting result in this connection is obtained if we rewrite (188) in the form

$$\Gamma(x, \xi; \lambda) = K(x, \xi) + \lambda \sum_{n=0}^{\infty} \lambda^n K_{n+2}(x, \xi)$$

$$= K(x, \xi) + \lambda \sum_{n=0}^{\infty} \lambda^n \int_a^b K(x, \xi_1) K_{n+1}(\xi_1, \xi)\, d\xi_1$$

and hence, again referring to (188), deduce the relation

$$\Gamma(x, \xi; \lambda) = K(x, \xi) + \lambda \int_a^b K(x, \xi_1)\Gamma(\xi_1, \xi; \lambda)\, d\xi_1,$$

or, with a change in notation,

$$\Gamma(x, y; \lambda) = K(x, y) + \lambda \int_a^b K(x, \xi)\Gamma(\xi, y; \lambda)\, d\xi. \tag{190}$$

Thus it follows that *the resolvent kernel* Γ, *considered as a function of the two variables* x *and* y *and the parameter* λ, *is the solution of equation* (186) *when the prescribed function* F *is replaced by the kernel* K, *considered as a function of* x *and* y.

In order to illustrate the actual determination of the resolvent kernel in a simple case, we again consider equation (173). With

$$K(x, \xi) = 1 - 3x\xi,$$

there follows

$$K_2(x, \xi) = \int_0^1 (1 - 3x\xi_1)(1 - 3\xi_1\xi)\, d\xi_1 = 1 - \frac{3}{2}(x + \xi) + 3x\xi$$

and, similarly,

$$K_3(x, \xi) = \int_0^1 K(x, \xi_1) K_2(\xi_1, \xi)\, d\xi_1 = \frac{1}{4}(1 - 3x\xi).$$

Since, in this special case, we therefore have $K_3 = K_1/4$, it follows easily that $K_n = K_{n-2}/4$ for $n \geq 3$, and hence we have

$$\Gamma = K_1 + \lambda K_2 + \lambda^2 K_3 + \cdots$$

$$= \left(1 + \frac{\lambda^2}{4} + \frac{\lambda^4}{16} + \cdots\right) K_1 + \lambda \left(1 + \frac{\lambda^2}{4} + \frac{\lambda^4}{16} + \cdots\right) K_2$$

or

$$\Gamma(x, \xi; \lambda) = \frac{1}{1 - \lambda^2/4}\left[(1 + \lambda) - \frac{3}{2}\lambda(x + \xi) - 3(1 - \lambda)x\xi\right] \tag{191}$$

$$(|\lambda| < 2).$$

The introduction of this function into (189), with $F(x) = 1$, leads again to the solution (176).

It is important to notice that the result obtained is correct for all values of λ except $\lambda = \pm 2$. That is, the resolvent kernel is correctly given by (191) for all such values of λ. However, the series involved in the equation preceding (191) converges only when $|\lambda| < 2$. It happens that we are able to sum that series explicitly in the present example, and that the resultant function correctly represents the resolvent kernel for *all* values of λ other than characteristic values.

4.11. Fredholm theory. It is possible to express the resolvent kernel $\Gamma(x, \xi; \lambda)$ as the *ratio* of *two* infinite series of powers of λ, in such a way that *both series converge for all values of* λ. The derivation of the basic equations, due originally to Fredholm, involves considerable algebraic manipulation and is not considered here.

If the resolvent kernel is expressed as the ratio

$$\Gamma(x, \xi; \lambda) = \frac{D(x, \xi; \lambda)}{\Delta(\lambda)}, \tag{192}$$

where

$$D(x, \xi; \lambda) = K(x, \xi) - \frac{\lambda}{1!} D_1(x, \xi) + \frac{\lambda^2}{2!} D_2(x, \xi) - \cdots \tag{193}$$

and

$$\Delta(\lambda) = 1 - \frac{\lambda}{1!} C_1 + \frac{\lambda^2}{2!} C_2 - \cdots, \tag{194}$$

it is found that the coefficients C_n and the functions $D_n(x, \xi)$ can be determined successively by the following sequence of calculations:

$$C_1 = \int_a^b K(x, x)\, dx, \quad D_1(x, \xi) = C_1 K(x, \xi) - \int_a^b K(x, \xi_1) K(\xi_1, \xi)\, d\xi_1;$$

$$C_2 = \int_a^b D_1(x, x)\, dx, \quad D_2(x, \xi) = C_2 K(x, \xi) - 2 \int_a^b K(x, \xi_1) D_1(\xi_1, \xi)\, d\xi_1;$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$C_n = \int_a^b D_{n-1}(x, x)\, dx,$$

$$D_n(x, \xi) = C_n K(x, \xi) - n \int_a^b K(x, \xi_1) D_{n-1}(\xi_1, \xi)\, d\xi_1. \tag{195}$$

The solution of the equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi \tag{196}$$

is then obtained by introducing (192) into (189), in the form

$$y(x) = F(x) + \lambda \frac{\int_a^b D(x, \xi; \lambda)F(\xi)\, d\xi}{\Delta(\lambda)}. \tag{197}$$

In those cases when $K(x, \xi)$ is *separable,* this result is identical in form with the solution obtained by the methods of Section 4.6. The series (193) and (194) then each involve only a finite number of terms.

More generally, if the ratio of the two power series involved in (197) were expressed as a single power series in λ (by division or otherwise) the result would reduce to the series (158). However, the result of this operation would converge only for small values of $|\lambda|$ (when $|\lambda| < |\lambda_1|$), whereas the separate series expansions of the numerator and denominator in the last term of (197) each converge for *all* values of λ.

The denominator $\Delta(\lambda)$ vanishes only when λ takes on a characteristic value, in which case either no solution or infinitely many solutions of (196) exist, and (197) is no longer valid.

Despite the generality of the solution just described, the practical usefulness of the result is limited by the fact that the relevant calculations usually involve a prohibitive amount of labor unless $K(x, \xi)$ is separable (and hence the simpler methods of Section 4.6 are usually preferable). Nevertheless, the rigorous development of the underlying theory has led to valuable information concerning existence and uniqueness of solutions of (196). In the following paragraph we summarize certain known facts which generalize results already obtained in the special cases when the kernel is either separable or symmetric (see Reference 2).

The equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi)y(\xi)\, d\xi, \tag{198}$$

where $F(x)$ and $K(x, \xi)$ are continuous in (a, b), possesses one and only one continuous solution for any fixed value of λ which is not a characteristic value. If λ_c is a characteristic number of multiplicity r, that is, if the associated homogeneous equation

$$y(x) = \lambda_c \int_a^b K(x, \xi)y(\xi)\, d\xi \tag{199}$$

possesses r linearly independent nontrivial solutions $\phi_1, \phi_2, \ldots, \phi_r$, *then the associated transposed homogeneous equation*

$$y(x) = \lambda_c \int_a^b K(\xi, x) y(\xi)\, d\xi \tag{200}$$

also possesses r linearly independent nontrivial solutions $\psi_1, \psi_2, \ldots, \psi_r$. *In this exceptional case,* (198) *possesses no solution unless* $F(x)$ *is orthogonal to each of the characteristic functions* $\psi_1, \psi_2, \ldots, \psi_r$,

$$\int_a^b F(x)\psi_k(x)\, dx = 0 \qquad (k = 1, 2, \cdots, r). \tag{201}$$

Finally, if $\lambda = \lambda_c$ *and* (201) *is satisfied, then the solution of* (198) *is determinate only within an additive linear combination* $c_1\phi_1 + c_2\phi_2 + \cdots + c_r\phi_r$, *where the r constants* c_n *are arbitrary.*

When $K(x, \xi)$ is *symmetric,* equations (199) and (200) are identical and the preceding results reduce to those given by the Hilbert-Schmidt theory of Section 4.8.

It is useful to notice the complete analogy between the preceding results and the corresponding results relevant to existence and uniqueness of solutions of sets of n linear algebraic equations in n unknowns (see Section 1.10). Indeed, the plausibility of these statements was first suggested by the possibility of considering a Fredholm integral equation as the limit of such a set of equations as the number n of equations and unknowns becomes infinite.

A *Volterra* integral equation, of the form

$$y(x) = F(x) + \lambda \int_a^x K(x, \xi) y(\xi)\, d\xi, \tag{202}$$

can be considered as a special form of a Fredholm equation, with a kernel given by the expressions

$$\bar{K}(x, \xi) = \begin{cases} 0 & \text{when} \quad x < \xi, \\ K(x, \xi) & \text{when} \quad x \geqq \xi. \end{cases} \tag{203}$$

However, unless $K(x, x) = 0$, the modified kernel $\bar{K}(x, \xi)$ is *discontinuous* when $x = \xi$. The results of Section 4.9 show that *if* $F(x)$ *and* $K(x, \xi)$ *are continuous the Volterra equation* (202) *possesses one and only one continuous solution, and that solution is given by the series* (169) *for any value of* λ. In particular, when $F(x) \equiv 0$ the only possible continuous solution of (202) is then the *trivial* solution $y(x) \equiv 0$.

4.12. Singular integral equations. An integral equation in which the range of integration is infinite, or in which the kernel $K(x, \xi)$ is discontinuous, is called a *singular* integral equation. Thus, in illustration, the equations

$$F(x) = \int_0^\infty \sin(x\xi)\, y(\xi)\, d\xi, \tag{204}$$

$$F(x) = \int_0^\infty e^{-x\xi}\, y(\xi)\, d\xi, \tag{205}$$

and

$$F(x) = \int_0^x \frac{y(\xi)}{\sqrt{x-\xi}}\, d\xi \tag{206}$$

are all singular integral equations of the first kind. Except in certain special cases, theoretical information concerning singular equations is not yet present in the literature. As will be seen, such equations may possess very unusual properties. The three preceding examples were chosen here because of the fact that they have been studied rather extensively.

The function $F(x)$ defined by the right-hand member of (204) may be recognized as the *Fourier sine transform* of $y(x)$. If $F(x)$ is piecewise differentiable when $x > 0$, and if $\int_0^\infty |F(x)|\, dx$ exists, then it is known that equation (204) can be inverted uniquely in the form

$$y(x) = \frac{2}{\pi}\int_0^\infty \sin(x\xi)\, F(\xi)\, d\xi \qquad (x > 0). \tag{207}$$

This result leads to an interesting property of the homogeneous integral equation

$$y(x) = \lambda \int_0^\infty \sin(x\xi)\, y(\xi)\, d\xi, \tag{208}$$

associated with (204), and obtained from (204) by replacing $F(x)$ by $y(x)/\lambda$, since the corresponding inversion of (208) is then of the form

$$y(x) = \frac{2}{\pi\lambda}\int_0^\infty \sin(x\xi)\, y(\xi)\, d\xi. \tag{209}$$

Unless $y(x) \equiv 0$, equations (208) and (209) are compatible only if

$$\lambda = \pm\sqrt{\frac{2}{\pi}}. \tag{210}$$

Thus we conclude that if (208) possesses characteristic numbers, those numbers can only be $\lambda = \sqrt{2/\pi}$ and $\lambda = -\sqrt{2/\pi}$.

That these values of λ *are* actually characteristic values follows from the verifiable relation

$$\sqrt{\frac{\pi}{2}}\, e^{-ax} \pm \frac{x}{a^2 + x^2} = \pm \sqrt{\frac{2}{\pi}} \int_0^\infty \sin(x\xi) \left[\sqrt{\frac{\pi}{2}}\, e^{-a\xi} \pm \frac{\xi}{a^2 + \xi^2} \right] d\xi, \tag{211}$$

when $x > 0$ and $a > 0$. This equation states that when $\lambda = \sqrt{2/\pi}$ equation (208) is satisfied by the function

$$y_1(x) = \sqrt{\frac{\pi}{2}}\, e^{-ax} + \frac{x}{a^2 + x^2} \qquad (x > 0), \tag{212}$$

for any positive constant value of a, whereas when $\lambda = -\sqrt{2/\pi}$ the function

$$y_2(x) = \sqrt{\frac{\pi}{2}}\, e^{-ax} - \frac{x}{a^2 + x^2} \qquad (x > 0) \tag{213}$$

is a solution for any positive value of a. Thus, the two characteristic values of λ are here of *infinite multiplicity;* that is, each value corresponds to infinitely many independent characteristic functions. This situation is in contrast with the fact that any characteristic number of a *nonsingular* Fredholm equation corresponds only to a *finite* number of independent characteristic functions.

The function $F(x)$ defined by the right-hand member of equation (205) is the *Laplace transform* of the function $y(x)$. It is known that, while not all functions can be Laplace transforms of other functions, there cannot be two distinct functions with the same transform. Thus for a prescribed function $F(x)$, if (205) possesses a solution then that solution is unique, and it can be determined by known methods. In order to establish an unusual property of the associated homogeneous equation

$$y(x) = \lambda \int_0^\infty e^{-x\xi}\, y(\xi)\, d\xi \qquad (x > 0), \tag{214}$$

we notice that, in accordance with the definition of the *Gamma function,* we have the relation

$$\int_0^\infty e^{-x\xi}\, \xi^{a-1}\, d\xi = \Gamma(a) x^{-a} \qquad (a > 0). \tag{215}$$

The result of replacing a by $1 - a$ is then of the form

$$\int_0^\infty e^{-x\xi}\,\xi^{-a}\,d\xi = \Gamma(1-a)x^{a-1} \qquad (a < 1). \tag{216}$$

If (215) is divided by $\sqrt{\Gamma(a)}$, and (216) is divided by $\sqrt{\Gamma(1-a)}$, and if the resultant equations are added to each other, the truth of the equation

$$\int_0^\infty e^{-x\xi}[\sqrt{\Gamma(1-a)}\,\xi^{a-1} + \sqrt{\Gamma(a)}\,\xi^{-a}]\,d\xi$$
$$= \sqrt{\Gamma(a)\Gamma(1-a)}\,[\sqrt{\Gamma(1-a)}\,x^{a-1} + \sqrt{\Gamma(a)}\,x^{-a}]$$
$$(0 < a < 1) \tag{217}$$

is established. This equation is identified with (214) by writing

$$\lambda = \frac{1}{\sqrt{\Gamma(a)\Gamma(1-a)}} \qquad (0 < a < 1) \tag{218}$$

and

$$y(x) = \sqrt{\Gamma(1-a)}\,x^{a-1} + \sqrt{\Gamma(a)}\,x^{-a} \qquad (x > 0). \tag{219}$$

It thus follows that for any value of the parameter a such that $0 < a < 1$ a value of λ is determined by (218), corresponding to which (214) possesses a nontrivial solution specified by (219). In consequence of the identity

$$\Gamma(a)\Gamma(1-a) = \frac{\pi}{\sin \pi a} \qquad (0 < a < 1),$$

equation (218) can be written in the form

$$\lambda = \sqrt{\frac{\sin \pi a}{\pi}} \qquad (0 < a < 1), \tag{220}$$

from which it follows that *all values of* λ *in the interval* $0 < \lambda \leqq 1/\sqrt{\pi}$ *are characteristic values for the singular integral equation* (214).

This situation is in contrast with the fact that the characteristic values of λ for a *nonsingular* equation are *discretely* distributed, and cannot constitute a *continuous* "spectrum."

It can be shown further that all values of λ in the interval $-1/\sqrt{\pi} \leqq \lambda < 0$ are also characteristic values for equation (214).

Other singular Fredholm equations may possess only discretely

distributed characteristic numbers, or they may possess both a discrete and a continuous spectrum of characteristic numbers.

Equation (206), in which the range of integration is finite but the kernel is unbounded, is considered in the following section.

4.13. Special devices. In this section we present techniques which are useful in dealing with certain special types of integral equations.

1. *Transforms.* If a relationship of the form

$$y(x) = \int_a^b \int_a^b \Gamma(x, \xi_1) K(\xi_1, \xi) y(\xi)\, d\xi\, d\xi_1 \tag{221}$$

is known to be valid (for a suitably restricted class of functions y) and if the double integral can be evaluated as an iterated integral, then it follows that if

$$F(x) = \int_a^b K(x, \xi) y(\xi)\, d\xi \tag{222}$$

we have also

$$y(x) = \int_a^b \Gamma(x, \xi) F(\xi)\, d\xi. \tag{223}$$

Thus, if (222) is considered as an integral equation in y, a solution is given by (223), whereas if (223) is considered as an integral equation in F a solution is given by (222). It is conventional to refer to one of the functions as the *transform* of the second function, and to the second function as an *inverse transform* of the first. The correspondence may or may not be unique. Thus, for example, the Fourier sine-integral formula

$$y(x) = \frac{2}{\pi} \int_0^\infty \int_0^\infty \sin(x\xi_1) \sin(\xi_1 \xi)\, y(\xi)\, d\xi\, d\xi_1$$

leads to the reciprocal relations (204) and (207).

2. *The convolution.* The function defined by the integral

$$\int_0^x u(x - \xi) v(\xi)\, d\xi \tag{224}$$

is known as the *convolution* of $u(x)$ and $v(x)$. The known fact that *the Laplace transform of the convolution of u and v is equal to the product of the transforms of u and v* permits the reduction of the problem of solving the special Volterra equation

$$y(x) = F(x) + \int_0^x K(x - \xi) y(\xi)\, d\xi \tag{225}$$

to the problem of determining an inverse Laplace transform. If we denote the Laplace transform of a function $f(x)$ by $\mathfrak{L}\, f$, the result of taking the transforms of the equal members of (225) takes the form

$$\mathfrak{L}\, y(x) = \mathfrak{L}\, F(x) + \mathfrak{L}\, K(x)\, \mathfrak{L}\, y(x),$$

and hence there follows

$$\mathfrak{L}\, y(x) = \frac{\mathfrak{L}\, F(x)}{1 - \mathfrak{L}\, K(x)}. \tag{226}$$

The right-hand member of (226) is calculable, and it remains only to determine (by use of tables or otherwise) its inverse transform. Equations of the form (225) occur rather frequently in practice.

3. *Volterra equations of the first kind.* It is often possible to reduce an integral equation of the form

$$F(x) = \int_0^x K(x, \xi) y(\xi)\, d\xi \tag{227}$$

to an equation of the *second* kind. Such a reduction is desirable since the method of successive substitutions is then applicable. Under the assumption that the kernel is continuously differentiable when $\xi \leqq x$, two different procedures are available. First, by differentiating the equal members of (227) we obtain the relation

$$F'(x) = K(x, x) y(x) + \int_0^x \frac{\partial K(x, \xi)}{\partial x}\, y(\xi)\, d\xi.$$

If $K(x, x)$ is never zero, this equation can be put in the form

$$y(x) = \bar{F}(x) + \int_0^x \bar{K}(x, \xi) y(\xi)\, d\xi, \tag{228}$$

where

$$\bar{F}(x) = \frac{F'(x)}{K(x, x)}, \quad \bar{K}(x, \xi) = -\frac{1}{K(x, x)} \frac{\partial K(x, \xi)}{\partial x}. \tag{229}$$

Alternatively, if we define the function

$$Y(x) = \int_0^x y(\xi)\, d\xi, \tag{230}$$

equation (227) takes the form

$$F(x) = \int_0^x K(x, \xi) Y'(\xi)\, d\xi$$

and an integration by parts leads to the relation

$$F(x) = K(x, x)Y(x) - \int_0^x \frac{\partial K(x, \xi)}{\partial \xi} Y(\xi)\, d\xi.$$

If $K(x, x) \neq 0$, we may rewrite this equation in the form

$$Y(x) = \bar{F}(x) + \int_0^x \bar{K}(x, \xi) Y(\xi)\, d\xi, \tag{231}$$

where

$$\bar{F}(x) = \frac{F(x)}{K(x, x)}, \quad \bar{K}(x, \xi) = \frac{1}{K(x, x)} \frac{\partial K(x, \xi)}{\partial \xi}. \tag{232}$$

The solution of (227) is then related to the solution of (231) by the equation $y(x) = Y'(x)$.

4. *Abel's equation.* The Volterra equation

$$F(x) = \int_0^x \frac{y(\xi)}{\sqrt{x - \xi}}\, d\xi \tag{233}$$

is known as *Abel's integral equation.* It can be solved, under appropriate restrictions on the prescribed function F, by an indirect method in which we divide both sides of (233) by $\sqrt{s - x}$, where s is a parameter, and integrate the results with respect to x over $(0, s)$. This procedure leads to the equation

$$\int_0^s \frac{F(x)}{\sqrt{s - x}}\, dx = \int_0^s \left\{ \int_0^x \frac{y(\xi)}{\sqrt{x - \xi}}\, d\xi \right\} \frac{dx}{\sqrt{s - x}}. \tag{234a}$$

If the order of integration in the right-hand member is inverted, and the limits of integration are modified accordingly, this equation becomes

$$\int_0^s \frac{F(x)}{\sqrt{s - x}}\, dx = \int_0^s \left\{ \int_\xi^s \frac{dx}{\sqrt{(x - \xi)(s - x)}} \right\} y(\xi)\, d\xi. \tag{234b}$$

The success of this special method depends upon the fact that the inner integral on the right can be evaluated by elementary methods* to give the *constant* value

$$\int_\xi^s \frac{dx}{\sqrt{(x - \xi)(s - x)}} = \pi.$$

* With $x = (s - \xi)t + \xi$, this integral takes the form $\int_0^1 dt/\sqrt{t(1 - t)}$.

Hence (234b) is equivalent to the relation

$$\int_0^s y(\xi)\,d\xi = \frac{1}{\pi}\int_0^s \frac{F(x)}{\sqrt{s-x}}\,dx$$

or, with a more convenient notation,

$$\int_0^x y(\xi)\,d\xi = \frac{1}{\pi}\int_0^x \frac{F(\xi)}{\sqrt{x-\xi}}\,d\xi. \tag{235}$$

By differentiating this relation, we then obtain the desired solution

$$y(x) = \frac{1}{\pi}\frac{d}{dx}\int_0^x \frac{F(\xi)}{\sqrt{x-\xi}}\,d\xi. \tag{236}$$

Unless F is prescribed in such a way that the right-hand member of (236) exists and is continuous, the equation (233) does not possess a continuous solution. A more direct derivation of (236) can be accomplished by the use of Laplace transforms (see Problem 62).

It is of some interest to consider the mechanical problem which led Abel to consider this equation. Suppose that a particle of mass m starts from rest at the time $t = 0$, and slides to the ground along a smooth curve in a vertical plane under the action of gravity. If the initial point is at height x above the ground and if the height is ξ at time t, then the speed at time t is given by $\sqrt{2g(x-\xi)}$, regardless of the shape of the curve. However, the *time of descent* will depend upon this shape. If distance along the curve from the terminal point at time t is denoted by $s(\xi)$, there follows

$$\frac{ds}{dt} = -\sqrt{2g(x-\xi)}$$

and hence the time of descent is given by

$$T = \frac{1}{\sqrt{2g}}\int_0^x \frac{s'(\xi)}{\sqrt{x-\xi}}\,d\xi. \tag{237}$$

For a specified curve, this relation permits the calculation of T as a function of the initial height x. Abel considered the converse problem, in which the time of descent is specified as a function of x, and the curve is to be determined. Equation (237) reduces to (233) if we write $F(x) = \sqrt{2g}\,T(x)$ and $y(x) = s'(x)$.

The more general equation

$$F(x) = \int_0^x \frac{y(\xi)}{(x-\xi)^\alpha}\, d\xi \qquad (0 < \alpha < 1), \tag{238}$$

which was also considered by Abel, can be solved in a similar way (see Problem 60).

4.14. Iterative approximations to characteristic functions. Methods analogous to those given in Section 1.23, for the approximate determination of characteristic numbers and functions, can be applied to the homogeneous equation

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi, \tag{239}$$

where $K(x, \xi)$ is symmetric. If this equation is written in the operational form $y = \lambda \mathcal{K} y$, it appears that here the parameter λ is analogous to the inverse parameter $1/\lambda$ in the matrix equation $\mathbf{a}\,\mathbf{x} = \lambda\,\mathbf{x}$ of Section 1.23. Consequently, whereas the methods of that section tend to determine the *largest* characteristic value of λ, the analogous procedures in the present case tend to determine the characteristic number with *smallest* absolute value. Except in those cases when the kernel is separable, the integral equation (239), with a symmetric kernel, possesses an infinite set of characteristic numbers (see Problem 42), and it is known that this set does not possess a largest member, in terms of absolute value.

In order to approximate the fundamental characteristic function we choose an initial approximation $y^{(1)}(x)$ and calculate a corresponding approximation from the equation

$$y(x) = \lambda \int_a^b K(x, \xi) y^{(1)}(\xi)\, d\xi \equiv \lambda f^{(1)}(x). \tag{240}$$

A convenient multiple of $f^{(1)}(x)$ is then taken as a new approximation $y^{(2)}(x)$, and the process is repeated until satisfactory convergence is indicated. In those cases when the kernel $K(x, \xi)$ is continuous and symmetric, it can be shown that the successive approximations $y^{(n)}(x)$ tend to a characteristic function $y_1(x)$ corresponding to the characteristic number λ_1 with smallest absolute value unless it happens that the initial approximation is orthogonal to that function. Further, the ratio of the input $y^{(n)}(x)$ to the output $f^{(n)}(x)$ in the nth cycle tends to λ_1 as n increases. The proof is completely analogous to that given in Section 1.24.

Estimates of the value λ_1 in the nth cycle are afforded by use of any of the following formulas (see Problem 78):

$$\lambda_1 \approx \frac{\int_a^b y^{(n)}(x)\,dx}{\int_a^b f^{(n)}(x)\,dx} \quad \text{or} \quad \frac{\int_a^b [y^{(n)}(x)]^2\,dx}{\int_a^b y^{(n)}(x) f^{(n)}(x)\,dx} \quad \text{or} \quad \frac{\int_a^b y^{(n)}(x) f^{(n)}(x)\,dx}{\int_a^b [f^{(n)}(x)]^2\,dx}. \tag{241a,b,c}$$

The approximation given by the ratio (241b) is in general more nearly accurate than that given by (241a), whereas (241c) is in general still more efficient. [Compare equations (232a,b) of Section 1.24.]

If the characteristic function $y_1(x)$ were known *exactly*, and the next higher characteristic quantities were required, a sequence of approximations tending toward $y_2(x)$ would be obtained by starting with an initial approximation which is orthogonal to $y_1(x)$ over (a, b). That is, we would choose a convenient function $F(x)$ and take

$$y^{(1)}(x) = F(x) - c\,y_1(x), \tag{242}$$

where c is determined by the equation

$$c \int_a^b [y_1(x)]^2\,dx = \int_a^b F(x) y_1(x)\,dx, \tag{243}$$

so that "the y_1-component of F is subtracted from F." Since $y_1(x)$ is not known exactly, its approximation must be substituted for it in (242) and (243). Convergence to λ_2 and $y_2(x)$ will generally then obtain if before each cycle the initial approximation is (approximately) "cleared" of $y_1(x)$ before substitution into (239). Once λ_2 and $y_2(x)$ are satisfactorily approximated, the successive initial approximations in the next stage must be cleared of both $y_1(x)$ and $y_2(x)$, and the process may be continued indefinitely. However, it is found that unless the fundamental characteristic functions are determined to a high degree of accuracy, the accuracy and rate of convergence of following calculations may be seriously impaired.

Information concerning the convergence of the preceding process in the more general case of a nonsymmetric kernel is limited. However, in the case of the equation

$$y(x) = \lambda \int_a^b G(x, \xi) r(\xi) y(\xi)\,d\xi, \tag{244}$$

where G is symmetric and $r(x)$ is positive in (a, b), the basic theory differs from that associated with (239) only in that the characteristic functions corresponding to distinct characteristic numbers are orthogonal *with respect to the weighting function* $r(x)$ (see Problem 38). The iterative procedure outlined above is accordingly modified only to the extent that the weighting function r is to be introduced into the integrals appearing in equations (239), (240), (241), and (243). [Compare equations (256a,b) of Section 1.25.]

4.15. Approximation of Fredholm equations by sets of algebraic equations. It has already been pointed out that a Fredholm integral equation can be considered as the limit of a set of n algebraic equations, as the number of equations increases without limit. Use can be made of this fact to obtain approximate solutions of such integral equations.

For this purpose, we recall first that a definite integral of the form

$$I = \int_a^b f(\xi)\, d\xi \tag{245}$$

is defined as a limit of the form

$$I = \lim_{n \to \infty} \sum_{k=1}^{n} f(x_k)(\Delta x)_k, \tag{246}$$

where the interval (a, b) is divided into n subintervals of lengths $(\Delta x)_1, \ldots, (\Delta x)_n$, and x_k is a point of the kth subinterval. An approximate evaluation can be obtained by not proceeding to the limit, and hence by expressing I approximately as the weighted sum of the ordinates $f(x_k)$ at n conveniently chosen points $x_1, x_2 \ldots, x_n$ of the interval (a, b):

$$I \approx \sum_{k=1}^{n} D_k f(x_k), \tag{247}$$

where D_k is the "weighting coefficient" associated with the point x_k.

The coefficient D_k may be identified with the length $(\Delta x)_k$ of the subinterval associated with the point x_k, as is suggested by (246). However, when the points $x_1, x_2, \ldots, x_n$ are equally spaced, more nearly accurate approximations are generally obtained by choosing these coefficients in accordance with a formula such as the *trapezoidal rule* or *Simpson's rule*. More elaborate formulas are also

available in the literature. If the points x_1 and x_n are identified with the end points $x = a$ and $x = b$, respectively, and a uniform spacing h is chosen, so that

$$(n - 1)h = b - a, \tag{248}$$

we recall that the *trapezoidal rule* gives

$$\{D_1, D_2, D_3, D_4, \cdots, D_{n-2}, D_{n-1}, D_n\} = h\{\tfrac{1}{2}, 1, 1, 1, \cdots, 1, 1, \tfrac{1}{2}\}. \tag{249}$$

According to *Simpson's rule*, which is applicable only if n is *odd*, the weighting coefficients are of the form

$$\{D_1, D_2, D_3, D_4, \cdots, D_{n-2}, D_{n-1}, D_n\} = \frac{h}{3}\{1, 4, 2, 4, \cdots, 2, 4, 1\}, \tag{250}$$

when $n = 5, 7, 9, \ldots$, and are of the form

$$\{D_1, D_2, D_3\} = \frac{h}{3}\{1, 4, 1\} \tag{250a}$$

in the special case when $n = 3$.*

In the same way, the integral equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi)y(\xi)\, d\xi \tag{251}$$

can be approximated in the form

$$y(x) \approx F(x) + \lambda \sum_{k=1}^{n} D_k K(x, x_k)y(x_k), \tag{252}$$

where the points x_k are n conveniently chosen points in the interval (a, b), and the constants D_k are corresponding weighting coefficients. If we now require that the two members of (252) be equal *at each of the n chosen points*, we obtain the n linear equations

$$y(x_i) = F(x_i) + \lambda \sum_{k=1}^{n} D_k K(x_i, x_k)y(x_k) \qquad (i = 1, 2, \cdots, n), \tag{253}$$

* It is recalled that the trapezoidal rule results from approximating the integrand by joining the ordinates at successive division points by straight lines; Simpson's rule results from passing *parabolas* through successive sets of *three* ordinates, and is generally more nearly accurate.

in the n unknowns $y(x_1), \ldots, y(x_n)$ which specify approximate values of the unknown function $y(x)$ at the n points.

If we introduce the abbreviations

$$y_i = y(x_i), \qquad F_i = F(x_i), \qquad K_{ij} = K(x_i, x_j), \tag{254}$$

where K_{ij} is hence the value of $K(x, \xi)$ when $x = x_i$ and $\xi = x_j$, this set of equations can be written in the form

$$y_i = F_i + \lambda \sum_{k=1}^{n} K_{ik} D_k y_k \qquad (i = 1, 2, \cdots, n), \tag{255}$$

Thus, if we consider the numbers y_i and F_i as components of the vectors $\mathbf{y}$ and $\mathbf{F}$, and define the matrix $\mathbf{K} = [K_{ij}]$, the set of equations (253) can be written concisely in the form

$$\mathbf{y} = \mathbf{F} + \lambda \mathbf{K} \mathbf{D} \mathbf{y}.$$

Here $\mathbf{D} = [D_i \delta_{ij}]$ is a diagonal matrix, and the product $\mathbf{K} \mathbf{D}$ is the matrix obtained by multiplying successive *columns* of $\mathbf{K}$ by successive weighting coefficients. Hence the required set of equations is of the form

$$\mathbf{a} \mathbf{y} = \mathbf{F} \quad \text{where} \quad \mathbf{a} = \mathbf{I} - \lambda \mathbf{K} \mathbf{D}, \tag{256}$$

and where $\mathbf{I}$ is the unit matrix of order n.

To illustrate the use of this approximate procedure, we apply it to the solution of the integral equation

$$y(x) = x + \int_0^1 K(x, \xi) y(\xi)\, d\xi, \tag{257}$$

where the kernel is of the form defined by (23),

$$K(x, \xi) = \begin{cases} x(1 - \xi) & \text{when} \quad x < \xi, \\ \xi(1 - x) & \text{when} \quad x > \xi. \end{cases} \tag{258}$$

In this particular example, the integral equation can be reduced to the differential equation $d^2y/dx^2 + y = 0$ with the end conditions $y(0) = 0$, $y(1) = 1$, so that the *exact* solution is obtainable in the form

$$y(x) = \frac{\sin x}{\sin 1}. \tag{259}$$

For simplicity, we take $n = 5$ equally spaced points, so that

$$x_1 = 0, \quad x_2 = \tfrac{1}{4}, \quad x_3 = \tfrac{1}{2}, \quad x_4 = \tfrac{3}{4}, \quad x_5 = 1. \tag{260}$$

The corresponding matrix $\mathbf{K}$ is then easily determined in the form

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{9}{16} & \frac{1}{8} & \frac{1}{16} & 0 \\ 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 \\ 0 & \frac{1}{16} & \frac{1}{8} & \frac{3}{16} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{261}$$

If the weighting coefficients of the trapezoidal rule are used with $h = \frac{1}{4}$, the matrix of coefficients of the linear equations corresponding to (256) with $\lambda = 1$ is then obtained, in the present special case, in the form $\mathbf{I} - \frac{1}{4}\mathbf{K}$. The required equations then follow:

$$\left.\begin{aligned} y_1 &= 0, \\ \tfrac{61}{64}y_2 - \tfrac{1}{32}y_3 - \tfrac{1}{64}y_4 &= \tfrac{1}{4}, \\ -\tfrac{1}{32}y_2 + \tfrac{15}{16}y_3 - \tfrac{1}{32}y_4 &= \tfrac{1}{2}, \\ -\tfrac{1}{64}y_2 - \tfrac{1}{32}y_3 + \tfrac{61}{64}y_4 &= \tfrac{3}{4}, \\ y_5 &= 1 \end{aligned}\right\}. \tag{262}$$

The solution of this set of equations, when the results are rounded off to four decimal places, is obtained as follows:

$$y_1 = 0, \quad y_2 = 0.2943, \quad y_3 = 0.5702, \quad y_4 = 0.8104, \quad y_5 = 1. \tag{263}$$

These approximate values of the solution $y(x)$ at the points $x = 0$, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$, and 1 may be compared with the true values which are obtained from (259) as follows:

$$y_1 = 0, \quad y_2 = 0.2940, \quad y_3 = 0.5697, \quad y_4 = 0.8100, \quad y_5 = 1. \tag{264}$$

Because of the presence of a *corner* in the graph of the integrand in (257), when $x = \xi$, it happens that in this case the use of Simpson's rule is found to give less nearly accurate results.

The preceding method can clearly be applied equally well to the approximate solution of integral equations of the *first* kind, and to the treatment of *characteristic-value* problems. In the latter case, a corresponding problem of the type considered in Chapter 1 is obtained, and the iterative methods developed in that chapter are applicable.

It is important to notice that the present method is particularly useful when the kernel $K(x, \xi)$ is not given analytically, but is specified by empirical data. In this case, the matrix K is precisely a table of values of the empirical influence function. An obvious disadvantage of the method consists in the fact that the approximate solution is obtained only for the n points $x_1, \ldots, x_n$, and must be determined at intermediate points by interpolation or by merely plotting the calculated ordinates and joining them by a smooth curve, or by evaluating the right-hand member of (252).

4.16. Approximate methods of undetermined coefficients. Other numerical methods for obtaining approximate solutions of integral equations also generally consist in reducing the problem to the consideration of a finite set of algebraic equations. In particular, the solution of the equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi \tag{265}$$

may be approximated by a linear combination of n suitably chosen functions $\phi_1, \phi_2, \ldots, \phi_n$, of the form

$$y(x) \approx \sum_{k=1}^{n} A_k \phi_k(x), \tag{266}$$

where the n constants of combination are to be determined in such a way that (265) is satisfied as nearly as possible (in some sense) by (266) over the interval (a, b).

The requirement that (266) approximately satisfy (265) takes the form

$$\sum_{k=1}^{n} A_k \phi_k(x) \approx F(x) + \lambda \sum_{k=1}^{n} A_k \int_a^b K(x, \xi) \phi_k(\xi)\, d\xi \qquad (a \leqq x \leqq b). \tag{267}$$

With the convenient abbreviation

$$\Phi_k(x) = \int_a^b K(x, \xi) \phi_k(\xi)\, d\xi, \tag{268}$$

this condition becomes merely

$$\sum_{k=1}^{n} A_k [\phi_k(x) - \lambda \Phi_k(x)] \approx F(x) \qquad (a \leqq x \leqq b). \tag{269}$$

The coefficients $A_1, \ldots, A_n$ are then to be determined by a set of n conditions which tends to reduce the two members of (269) to an equality over the interval (a, b). Several procedures which lead to such sets of conditions are outlined and illustrated in the sections which follow.

In practical cases, advance information concerning the nature of the behavior of the unknown function $y(x)$ is frequently at hand, and the choice of the approximating functions is motivated by this knowledge. It is frequently convenient to take the approximation in the form of a polynomial of degree n, so that $\phi_k(x)$ is identified with x^k. However, if it happens that one or more of the end values $y(a)$ and $y(b)$ is known in advance (or obtainable by inspection from the integral equation, as indeed was the case in the example of the preceding section), it may be desirable to take the assumed approximation in the form

$$y(x) \approx \phi_0(x) + \sum_{k=1}^{n} A_k \phi_k(x),$$

where ϕ_0 is chosen in such a way that it assumes the known end values, and the remaining ϕ's are made to *vanish* at the corresponding end or ends of the interval.

The dependability of the approximation obtained can be judged to some extent by comparing the resultant left-hand member of (269) with the right-hand member. It should be pointed out, however, that situations unfortunately exist in which a *large* change in the function $y(x)$ may correspond to a *small* change in the function

$$y(x) - \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi.$$

In such cases, it may happen that the integral equation is very nearly satisfied over the interval (a, b) by an "approximation" $\bar{y}(x)$, in the sense that the difference between the two members of (269) is then everywhere small relative to either of those members, but nevertheless $\bar{y}(x)$ may differ appreciably from the exact solution $y(x)$ over that interval.

A somewhat more satisfactory estimate of dependability is obtained by comparing the result of an n-term approximation with the result of an $(n + 1)$-term approximation, where the constants of combination are determined by the same technique in both cases.

4.17. The method of collocation. If we introduce the abbreviation

$$f_k(x) = \phi_k(x) - \lambda \Phi_k(x) \equiv \phi_k(x) - \lambda \int_a^b K(x, \xi)\phi_k(\xi)\, d\xi, \quad (270)$$

the requirement that (266) approximately satisfy (265) can be expressed in the form

$$\sum_{k=1}^{n} A_k f_k(x) \approx F(x) \qquad (a \leqq x \leqq b). \quad (271)$$

A set of n conditions for the determination of the n constants of combination can be obtained most simply by requiring that (271) be an *equality* at n distinct points in the interval (a, b). If we denote these points by x_i $(i = 1, 2, \ldots, n)$, the resultant conditions are then of the form

$$\sum_{k=1}^{n} A_k f_k(x_i) = F(x_i) \qquad (i = 1, 2, \cdots, n). \quad (272)$$

The matrix of the coefficients of the A's in this set of equations is then given by

$$\mathbf{f} = [f_{ij}] \quad \text{where} \quad f_{ij} = f_j(x_i), \quad (273)$$

that is, the set of equations can be expressed in the matrix form

$$\mathbf{f}\,\mathbf{A} = \mathbf{F} \quad (274)$$

where $\mathbf{A} = \{A_i\}$ and $\mathbf{F} = \{F_i\}$.

To illustrate the method, we again consider the integral equation (257), where $K(x, \xi)$ is defined by (258). For simplicity, we assume a three-term approximation of the polynomial form

$$y(x) \approx A_1 + A_2 x + A_3 x^2. \quad (275)$$

With $\phi_1 = 1$, $\phi_2 = x$, and $\phi_3 = x^2$, and with the notation of (268), there follows by direct integration

$$\Phi_1 = \tfrac{1}{2}x(1 - x), \quad \Phi_2 = \tfrac{1}{6}x(1 - x^2), \quad \Phi_3 = \tfrac{1}{12}x(1 - x^3). \quad (276)$$

Thus, with the notation of (270), equation (271) here takes the form

$$A_1\left[1 - \frac{x}{2}(1 - x)\right] + A_2\left[x - \frac{x}{6}(1 - x^2)\right] + A_3\left[x^2 - \frac{x}{12}(1 - x^3)\right] \approx x \qquad (0 \leqq x \leqq 1). \quad (277)$$

If we require that this relation be an equality at the three points $x = 0$, $x = \frac{1}{2}$, and $x = 1$, we obtain the conditions

$$\left.\begin{aligned} A_1 &= 0, \\ \tfrac{7}{8}A_1 + \tfrac{7}{16}A_2 + \tfrac{41}{192}A_3 &= \tfrac{1}{2}, \\ A_1 + A_2 + A_3 &= 1 \end{aligned}\right\}. \tag{278}$$

The solution of this set of equations, with the results rounded to four decimal places, is then given by

$$A_1 = 0, \quad A_2 = 1.2791, \quad A_3 = -0.2791, \tag{279}$$

so that the desired approximate solution is of the form

$$y(x) \approx 1.2791x - 0.2791x^2. \tag{280}$$

A comparison with the exact solution is postponed until Section 4.19 (page 458).

In those cases where the integrals defining the Φ's in (268) are not readily evaluated, or where $K(x, \xi)$ is defined empirically, the integrals may be evaluated approximately as weighted sums of ordinates (see Problem 89).

4.18. The method of weighting functions. A second method of obtaining n conditions for the determination of the constants, which is often associated with the name of Galerkin, consists in requiring that the difference between the two members of (271) be *orthogonal* to n linearly independent functions $\psi_i(x)$ $(i = 1, 2, \ldots, n)$ over the interval (a, b).

Thus, the conditions obtained in this way are of the form

$$\sum_{k=1}^{n} A_k \int_a^b \psi_i f_k \, dx = \int_a^b \psi_i F \, dx \qquad (i = 1, 2, \cdots, n), \tag{281}$$

or, equivalently,

$$\mathbf{a}\,\mathbf{A} = \mathbf{c} \quad \text{where} \quad a_{ij} = \int_a^b \psi_i f_j \, dx \quad \text{and} \quad c_i = \int_a^b \psi_i F \, dx. \tag{282}$$

The procedure actually consists in weighting the two members of (271) by each of the functions $\psi_i(x)$, and requiring that the integrals of the weighted members be equal. A particularly convenient choice of the n weighting functions is the set $1, x, x^2, \cdots, x^{n-1}$. In this case the graphical representations of the two members

of (271) are required to determine areas with the x-axis which are equal, and whose first $n-1$ *moments* are equal. It is desirable to choose the functions ψ_i as n members of a *complete* set of functions (see Section 1.28), since then the relation (271) must necessarily tend to an equality over (a, b) as n increases without limit. It is often convenient to identify the *weighting* functions ψ_i with the *approximating* functions ϕ_i.

In illustration, the application of this procedure to the example considered previously, with the weighting functions 1, x, and x^2, leads to the conditions

$$\left.\begin{aligned} \tfrac{11}{12}A_1 + \tfrac{11}{24}A_2 + \tfrac{37}{120}A_3 &= \tfrac{1}{2}, \\ \tfrac{11}{24}A_1 + \tfrac{14}{45}A_2 + \tfrac{17}{72}A_3 &= \tfrac{1}{3}, \\ \tfrac{37}{120}A_1 + \tfrac{17}{72}A_2 + \tfrac{321}{1680}A_3 &= \tfrac{1}{4} \end{aligned}\right\} \tag{283}$$

These equations are obtained by multiplying the two members of (277) successively by 1, x, and x^2, and equating the integrals of the results over (0, 1). The solution is found to be

$$A_1 = -0.0088, \quad A_2 = 1.2968, \quad A_3 = -0.2798, \tag{284}$$

leading to the approximation

$$y(x) \approx -0.0088 + 1.2968x - 0.2798x^2. \tag{285}$$

A comparison with the exact solution, and with the approximation (280), is presented in Section 4.19 (page 458).

4.19. The method of least squares. The accuracy obtained by the procedures of the two preceding sections will in general depend upon the choice of appropriate points of collocation or weighting functions. A method which avoids this dependence upon the judgment of the computer is next presented.

In place of requiring that the integral equation be satisfied *exactly* at a number of points equal to the number of undetermined coefficients (Section 4.17), we may require that the integral of the square of the difference between the two members, over (a, b), be as small as possible. Thus the basic condition is of the form

$$\int_a^b \Big[\sum_{k=1}^{n} A_k f_k(x) - F(x)\Big]^2 dx = \text{minimum}, \tag{286}$$

with the notation of (270) and (271). In order that (286) be satisfied, the derivative of the left-hand member with respect to each parameter A_i must vanish, so that we must have

$$\int_a^b f_i(x)\left[\sum_{k=1}^{n} A_k f_k(x) - F(x)\right] dx = 0 \quad (i = 1, 2, \cdots, n). \tag{287}$$

These conditions take the form

$$\sum_{k=1}^{n} A_k \int_a^b f_i f_k \, dx = \int_a^b f_i F \, dx \quad (i = 1, 2, \cdots, n), \tag{288}$$

and hence are equivalent to the conditions (281) where the weighting functions $\psi_i(x)$ are identified with the functions $f_i(x)$. Thus it follows that *if the integral equation is to be satisfied as well as possible over* (a, b) *in the least-squares sense, the weighting functions of the preceding section must be identified with the functions* $f_i(x)$.

In many practical cases, the functions f_i are such that the integrations involved in (288) are not feasible. Therefore, a modification which incorporates most of the advantages of this method over the collocation procedure, with only a small increase in the amount of calculation involved, is now formulated.

If the integral in (286) and (287) is approximated by a weighted sum of the relevant ordinates at N conveniently chosen points, the resultant minimal conditions (287) take the form

$$\sum_{r=1}^{N} D_r f_i(x_r)\left[\sum_{k=1}^{n} A_k f_k(x_r) - F(x_r)\right] = 0 \quad (i = 1, 2, \cdots, n), \tag{289}$$

where the numbers D_r $(r = 1, 2, \ldots, N)$ are appropriate weighting coefficients, associated with the points $x_1, x_2, \ldots, x_N$ involved in the approximate integration. These conditions can also be expressed in the form

$$\sum_{k=1}^{n} A_k\left[\sum_{r=1}^{N} D_r f_i(x_r) f_k(x_r)\right] = \sum_{r=1}^{N} D_r f_i(x_r) F(x_r) \quad (i = 1, 2, \cdots, n). \tag{290}$$

In spite of the rather formidable appearance of this set of conditions, the coefficients in the set of linear algebraic equations which it represents can be obtained very simply by matrix multiplication, as is next shown.

Equation (290) can be written in the abbreviated form

$$\sum_{k=1}^{n} A_k \alpha_{ik} = \beta_i \qquad (i = 1, 2, \cdots, n), \tag{291}$$

where

$$\alpha_{ik} = \sum_{r=1}^{N} D_r f_{ri} f_{rk} \tag{292a}$$

and

$$\beta_i = \sum_{r=1}^{N} D_r f_{ri} F_r, \tag{292b}$$

and where we have written

$$f_{ij} = f_j(x_i) \tag{293}$$

in accordance with the notation of (273). With a change of indices, equation (292a) can be written in the form

$$\alpha_{ij} = \sum_{k=1}^{N} f_{ki} D_k f_{kj} \qquad (i, j = 1, 2, \cdots, n), \tag{294}$$

and hence can be expressed by the matrix equation

$$\boldsymbol{\alpha} = \mathbf{f}^T \mathbf{D} \mathbf{f} \tag{295}$$

where $\boldsymbol{\alpha} = [\alpha_{ij}]$, $\mathbf{f} = [f_{ij}] = [f_j(x_i)]$, and $\mathbf{D} = [D_i \delta_{ij}]$. In the same way, we obtain also

$$\boldsymbol{\beta} = \mathbf{f}^T \mathbf{D} \mathbf{F} \tag{296}$$

where $\boldsymbol{\beta} = \{\beta_i\}$ and $\mathbf{F} = \{F_i\}$. Further, since $\mathbf{D}$ is a diagonal matrix, there follows also

$$\mathbf{f}^T \mathbf{D} = (\mathbf{D} \mathbf{f})^T,$$

so that (295) and (296) become

$$\boldsymbol{\alpha} = (\mathbf{D} \mathbf{f})^T \mathbf{f}, \quad \boldsymbol{\beta} = (\mathbf{D} \mathbf{f})^T \mathbf{F}. \tag{297}$$

If we notice that $\mathbf{f}$ is the matrix of the coefficients of the A's in the equations

$$\sum_{k=1}^{n} A_k f_k(x_i) = F(x_i) \qquad (i = 1, 2, \cdots, N), \tag{298}$$

it follows that the matrix of coefficients $\boldsymbol{\alpha}$ in (290) can be obtained by multiplying the matrix of coefficients $\mathbf{f}$ in (298) by the transpose of the matrix $\mathbf{D} \mathbf{f}$, and the column of right-hand members $\boldsymbol{\beta}$ in (290)

can be obtained similarly by multiplying the corresponding column **F** in (298) by the same matrix.

This result leads to the following procedure for determining the n linear equations represented by (290):

1. Choose N points $x_1, x_2, \ldots, x_N$ in the interval (a, b) and write down the N equations (298) which would require that the integral equation be satisfied at those points.

2. Denote the $N \times n$-matrix of coefficients in this set of equations by **f**, and form an associated "weighting matrix" $\mathbf{f}^* = \mathbf{D}\mathbf{f}$ by multiplying the ith *row* of **f** by the weighting coefficient D_i associated with the point x_i in an approximate integration scheme involving the N points.

3. Multiply the *augmented matrix* of (298) by the *transpose* of the weighting matrix $\mathbf{f}^*$. The resultant matrix is the augmented matrix of the required set of n linear equations which determines the constants $A_1, A_2, \ldots, A_n$.

We may notice that, since (290) is homogeneous in the D's, these weighting coefficients may be multiplied by any convenient common factor in the formation of $\mathbf{f}^*$. Thus if the formula of the trapezoidal rule is used, the successive coefficients are conveniently taken to be merely $\frac{1}{2}, 1, 1, \ldots, 1, 1, \frac{1}{2}$, so that the elements of the first and last rows of **f** are divided by two and the remaining entries are unchanged. Similarly, if Simpson's rule is used, the coefficients can be taken as $\frac{1}{2}, 2, 1, 2, 1, \ldots, 1, 2, \frac{1}{2}$.

As may be expected, this method leads to a set of equations equivalent to the original set when $N = n$, that is, when the number of chosen points x_i is equal to the number of A's to be determined. However, when $N > n$, it permits us to choose a number of points greater than n and to require that the integral equation be satisfied as nearly as possible at those points, rather than to require that it be satisfied *exactly* at n points. The weighting coefficients D_i weight the errors committed in failing to satisfy the equation at the respective points x_i in proportion to the influence of the ordinate at x_i in the integration of the squared error over (a, b). Whereas the N equations (298) are in general incompatible, this procedure affords the "best possible" solution in a least-squares sense.

It should be noticed that, by substitution of the calculated A's into the left-hand members of the equations (298), the difference between the two members of the integral equation can be readily

calculated at the N chosen points, to give an indication of the dependability of the solution. In physically motivated problems, it is often clear from the nature of the relevant physical phenomenon that small errors in the satisfaction of the integral equation necessarily imply also small errors in the unknown function. However, as was mentioned earlier, this situation does not *always* exist.

The present procedure differs from the collocation procedure of Section 4.17 in that, first, more than n equations are formed initially; second, a weighting matrix must be determined; and, third, an additional matrix multiplication is involved. Since additional equations would be needed in any case for the purpose of investigating the degree of satisfaction of the integral equation, this feature involves no additional calculation. As was shown, the formation of the weighting matrix need involve only multiplication or division of certain elements in the original coefficient matrix by a factor of two if the formulas of the trapezoidal rule or of Simpson's rule are used. The principal source of increased labor is involved in the matrix multiplication. However, the relevant operations are particularly well adapted to the use of automatic desk calculators, each element of the product matrix being determined by a single continuous sequence of machine operations.

In those cases when N is large (so that the matrix $\mathbf{f}$ possesses a large number of *rows*) it is often inconvenient to actually write the weighting matrix $\mathbf{f}^*$ in transposed form. In such cases it may be preferable to merely write the matrix $\mathbf{f}^*$ to the left of the original augmented matrix, without transposing its rows and columns, and to determine the product by column-into-column (rather than row-into-column) multiplication. The element in the ith row and jth column of the product matrix is then formed from the ith column of the first factor and the jth column of the second factor.

It is useful to notice that it follows from (294) that

$$\alpha_{ji} = \alpha_{ij}, \tag{299}$$

so that the coefficient matrix of the final set of equations is *symmetric*. This means that all elements below the principal diagonal of $\boldsymbol{\alpha}$ need not be calculated directly, but may be written down by symmetry once the remaining elements have been determined. The symmetry of the coefficient matrix also permits an appreciable

reduction of labor in the actual solution of the corresponding set of equations (see Appendix, page 504).

To illustrate the procedure just considered, we again deal with the example of the preceding sections. If we choose the five points $x = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and 1 as the points x_i, the five equations corresponding to (298) are obtained by equating the two members of (277) for those five values of x, in the form

$$\left.\begin{aligned} A_1 &= 0, \\ \tfrac{29}{32}A_1 + \tfrac{27}{128}A_2 + \tfrac{43}{1024}A_3 &= \tfrac{1}{4}, \\ \tfrac{7}{8}A_1 + \tfrac{7}{16}A_2 + \tfrac{41}{192}A_3 &= \tfrac{1}{2}, \\ \tfrac{29}{32}A_1 + \tfrac{89}{128}A_2 + \tfrac{539}{1024}A_3 &= \tfrac{3}{4}, \\ A_1 + A_2 + A_3 &= 1 \end{aligned}\right\}. \tag{300}$$

If five decimal places are retained in the calculations, the augmented matrix of the required set of three equations is then obtained as follows:

$$\begin{bmatrix} 0.50000 & 0 & 0 \\ 1.81250 & 0.42188 & 0.08398 \\ 0.87500 & 0.43750 & 0.21354 \\ 1.81250 & 1.39062 & 1.05274 \\ 0.50000 & 0.50000 & 0.50000 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.90625 & 0.21094 & 0.04199 & 0.25000 \\ 0.87500 & 0.43750 & 0.21354 & 0.50000 \\ 0.90625 & 0.69531 & 0.52637 & 0.75000 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 5.05078 & 2.52539 & 1.71700 & 2.75000 \\ 2.52539 & 1.74731 & 1.34312 & 1.86718 \\ 1.71700 & 1.34312 & 1.10326 & 1.41732 \end{bmatrix}. \tag{301}$$

The second factor in the product is merely the augmented matrix of (300). In forming the weighting matrix which precedes it, we have used the weighting coefficients $\frac{1}{2}$, 2, 1, 2, $\frac{1}{2}$ corresponding to the formula of Simpson's rule. The corresponding set of equations,

$$\left.\begin{aligned} 5.05078A_1 + 2.52539A_2 + 1.71700A_3 &= 2.75000, \\ 2.52539A_1 + 1.74731A_2 + 1.34312A_3 &= 1.86718, \\ 1.71700A_1 + 1.34312A_2 + 1.10326A_3 &= 1.41732 \end{aligned}\right\}, \tag{302}$$

then is found to possess the solution

$$A_1 = -0.0079, \quad A_2 = 1.2939, \quad A_3 = -0.2783, \tag{303}$$

leading to the approximation

$$y(x) \approx -0.0079 + 1.2939x - 0.2783x^2. \tag{304}$$

In the following table we compare the results of (A) three-point collocation [equation (280)], (B) use of weighting functions 1, x, and x^2 [equation (285)], and (C) five-point least squares with Simpson's rule [equation (304)], with the exact solution given by equation (259):

		Approximate Solutions			
x	$y(x)$	(A)	(B)	(C)	$\bar{y}(x)$
0	0	0	−0.0088	−0.0079	−0.0090
0.1	0.1186	0.1251	0.1181	0.1187	0.1180
0.2	0.2361	0.2446	0.2394	0.2398	0.2393
0.3	0.3512	0.3586	0.3550	0.3552	0.3550
0.4	0.4628	0.4670	0.4651	0.4652	0.4652
0.5	0.5697	0.5698	0.5696	0.5695	0.5697
0.6	0.6710	0.6670	0.6685	0.6683	0.6686
0.7	0.7656	0.7586	0.7618	0.7615	0.7619
0.8	0.8526	0.8446	0.8495	0.8491	0.8496
0.9	0.9309	0.9251	0.9316	0.9312	0.9316
1.0	1.0000	1.0000	1.0081	1.0077	1.0081

For the purpose of further comparison, there are included in the last column of the table the values of the parabola $\bar{y} = A_1 + A_2x + A_3x^2$ which gives the best least-squares approximation to the exact solution itself, over the interval (0, 1). The coefficients were determined in such a way that the integral

$$\int_0^1 \left[\frac{\sin x}{\sin 1} - (A_1 + A_2x + A_3x^2)\right]^2 dx$$

takes on a minimum value, and are defined by the equation

$$\bar{y}(x) = -0.0090 + 1.2976x - 0.2805x^2. \tag{305}$$

The example considered was chosen for the purpose of simplicity, and also for the reason that the exact solution is known and can be reasonably well approximated by a parabola over the relevant interval. It may be noticed that approximation (B) agrees very

closely with the best possible parabolic approximation $\bar{y}$, that it affords a better approximation to the exact solution y than does the collocation approximation (A), and that the five-point least-squares approximation (C) is only slightly less accurate than (B).

Because of the simplicity of the coefficient functions $f_i(x)$ appearing in (277), in the present case, the formulation of equations (283) was very easily accomplished. Indeed, the amount of relevant calculation was less than that involved in the formation of the approximate least-squares equations (302). In more involved problems, in which the integrals involved in (281) frequently must be evaluated by approximate methods, the modified least-squares procedure of the present section is usually preferable because of the fact that the relevant numerical calculations are carried out in a systematic way.

In view of the fact that the integral equation specified by (257) and (258) implies the obvious end conditions $y(0) = 0$, $y(1) = 1$, it is to be expected that a three-parameter approximation of the form

$$y(x) \approx x + x(1 - x)(B_1 + B_2x + B_3x^2)$$

would lead to much more nearly accurate results.

When polynomial approximation is used in connection with the method of collocation, or the modified least-squares procedure, the calculations involved can be further systematized if the approximating polynomial is expressed in the so-called *Lagrangian form*. Details may be found in References 5 and 6.

4.20. Approximation of the kernel. As was mentioned in Section 4.6, it is sometimes convenient to approximate the *kernel* of a Fredholm integral equation by a polynomial in x and ξ, or by a *separable kernel* of more general form, and to solve the resultant approximate equation by the methods of that section.

Thus, for example, the kernel (258) could be approximated by a three-parameter polynomial, of the form $A_1 + A_2x + A_3x^2$ or of the more appropriate form $x(1 - x)(B_1 + B_2x + B_3x^2)$, where the A's or B's are determined as functions of ξ by three-point collocation, the use of appropriate weighting functions, or the use of least-square techniques.

To illustrate the procedure, we assume a crude approximation of the form

$$K(x, \xi) \approx B\,x(1 - x). \tag{306}$$

Noticing that the approximation is exact at the end-points $x = 0$ and $x = 1$, we determine the coefficient B in such a way that the integral of the kernel over (0, 1) is equal to the integral of its approximation over that interval:

$$\int_0^1 K(x, \xi)\, dx = B \int_0^1 x(1 - x)\, dx. \tag{307}$$

Direct calculation then gives the determination

$$B = 3\xi(1 - \xi), \tag{308}$$

and the introduction of the corresponding approximate kernel into (257) leads to the approximating integral equation

$$y(x) = x + 3x(1 - x) \int_0^1 \xi(1 - \xi) y(\xi)\, d\xi. \tag{309}$$

Following the method of Section 4.6, we introduce the abbreviation

$$c = \int_0^1 x(1 - x) y(x)\, dx, \tag{310}$$

and rewrite (309) in the form

$$y(x) = x + 3cx(1 - x). \tag{311}$$

In order to determine c, we multiply the equal members of (311) by $x(1 - x)$ and integrate the results over (0, 1), to obtain the condition

$$c = \int_0^1 x^2(1 - x)\, dx + 3c \int_0^1 x^2(1 - x)^2\, dx,$$

and the evaluation

$$c = \tfrac{5}{54}. \tag{312}$$

Hence the desired approximate solution (311) is obtained in the form

$$y(x) \approx x + \tfrac{5}{18}x(1 - x) \doteq 1.2778x - 0.2778x^2. \tag{313}$$

The approximation very nearly coincides with that of (280).

More generally, it is easily seen that a kernel approximation of the special form

$$K(x, \xi) \approx x\xi(1 - x)(1 - \xi)(a_1 + a_2 x\xi + a_3 x^2\xi^2 + \cdots)$$

would lead to an approximate solution of the form

$$y(x) \approx x + x(1 - x)(c_1 + c_2x + c_3x^2 + \cdots)$$

n the case of the present example.

REFERENCES

. Lovitt, W. V.: *Linear Integral Equations*, McGraw-Hill Book Company Inc., New York, 1924.

. Courant, R., and D. Hilbert: *Methoden der mathematischen Physik*, Interscience Publishers, Inc., New York, 1943.

. Frank, Ph., and R. von Mises: *Die Differential- und Integralgleichungen der Mechanik und Physik*, Rosenberg, New York, 1943.

. Whittaker, E. T., and G. N. Watson: *A Course in Modern Analysis*, Cambridge University Press, London, 1927.

. Crout, P. D.: "An Application of Polynomial Approximation to the Solution of Integral Equations Arising in Physical Problems," *J. Math. Phys.*, Vol. 19, No. 1 (1940).

. Crout, P. D., and F. B. Hildebrand: "A Least Square Procedure for Solving Integral Equations by Polynomial Approximation," *J. Math. Phys.*, Vol. 20, No. 3 (1941).

PROBLEMS

lection 4.1.

1. (a) If $y''(x) = F(x)$, and y satisfies the initial conditions $y(0) = y_0$ nd $y'(0) = y_0'$, show that

$$y(x) = \int_0^x (x - \xi)F(\xi)\, d\xi + y_0'x + y_0.$$

Notice that $y'(x) = \int_0^x F(x)\, dx + y_0'$, and use (10).]

(b) Verify directly that this expression satisfies the prescribed ifferential equation and initial conditions.

2. (a) If $y''(x) = F(x)$, and y satisfies the end conditions $y(0) = 0$ nd $y(1) = 0$, show that

$$y(x) = \int_0^x (x - \xi)F(\xi)\, d\xi - x\int_0^1 (1 - \xi)F(\xi)\, d\xi.$$

Set $y_0 = 0$ in the result of Problem 1, and determine y_0' so that $y(1) = 0$.]

(b) Show that the result of part (a) can be written in the form

$$y(x) = \int_0^1 K(x, \xi)F(\xi)\, d\xi,$$

where $K(x, \xi)$ is defined by the relations

$$K(x, \xi) = \begin{cases} \xi(x - 1) & \text{when} \quad \xi < x, \\ x(\xi - 1) & \text{when} \quad \xi > x. \end{cases}$$

(c) Verify directly that the expression obtained satisfies the prescribed differential equation and end conditions.

Section 4.2.

3. (a) Show that, if $y(x)$ satisfies the differential equation

$$\frac{d^2y}{dx^2} + x\, y = 1$$

and the conditions $y(0) = y'(0) = 0$, then y also satisfies the Volterra equation

$$y(x) = \int_0^x \xi(\xi - x)y(\xi)\, d\xi + \frac{1}{2}x^2.$$

(b) Prove that the converse of the preceding statement is also true.

4. Suppose that a sequence of approximate solutions is obtained for the integral equation of Problem 3, the $(n + 1)$th approximation $y^{(n+1)}(x)$ being defined by substitution of the nth approximation into the right-hand member:

$$y^{(n+1)}(x) = \int_0^x \xi(\xi - x)y^{(n)}(\xi)\, d\xi + \frac{1}{2}x^2.$$

(a) Taking $y^{(0)}(x) = 0$, obtain the functions

$$y^{(1)}(x) = \tfrac{1}{2}x^2 \quad \text{and} \quad y^{(2)}(x) = \tfrac{1}{2}x^2 - \tfrac{1}{40}x^5$$

as the two succeeding approximations.

(b) Obtain the first two nonvanishing terms in the power-series solution of the problem considered in Problem 3(a), in the form

$$y(x) = a_0 + a_1x + a_2x^2 + \cdots,$$

and compare the result with that of part (a). [Notice that we must have $a_0 = a_1 = 0$, to satisfy the initial conditions, and determine the remaining a's by substitution in the differential equation.]

5. (a) Show that, if $y(x)$ satisfies the differential equation

$$x\frac{d^2y}{dx^2} + \frac{dy}{dx} + x\, y = x$$

when $x \geqq 0$, and if y, y', and y'' are finite at $x = 0$, then there must follow $y'(0) = 0$. [This conclusion follows directly from the differential equation.

Further, the most general solution which is finite at $x = 0$ is found to be $y(x) = 1 + c\,J_0(x)$, where $J_0(x)$ is the Bessel function of first kind, of order zero. In virtue of the fact that $J_0(0) = 1$ and $J_0'(0) = 0$, the value of $y(0)$ can be arbitrarily prescribed, whereas $y'(0)$ *cannot* be prescribed. This situation is a consequence of the fact that $x = 0$ is a *singular point* of the differential equation.]

(b) Show that the integral-equation formulation of equation (13) is not applicable to the problem of part (a), if the initial conditions are prescribed at point $x = 0$. [When the equation is written in the form of equation (11), the function $A(x)$ is not finite at $x = 0$, and hence the right-hand member of (14b) is undefined. The integration by parts which led to (13) was not legitimate in this case.]

6. By integrating the equal members of the differential equation of Problem 5(a) twice over the interval $(0, x)$, and simplifying the integrals $\int x\,y''\,dx$ and $\int x\,y'\,dx$ by integration by parts in successive steps, show that $y(x)$ must satisfy the integral equation

$$x\,y(x) = \int_0^x [\xi(\xi - x) + 1]y(\xi)\,d\xi + \frac{1}{6}x^3,$$

of the "third kind," regardless of the prescribed initial condition. [Notice that this equation hence must possess *infinitely many* solutions, each of the form $y(x) = 1 + c\,J_0(x)$, where c is an arbitrary constant. This situation is a consequence of the fact that, when this equation is written in the form (13), the kernel $K(x, \xi)$ becomes infinite when $x = 0$.]

7. Obtain an alternative integral-equation formulation of the problem described by equations (11) and (12), by first setting $y'' = u$, and showing that

$$y(x) = \int_a^x (x - \xi)u(\xi)\,d\xi + y_0'(x - a) + y_0$$

where $u(x)$ satisfies an integral equation

$$u(x) = \int_a^x [(\xi - x)B(x) - A(x)]u(\xi)\,d\xi + F(x).$$

8. Show that the application of the method of Section 4.2 to the problem $y'' + A\,y' + B\,y = 0$, $y(0) = y(1) = 0$, where A and B are constants, leads to the integral equation $y(x) = \int_0^1 K(x, \xi)y(\xi)\,d\xi$,

where $$K(x, \xi) = \begin{cases} B\,\xi(1 - x) + A\,x - A & \text{when } \xi < x, \\ B\,x(1 - \xi) + A\,x & \text{when } \xi > x. \end{cases}$$

[Notice that the kernel obtained in this way is nonsymmetric, and discontinuous at $\xi = x$, unless $A = 0$.]

(b) Divide the interval of integration in the left-hand member into the subintervals (a, ξ) and (ξ, b) [so that Green's formula of Problem 13(a) applies over each subinterval], and show that the resultant expression

$$-\int_a^{\xi} G_1(x, \xi)\, L\, y(x)\, dx - \int_{\xi}^{b} G_2(x, \xi)\, L\, y(x)\, dx,$$

can be written in the form

$$-p(\xi)\left[G_1(\xi, \xi)y'(\xi) - \frac{\partial G_1(x, \xi)}{\partial x}\bigg|_{x=\xi} y(\xi)\right] + p(\xi)\left[G_2(\xi, \xi)y'(\xi) - \frac{\partial G_2(x, \xi)}{\partial x}\bigg|_{x=\xi} y(\xi)\right] = y(\xi).$$

(c) By appropriately changing variables, deduce that $y(x)$ must therefore satisfy (36).

15. Suppose that a function $U(x)$ satisfies the equation $L\,y = 0$ in an interval (a, b), where L is the self-adjoint operator defined in Problem 13, and also satisfies certain homogeneous conditions at the ends of the interval. Prove that the equation $L\,y + \Phi = 0$ cannot possess a solution, valid everywhere in (a, b) and satisfying the *same* homogeneous conditions at the end points, unless $\Phi(x)$ is "orthogonal" to $U(x)$, that is, unless the condition

$$\int_a^b U(x)\Phi(x)\, dx = 0$$

is satisfied. [Assume that such a solution $y(x)$ exists. Multiply the equal members of the relation $\Phi = -L\,y$ by $U(x)$, integrate the results over (a, b), and use Green's formula of Problem 13(b).] Also, verify this result in the case of the problem $y'' + \Phi = 0$, $y'(0) = y'(1) = 0$, with $\Phi = 1$ and with $\Phi = 2x - 1$, noticing that here $U(x) =$ constant.

16. *The generalized Green's function.* Suppose that a problem, consisting of the differential equation $L\,y \equiv (p\,y')' + q\,y = 0$ and homogeneous conditions prescribed at the ends of an interval (a, b), is satisfied by a function $y = U(x)$, so that the Green's function defined by properties 1 to 4 of page 388 does not exist. The *generalized* Green's function is then defined as a function H which, when considered as a function of x for a fixed number ξ, possesses the following properties:

1. H satisfies the differential equation

$$L\,H = C\,U(x)U(\xi)$$

in the subintervals (a, ξ) and (ξ, b).

2. H satisfies the prescribed end conditions.

3. H is continuous at $x = \xi$.

4. The x-derivative of H possesses a jump of magnitude $-1/[p(\xi)]$ as the point $x = \xi$ is crossed in the positive x-direction.

5. H satisfies the condition

$$\int_a^b H(x, \xi)U(x)\,dx = 0.$$

Show, by the following steps, that (if such a function exists) the function H has the property that, if Φ is any function such that $\int_a^b U\Phi\,dx = 0$, a solution of the equation

$$L\,y + \Phi = 0,$$

subject to the same homogeneous end conditions, is of the form

$$y(x) = \int_a^b H(x, \xi)\Phi(\xi)\,d\xi.$$

(a) By writing $H = H_1$ when $x < \xi$ and $H = H_2$ when $x > \xi$, show that there follows

$$y(x) = \int_a^x H_2(x, \xi)\Phi(\xi)\,d\xi + \int_x^b H_1(x, \xi)\Phi(\xi)\,d\xi,$$

$$y'(x) = \int_a^x \frac{\partial H_2(x, \xi)}{\partial x}\,\Phi(\xi)\,d\xi + \int_x^b \frac{\partial H_1(x, \xi)}{\partial x}\,\Phi(\xi)\,d\xi,$$

and

$$y''(x) = \int_a^x \frac{\partial^2 H_2(x, \xi)}{\partial x^2}\,\Phi(\xi)\,d\xi + \int_x^b \frac{\partial^2 H_1(x, \xi)}{\partial x^2}\,\Phi(\xi)\,d\xi - \frac{1}{p(x)}\,\Phi(x).$$

[Make use of properties 3 and 4.]

(b) Verify that

$$L\,y(x) = \int_a^x [L\,H_2(x, \xi)]\Phi(\xi)\,d\xi + \int_x^b [L\,H_1(x, \xi)]\Phi(\xi)\,d\xi - \Phi(x)$$

$$= \int_a^b [C\,U(x)U(\xi)]\Phi(\xi)\,d\xi - \Phi(x)$$

$$= -\Phi(x).$$

[Make use of property 1 and the restriction on Φ. *Notice that satisfaction of property 5 is not necessary.* (See, however, Problem 19.)]

17. Suppose that $U(x)$ is a solution of the equation $L\,y = 0$ in the interval (a, b), and that $U(x)$ satisfies certain prescribed homogeneous conditions at the ends of that interval. Let $u(x)$ denote a function such that $L\,u(x) = U(x)$ when $a \leqq x \leqq \xi$, and let $v(x)$ denote a function such that $L\,v(x) = U(x)$ when $\xi \leqq x \leqq b$. Finally, suppose that $u(x)$ satisfies the prescribed condition at $x = a$, whereas $v(x)$ satisfies the prescribed condition at $x = b$.

(a) By setting $f = U$ and $g = u$ in Green's formula [Problem 13(a)], show that

$$\int_a^\xi [U(x)]^2\,dx = p(\xi)[U(\xi)u'(\xi) - U'(\xi)u(\xi)].$$

(b) In a similar way, show that

$$\int_\xi^b [U(x)]^2\,dx = -p(\xi)[U(\xi)v'(\xi) - U'(\xi)v(\xi)].$$

(c) Deduce that

$$p(\xi)\{[u'(\xi) - v'(\xi)]U(\xi) - [u(\xi) - v(\xi)]U'(\xi)\} = \int_a^b [U(x)]^2\,dx.$$

18. With the terminology of Problems 16 and 17, verify that, if the function $U(x)$ is normalized in such a way that

$$\int_a^b [U(x)]^2\,dx = 1,$$

then the function

$$H(x, \xi) = \alpha(\xi)U(x) + \begin{cases} v(\xi)U(x) + u(x)U(\xi) & \text{when } x < \xi, \\ u(\xi)U(x) + v(x)U(\xi) & \text{when } x > \xi \end{cases}$$

satisfies properties 1, 2, 3, and 4 of Problem 16, regardless of the form of the function $\alpha(\xi)$, so that the required generalized Green's function is obtained by determining $\alpha(\xi)$ by condition 5, in such a way that

$$\int_a^b H(x, \xi)U(x)\,dx = 0.$$

[Use the result of Problem 17 in investigating the satisfaction of property 4.]

19. (a) By writing $f(x) = H(x, r)$ and $g(x) = H(x, s)$ in Green's formula [Problem 13(a)], where $H(x, \xi)$ is the generalized Green's function relevant to the operator L in (a, b), show that

$$\int_a^b [H(x, r)\,L\,H(x, s) - H(x, s)\,L\,H(x, r)]\,dx = H(s, r) - H(r, s).$$

[Write $\int_a^b = \int_a^{r-} + \int_{r+}^{s-} + \int_{s+}^b$ and recall that $\partial H(x, r)/\partial x$ has a jump of magnitude $-1/[p(r)]$ at $x = r$, whereas $\partial H(x, s)/\partial x$ has a jump of magnitude $-1/[p(s)]$ at $x = s$.]

(b) Show that the left-hand member of the preceding equation can be written in the form

$$U(s)\int_a^b H(x, r)U(x)\,dx - U(r)\int_a^b H(x, s)U(x)\,dx,$$

and hence vanishes in consequence of the satisfaction of Property 5. Thus deduce that the generalized Green's function is symmetric:

$$H(x, \xi) = H(\xi, x).$$

[Notice that this proof applies also to the conventional Green's function (with $U = 0$), although symmetry in this case is obvious from the explicit form (30).]

20. Prove (by methods similar to those used in Problem 14) that the differential equation $Ly + \lambda y = 0$ with associated homogeneous conditions at the ends of the interval (a, b), together with the requirement

$$\int_a^b U(x)y(x)\,dx = 0,$$

where $U(x)$ is a function satisfying $LU = 0$ and the boundary conditions, implies the integral equation

$$y(x) = \lambda \int_a^b H(x, \xi)r(\xi)y(\xi)\,d\xi,$$

where H is the generalized Green's function.

21. Determine the generalized Green's function, relevant to the end conditions $y'(0) = y'(1) = 0$, for the expression $L\,y = d^2y/dx^2$, in the form

$$H(x, \xi) = \tfrac{1}{3} + \tfrac{1}{2}(x^2 + \xi^2) - \begin{cases} \xi & \text{when } x < \xi, \\ x & \text{when } x > \xi. \end{cases}$$

[Use either the basic properties of Problem 16 or the formula of Problem 18, with $U(x) = 1$.] Deduce also that the problem

$$\frac{d^2y}{dx^2} + \lambda y = 0, \qquad y'(0) = y'(1) = 0$$

is transformable to the integral equation

$$y(x) = \lambda \int_0^1 H(x, \xi)y(\xi)\,d\xi.$$

22. For the Legendre operator $L = \dfrac{d}{dx}\left[(1 - x^2)\dfrac{d}{dx}\right]$ there follows $p(x) = 1 - x^2$, and hence $p(\pm 1) = 0$. Thus appropriate end conditions for the expression $L\,y$ in the interval $(-1, 1)$ consist in the requirements that $y(-1)$ and $y(1)$ be *finite*. Noticing that the function $U(x) =$ constant satisfies the equation $L\,y = 0$ and these finiteness conditions (so that the conventional Green's function does not exist), obtain the generalized Green's function in the form

$$H(x, \xi) = \log 2 - \tfrac{1}{2} - \begin{cases} \frac{1}{2}\log[(1 + \xi)(1 - x)] & \text{when } x < \xi, \\ \frac{1}{2}\log[(1 - \xi)(1 + x)] & \text{when } x > \xi. \end{cases}$$

Deduce also that the problem

$$(1 - x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \lambda y = 0, \qquad y(\pm 1) \text{ finite}$$

transforms into the integral equation

$$y(x) = \lambda \int_{-1}^{1} H(x, \xi) y(\xi)\, d\xi.$$

[Notice that one must take $l'(x) = 1/\sqrt{2}$, if the result of Problem 18 is used.]

Section 4.4.

23. Obtain an explicit solution of the problem $y''(x) + \Phi_\epsilon(x) = 0$, in the interval $(0, 1)$, where $\Phi_\epsilon(x)$ vanishes outside the interval $(\xi - \epsilon, \xi + \epsilon)$ and is given by $1/(2\epsilon)$ inside that interval, and where $y(0) = 0$ and $y(1) = 0$. [Determine the general solution in each of the subintervals $(0, \xi - \epsilon)$, $(\xi - \epsilon, \xi + \epsilon)$, and $(\xi + \epsilon, 1)$, and determine the six constants of integration by satisfying the end conditions and requiring that y and y' be continuous at the transition points.] Show that the solution is of the form

$$y(x) = \begin{cases} x(1 - \xi) & \text{when } 0 < x < \xi - \epsilon, \\ \dfrac{x + \xi}{2} - x\xi - \dfrac{\epsilon}{4} - \dfrac{(x - \xi)^2}{4\epsilon} & \text{when } \xi - \epsilon < x < \xi + \epsilon, \\ (1 - x)\xi & \text{when } \xi + \epsilon < x < 1, \end{cases}$$

and notice that this form tends to the relevant Green's function of $L\,y = d^2y/dx^2$, subject to the prescribed end conditions, as $\epsilon \to 0$.

24. Suppose that $G(x, y; \xi, \eta)$ is the Green's function for the Laplacian expression $\nabla^2 w$ in a simple region R of the xy-plane, relevant to the requirement that w vanish along the boundary C, so that the solution of Poisson's equation $\nabla^2 w + \Phi(x, y) = 0$, subject to that boundary condition, is of the form

$$w(x, y) = \iint_R G(x, y; \xi, \eta)\Phi(\xi, \eta)\, d\xi\, d\eta,$$

and the corresponding solution of the equation $\nabla^2 w + \lambda w = 0$ also satisfies the integral equation

$$w(x, y) = \lambda \iint_R G(x, y; \xi, \eta) w(\xi, \eta)\, d\xi\, d\eta.$$

It can be shown (by a method analogous to that used in Problem 19) that G is *symmetric* in (x, y) and (ξ, η), so that $G(x, y; \xi, \eta) = G(\xi, \eta; x, y)$. Assuming this fact, obtain a further useful property of the Green's function, by the following steps:

(a) By applying Green's theorem of vector analysis, in the form

$$\iint_R (\phi_1 \nabla^2\phi_2 - \phi_2 \nabla^2\phi_1)\, dA = \oint_B \left(\phi_1 \frac{\partial \phi_2}{\partial n} - \phi_2 \frac{\partial \phi_1}{\partial n}\right) ds,$$

to the region R' of the $\xi\eta$-plane in Figure 4.2, with $\phi_1 = G(x, y; \xi, \eta)$ and $\phi_2 = \phi(\xi, \eta)$, where the interior of a small circle C_ϵ of radius ϵ about the point $P(x, y)$ is deleted from R, show that

$$\iint_{R'} (G\,\nabla^2\phi - \phi\,\nabla^2 G)\,d\xi\,d\eta = \oint_C \left(G\frac{\partial\phi}{\partial n} - \phi\frac{\partial G}{\partial n}\right) ds + \oint_{C_\epsilon} \left(G\frac{\partial\phi}{\partial n} - \phi\frac{\partial G}{\partial n}\right) ds,$$

where here $\nabla^2 = \dfrac{\partial^2}{\partial\xi^2} + \dfrac{\partial^2}{\partial\eta^2}$, and where the normal differentiation is with respect to the coordinates ξ and η.

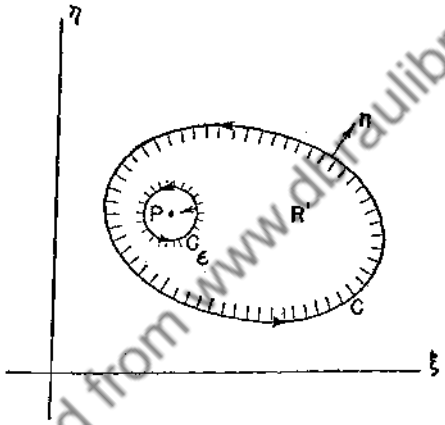


FIGURE 4.2

(b) Suppose that ϕ *satisfies* Laplace's equation, $\nabla^2\phi = 0$, everywhere inside R. Noticing that also $\nabla^2 G = 0$ except at the point P, and that the normal derivatives calculated along C_ϵ are along the *inward* normal relative to C_ϵ, show that, as $\epsilon \to 0$, there follows

$$\oint_C \left(G\frac{\partial\phi}{\partial n} - \phi\frac{\partial G}{\partial n}\right) ds = \lim_{r\to 0} \int_0^{2\pi} \left(G\frac{\partial\phi}{\partial r} - \phi\frac{\partial G}{\partial r}\right) r\,d\theta,$$

where r denotes radial distance outward from the point P (so that $\partial/\partial n = -\partial/\partial r$ on C_ϵ), and θ denotes angular position along C_ϵ. Show also that the limit on the right is given formally by $\phi(x, y)$, in consequence of the properties of the Green's function. [See equation (61′).]

(c) Deduce that, *since* G vanishes along the boundary C, there follows

$$\phi(x, y) = -\oint_C \phi\frac{\partial G}{\partial n}\,ds,$$

where the normal differentiation is with respect to the coordinates ξ and η, so that the value of ϕ at an interior point of R is thus expressed in terms of prescribed values of ϕ along the boundary C with the help of the normal derivative of the Green's function along C, and the solution of the Dirichlet problem for the interior of the region R is obtained.

25. Suppose that a function $f(z)$ of the complex variable $z = x + iy$ can be found with the following properties (Figure 4.3):

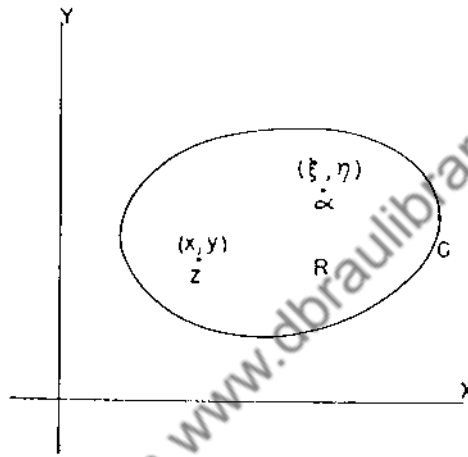


FIGURE 4.3

1. $f(z)$ is analytic everywhere inside a region R of the xy-plane and on the boundary C.

2. $|f(z)| = 1$ at all points of C and $|f(z)| < 1$ inside C.

3. $f(z)$ possesses a simple zero at the point (ξ, η), that is, at the complex point $\alpha = \xi + i\eta$, and differs from zero elsewhere in R and on C.

(a) Show that the function

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log |f(z)|$$

$$= -\frac{1}{2\pi} (\mathrm{R}_e [\log f(z)]$$

is the Green's function of the Laplacian expression $\nabla^2 w$ relevant to the requirement that w vanish along the boundary of R. [Recall that the real and imaginary parts of $f(z)$ satisfy Laplace's equation at points where $f(z)$ is analytic, and that $\log f(z)$ is analytic when $f(z)$ is analytic and $f(z) \neq 0$. Also notice that here we may write $f(z) = (z - \alpha)\phi(z)$, where $\phi(z)$ is analytic *and nonzero* everywhere in R and on C.]

(b) Show that the function $f(z)$ defined in part (a) has the property that the relation $w = f(z)$ maps the interior of R into the interior of the

unit circle $| w | = 1$ of the w-plane (Figure 4.4), with the point $\alpha = \xi + i\eta$ being mapped into the origin. [Notice that the requirement that $f(z)$

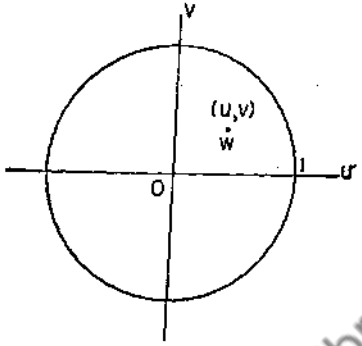


FIGURE 4.4

possess only a *simple* zero at $z = \alpha$ insures that $f'(z) \neq 0$, so that the mapping is indeed one to one.]

26. (a) Verify that the mapping

$$w = e^{ia}\frac{z - \alpha}{1 - \bar{\alpha}z} \equiv -\frac{e^{ia}}{\bar{\alpha}}\frac{z - \alpha}{z - 1/\bar{\alpha}},$$

where $\alpha = \xi + i\eta$ and $\bar{\alpha} = \xi - i\eta$ and a is any real constant, maps the boundary and interior of the unit circle in the z-plane into the boundary and interior of the unit circle in the w-plane if α is inside the unit circle (that is, if $| \alpha | < 1$), and that the point α maps into the origin. [Calculate $w\bar{w}$, and show that $w\bar{w} \equiv | w |^2$ is unity when $z\bar{z} \equiv | z |^2 = 1$. Also, show that the point $1/\bar{\alpha}$ is the image of the point α in an inversion relative to the unit circle (see Figure 4.5) and hence deduce from the second form of w

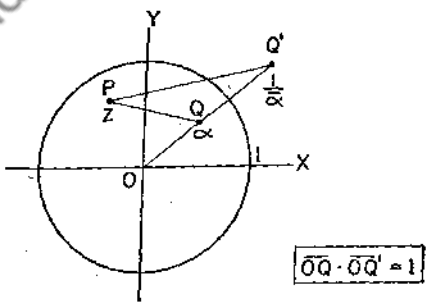


FIGURE 4.5

that $| w | < 1$ when z and α are inside the unit circle, recalling that $| z - \alpha |$ is the *distance* between the points z and α.]

(b) Deduce that the Green's function of $\nabla^2 w$ for the interior of the unit circle, relevant to the requirement that w vanish along the boundary, is of the form

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log \left| \frac{z - \alpha}{1 - \bar{\alpha} z} \right|,$$

where $\alpha = \xi + i\eta$, $\bar{\alpha} = \xi - i\eta$, and $z = x + iy$.

27. (a) With the introduction of the polar representations

$$z = \rho e^{i\theta}, \qquad \alpha = \beta e^{i\phi},$$

so that $x = \rho \cos \theta$, $y = \rho \sin \theta$, $\xi = \beta \cos \phi$, and $\eta = \beta \sin \phi$, show that the Green's function of Problem 26 takes the form

$$G(\rho, \theta; \beta, \phi) = \frac{1}{4\pi} \log \frac{1 - 2\beta\rho \cos(\theta - \phi) + \beta^2\rho^2}{\beta^2 - 2\beta\rho \cos(\theta - \phi) + \rho^2}.$$

(b) Deduce that the formal solution of the problem $\nabla^2 w + \Phi = 0$ inside the unit circle $\rho = 1$, where $w(1, \theta) = 0$, is of the form

$$w(\rho, \theta) = \int_0^1 d\beta \int_0^{2\pi} G(\rho, \theta; \beta, \phi) \Phi(\beta, \phi)\, d\phi,$$

and that the formal solution of the Dirichlet problem $\nabla^2 w = 0$ in the same region, when $w(1, \theta)$ is prescribed, is of the form

$$w(\rho, \theta) = -\int_0^{2\pi} \left[\frac{\partial G}{\partial \beta}\right]_{\beta=1} w(1, \phi)\, d\phi$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \frac{1 - \rho^2}{1 - 2\rho \cos(\theta - \phi) + \rho^2} w(1, \phi)\, d\phi.$$

[The last result is the well-known *Poisson integral formula*, relevant to the unit circle.]

28. Suppose that an analytic function $F(z)$ maps the interior of a region R into the interior of the unit circle, but does not necessarily map the point $z = \alpha = \xi + i\eta$ into the origin. Use the result of Problem 26 to show that the Green's function described in that problem is of the form

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log \left| \frac{F(z) - F(\alpha)}{1 - \overline{F(\alpha)} F(z)} \right|.$$

[Let the relation $t = F(z)$ map R into the unit circle of a t-plane, so that $z = \alpha$ maps into $t = F(\alpha)$, and map the t-plane into the w-plane by the mapping of Problem 26(a).]

29. (a) Verify that the function

$$f(z) = e^{i\alpha} \frac{z - \alpha}{z - \bar{\alpha}},$$

where a is a real constant, maps the upper half-plane ($y > 0$) into the interior of the unit circle when α is in the upper half-plane, and deduce from Problem 25 that the Green's function for the upper half-plane is of the form

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log \left| \frac{z - \alpha}{z - \bar{\alpha}} \right| = -\frac{1}{2\pi} \log \left| \frac{(x - \xi) + i(y - \eta)}{(x - \xi) + i(y + \eta)} \right|$$

$$= \frac{1}{4\pi} \log \frac{(x - \xi)^2 + (y + \eta)^2}{(x - \xi)^2 + (y - \eta)^2}.$$

(b) Deduce that the formal solution of the problem $\nabla^2 w + \Phi = 0$ in the upper half-plane, where $w(x, 0) = 0$, is of the form

$$w(x, y) = \int_{-\infty}^{\infty} d\xi \int_0^{\infty} G(x, y; \xi, \eta)\Phi(\xi, \eta)\, d\eta,$$

and that the formal solution of the problem $\nabla^2 w = 0$ in the same region, where $w(x, 0) = \phi(x)$, is of the form

$$w(x, y) = -\int_{-\infty}^{\infty} \left[\frac{\partial G}{\partial(-\eta)} \right]_{\eta=0} \phi(\xi)\, d\xi$$

$$= \frac{y}{\pi} \int_{-\infty}^{\infty} \frac{\phi(\xi)\, d\xi}{(\xi - x)^2 + y^2} \qquad (y > 0).$$

[See Problem 24(c). Notice that the *outward* normal, relative to the upper half-plane, is in the *negative* η-direction.]

30. Suppose that the analytic function $F(z)$ maps the interior of a region R into the upper half-plane. Use the result of Problem 29(a) to show that the Green's function for $\nabla^2 w$ in the region R, relevant to the requirement that w vanish on the boundary of R, is of the form

$$G(x, y; \xi, \eta) = -\frac{1}{2\pi} \log \left| \frac{F(z) - F(\alpha)}{F(z) - \overline{F(\alpha)}} \right|,$$

where $z = x + iy$ and $\alpha = \xi + i\eta$. [When the boundary of R is *polygonal*, the function $F(z)$ can be determined by the *Schwarz-Christoffel* transformation.]

Section 4.5.

31. (a) In a certain linear system, the effect e at time t, due to a unit cause at a time τ, is a function only of the elapsed time $t - \tau$. If the system is inactive when $t < 0$, show that the cause-effect relation is a Volterra equation of the first kind, of the form

$$e(t) = \int_0^t K(t - \tau)c(\tau)\, d\tau.$$

[Equations of this form are considered in Section 4.13.]

(b) Show that the equation of part (a) can also be written in the form

$$e(t) = \int_0^t K(\tau)c(t - \tau)\, d\tau,$$

by replacing τ by $t - \tau$. [Notice that $K(\tau)$ can be considered as the effect at time τ due to a unit cause at time $t = 0$.]

(c) Noticing that the corresponding homogeneous Volterra equation of the second kind,

$$c(t) = \lambda \int_0^t K(t - \tau)c(\tau)\, d\tau,$$

expresses the requirement that the effect instantaneously reproduce the cause at all times, within a constant multiplicative factor, consider the possibility of existence of nontrivial solutions of such an equation.

32. Figure 4.6 is a schematic representation of an optical system in which a distribution of illumination emanating from a one-dimensional

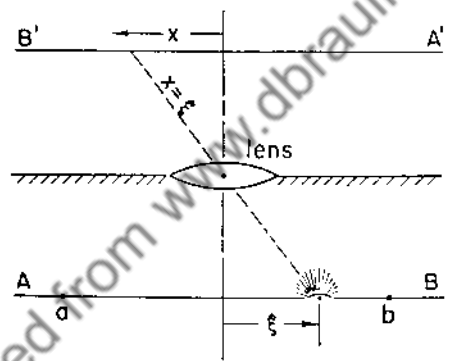


FIGURE 4.6

object along the line AB passes through a refracting lens and is projected into a one-dimensional (reversed) image along the line $A'B'$. With the notation of that figure, the light intensity at a point x, due to a unit source at ξ, is found to be a certain symmetrical function of the difference $x - \xi$, with a maximum at the point $x = \xi$, for a given lens.

(a) Formulate the problem of determining the object illumination distribution $I_o(\xi)$ over an interval $a \leqq \xi \leqq b$, corresponding to a prescribed image distribution $I_i(x)$ over the interval $a \leqq x \leqq b$, as an integral equation.

(b) Formulate the problem of determining those object distributions over $a \leqq \xi \leqq b$ which are magnified (and reversed), but not distorted, when projected on the line $A'B'$.

33. (a) If heat radiating from a unit point source is constrained to flow in a plane (as in a thin plate with insulated faces), show that the temperature T at a distance r from that point is given by

$$T = -\frac{1}{2\pi K}\log r + \text{constant},$$

where K is the thermal conductivity of the medium, in the absence of other sources or boundaries.

(b) Suppose that heat sources (and sinks) are continuously distributed along the closed boundary C of a region R in the xy-plane. Suppose also that the *net* heat supplied per second along the entire boundary is *zero*, so that temperatures at interior points of R do not change with time. Show that the steady-state temperature $T(x, y)$ at an interior point $P(x, y)$ can be expressed in the form

$$T(x, y) = -\frac{1}{2\pi K}\oint_C q(\sigma)\log r(x, y; \sigma)\, d\sigma + A,$$

where σ represents distance along C, from a reference point 0 to a point Q, at which point heat is being generated at the rate $q(\sigma)$ calories per second per unit length of arc, where $r(x, y; \sigma)$ represents the distance from P to Q (Figure 4.7), and where A is an undetermined constant. (Notice that the

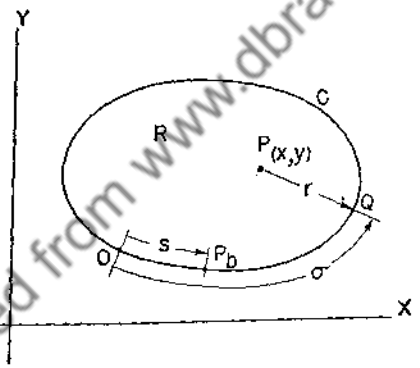


FIGURE 4.7

presence of A is in accordance with the fact that a *uniform* temperature distribution in R can be maintained without supplying heat along the boundary C.)

(c) Suppose that $q(\sigma)$ is not prescribed, but that the *temperature* of each boundary point P_b, at a distance s from 0 along C, is prescribed as $f(s)$. Denoting by $r(s; \sigma)$ the length of the chord $\overline{P_bQ}$, show that $q(\sigma)$ must satisfy the integral equation

$$f(s) = -\frac{1}{2\pi K}\oint_C q(\sigma)\log r(s; \sigma)\, d\sigma + A,$$

where the constant A is to be determined in such a way that $\oint_C q\, d\sigma = 0$.

34. By appropriately specializing the results of Problem 33, show that the steady-state temperature $T(\rho, \theta)$ at a point $x = \rho \cos \theta$, $y = \rho \sin \theta$, inside the boundary of a circular plate of radius a, with center at the origin (Figure 4.8), can be expressed in the form

$$T(\rho, \theta) = -\frac{1}{2\pi K} \int_0^{2\pi} q(\phi) \log \sqrt{\rho^2 - 2a\rho \cos(\theta - \phi) + a^2}\, d\phi + A,$$

where, if the temperature along the boundary is prescribed as $T(a, \theta) = f(\theta)$, the function $q(\phi)$ satisfies the integral equation

$$f(\theta) = -\frac{1}{4\pi K} \int_0^{2\pi} q(\phi) \log \sin^2 \frac{\theta - \phi}{2}\, d\phi + A,$$

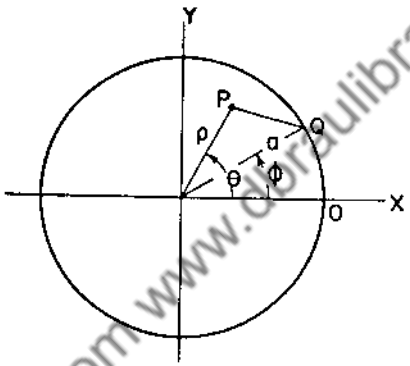


FIGURE 4.8

in which A is to be determined in such a way that the condition

$$\int_0^{2\pi} q(\phi)\, d\phi = 0$$

is satisfied.

By considering the result of integrating the equal members of the integral equation over $(0, 2\pi)$ with respect to θ [and using symmetry to show that the integral of the coefficient of $q(\phi)$ is independent of ϕ], show that A must then be taken as the mean value of the prescribed function $f(\theta)$ along the boundary,

$$A = \frac{1}{2\pi} \int_0^{2\pi} f(\theta)\, d\theta,$$

and verify further that A is also identified with the temperature $T(0, \theta)$ at the center of the circle. [In the case of a *circular* boundary, the Dirichlet problem can also be solved *directly* by use of the Poisson integral formula (see Problem 27). However, in the case of a general boundary, a formulation analogous to the preceding one (based on the results of Problem 33) is well adapted to numerical calculation.]

35. (a) If a unit heat source is present at a point Q, show that the rate of heat flow at a point P, per unit length of arc, across a curve C passing through P, is given by

$$-K\left.\frac{\partial T}{\partial n}\right|_P = \left.\frac{1}{2\pi}\frac{\partial}{\partial n}\log r\right|_P \equiv \left.\frac{1}{2\pi}\frac{\cos(n, r)}{r}\right|_P$$

where the differentiation is in the direction of the normal to C at the point P, where r denotes the distance from Q to P, and where (n, r) denotes the angle between QP and the normal.

(b) In the formulation of Problem 33, deduce that the net rate of heat flow *into* R across C at a boundary point P_b is given by

$$K\left.\frac{\partial T}{\partial n}\right|_{P_b} = \frac{1}{2}q(s) - \frac{1}{2\pi}\oint_C q(\sigma)\frac{\cos[n(\sigma), r(s;\sigma)]}{r(s;\sigma)}\,d\sigma,$$

where the angle involved in the integrand is that between the vector from the point Q to the point P_b and the *outward* normal at P_b. [The first term on the right corresponds to the fact that only one half of the heat generated by the source at P_b enters R; the second term subtracts the net rate of flow outward at P_b from the remaining sources along C. Notice that, in this formulation, when $K\,\partial T/\partial n$ is *prescribed* along C, the source distribution q is hence obtained as the solution of an integral equation of the *second* kind. Notice also that the right-hand member of the preceding equation is *not* the result of differentiating the right-hand member of the equation of Problem 33(b) under the integral sign, and evaluating the result at the boundary, but that the additional term $q(s)/2$ is present. *Inside* C, however, the integrand does not become infinite, and differentiation under the integral sign can then be justified, in general.]

(c) In the case of a *circular* boundary (see Problem 34), verify that the integral involved in part (b) vanishes, so that $q(\phi)/K$ is identified with *twice* the boundary value of $\partial T/\partial n$ in this case, and the first equation of Problem 34 then gives the solution of the *Neumann* problem for the circle explicitly, with A an arbitrary constant. [Show that the coefficient of $q(\sigma)$ in the integral of part (b) is then given by the value of

$$\frac{\partial}{\partial\rho}\log\sqrt{\rho^2 - 2a\rho\cos(\theta - \phi) + a^2}$$

when $\rho = a$, and that this expression has the *constant* value $1/(2a)$.]

Sections 4.6, 4.7.

36. (a) Show that the characteristic values of λ for the equation

$$y(x) = \lambda\int_0^{2\pi}\sin(x + \xi)\,y(\xi)\,d\xi$$

are $\lambda_1 = 1/\pi$ and $\lambda_2 = -1/\pi$, with corresponding characteristic functions of the form $y_1(x) = \sin x + \cos x$ and $y_2(x) = \sin x - \cos x$.

(b) Obtain the most general solution of the equation

$$y(x) = \lambda \int_0^{2\pi} \sin(x + \xi)\, y(\xi)\, d\xi + F(x)$$

when $F(x) = x$ and when $F(x) = 1$, under the assumption that $\lambda \neq \pm 1/\pi$.

(c) Prove that the equation

$$y(x) = \frac{1}{\pi} \int_0^{2\pi} \sin(x + \xi)\, y(\xi)\, d\xi + F(x)$$

possesses no solution when $F(x) = x$, but that it possesses infinitely many solutions when $F(x) = 1$. Determine all such solutions.

(d) Determine the most general form of the prescribed function $F(x)$, for which the integral equation

$$\int_0^{2\pi} \sin(x + \xi)\, y(\xi)\, d\xi = F(x),$$

of the first kind, possesses a solution.

37. Obtain an approximate solution of the integral equation

$$y(x) = \int_0^1 \sin(x\xi)\, y(\xi)\, d\xi + x^2,$$

by replacing $\sin(x\xi)$ by the first two terms of its power-series development

$$\sin(x\xi) = (x\xi) - \frac{(x\xi)^3}{3!} + \cdots.$$

Section 4.8.

38. Suppose that the kernel $K(x, \xi)$ of Section 4.8 is not necessarily symmetric, but is expressible in the form

$$K(x, \xi) = r(\xi)G(x, \xi),$$

where $r(\xi)$ is continuous in (a, b) and does not change sign in (a, b), and where $G(x, \xi)$ is symmetric. By appropriately modifying the treatments of that section, establish the following results:

(a) Two characteristic function $y_m(x)$ and $y_n(x)$, corresponding to distinct characteristic numbers λ_m and λ_n, are orthogonal over (a, b) with respect to the weighting function $r(x)$:

$$\int_a^b r(x) y_m(x) y_n(x)\, dx = 0.$$

(b) The characteristic numbers of $K(x, \xi)$ are all real.

(c) If equation (122) possesses a solution, then that solution is given by (130) or (130′), where the weighting function $r(x)$ is to be inserted in the integrands involved in (119) and in (131) or (131′).

(d) If equation (132) possesses a solution, then that solution is given by (144), where the weighting function $r(x)$ is to be inserted in the integrands involved in (119) and (140).

39. A complex kernel $K(x, \xi)$ [such as $e^{i(x-\xi)}$] which has the property

$$K(\xi, x) = \overline{K(x, \xi)},$$

where $\overline{K(x, \xi)}$ represents the complex conjugate of $K(x, \xi)$, is called a *Hermitian kernel*. By appropriately modifying the treatments of Section 4.8, establish the following results:

(a) Two characteristic functions $y_m(x)$ and $y_n(x)$, corresponding to distinct characteristic numbers λ_m and λ_n, are orthogonal over (a, b) in the Hermitian sense:

$$\int_a^b \overline{y_m(x)} y_n(x)\, dx = 0.$$

(b) The characteristic numbers associated with a Hermitian kernel over (a, b) are all real.

(c) Let the characteristic functions be normalized in the Hermitian sense:

$$\int_a^b \overline{\phi_n(x)} \phi_n(x)\, dx = 1.$$

Then, if equation (122) possesses a solution, that solution is given by (130), where (131) is replaced by the definition

$$f_n = \int_a^b \overline{\phi_n(x)} F(x)\, dx,$$

or by (130′), where (131′) is replaced by the definition

$$F_n \int_a^b \overline{y_n(x)} y_n(x)\, dx = \int_a^b \overline{y_n(x)} F(x)\, dx.$$

(d) If (132) possesses a solution, then that solution is given by (144), where f_n is defined in part (c).

40. If $u(x)$ is a characteristic function of a complex kernel $K(x, \xi)$ (which need not be Hermitian), corresponding to a characteristic number λ over (a, b), and $v(x)$ is a characteristic function of the transposed complex conjugate kernel $\overline{K(\xi, x)}$, corresponding to a characteristic number μ, show that

$$\int_a^b u(x) \overline{v(x)}\, dx = 0,$$

if $\lambda \neq \bar{\mu}$. [The kernel $\overline{K(\xi, x)}$ is called the *adjoint* of $K(x, \xi)$. Notice that real symmetric kernels and Hermitian kernels are *self-adjoint*.]

41. When λ is *nearly* equal to a characteristic number λ_m, to which there corresponds only one characteristic function, show that the solution of the equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi,$$

where $K(x, \xi)$ is symmetric, is given approximately by

$$y(x) \approx F(x) + \frac{\lambda}{\lambda_m - \lambda}\left(\int_a^b F\,\phi_m\,dx\right)\phi_m(x),$$

where $\phi_m(x)$ is the corresponding normalized characteristic function.

42. Assume that a symmetric kernel $K(x, \xi)$ can itself be expanded in a series of its orthogonalized and normalized characteristic functions,

$$K(x, \xi) = \sum_{n=1}^{\infty} a_n(\xi)\phi_n(x) \qquad (a \leqq x \leqq b),$$

where K is considered as a function of x for fixed values of ξ.

(a) Assuming also that term-by-term integration is permissible, show that the coefficient functions $a_n(\xi)$ must be of the form

$$a_n(\xi) = \frac{1}{\lambda_n}\,\phi_n(\xi),$$

where λ_n is the nth characteristic number, and hence obtain the so-called *bilinear expansion* of a symmetric kernel:

$$K(x, \xi) = \sum_{n=1}^{\infty} \frac{\phi_n(x)\phi_n(\xi)}{\lambda_n} \qquad (a \leqq x \leqq b).$$

It is known that this series converges (absolutely and uniformly) to $K(x, \xi)$ in (a, b) if $K(x, \xi)$ is continuous and symmetric, and if all except a finite number of its characteristic numbers are of the same sign. In most physically motivated problems, it is known in advance that all characteristic numbers are nonnegative, so that this result is then applicable.]

(b) Deduce that a continuous symmetric kernel which possesses only a finite number of linearly independent characteristic functions must be a *separable* kernel (in the sense of Section 4.6).

(c) Verify the result of part (a) in the case of the kernel involved in equation (88).

Section 4.9.

43. Consider the integral equation

$$y(x) = \lambda \int_0^1 x\xi\, y(\xi)\, d\xi + 1.$$

(a) Make use of equation (166) to show in advance that the iterative procedure of Section 4.9 will converge when $|\lambda| < 3$.

(b) Show that the iterative procedure leads formally to the expression

$$y(x) = 1 + x\left(\frac{\lambda}{2} + \frac{\lambda^2}{6} + \frac{\lambda^3}{18} + \cdots\right).$$

(c) Use the method of Section 4.6 to obtain the exact solution of the problem in the form

$$y(x) = 1 + \frac{3\lambda x}{2(3 - \lambda)} \qquad (\lambda \neq 3).$$

[Notice that the leading terms in the series expansion in powers of λ are given correctly by part (b), and that the estimate of part (a) happens to provide the actual convergence limit on λ.]

44. Deal with the integral equation

$$y(x) = \lambda \int_0^1 (x + \xi) y(\xi)\, d\xi + 1$$

as in Problem 43. Show also that the estimate afforded by (166) is slightly conservative in this case.

45. Show that, for sufficiently small values of $|\epsilon|$, an approximate solution of the equation

$$y(x) = \epsilon \int_0^a e^{-|x-\xi|} y(\xi)\, d\xi + 1$$

is afforded by the expression

$$y(x) \approx 1 + \epsilon\,[2 - e^{-x} - e^{-(a-x)}].$$

46. (a) Apply the method of Section 4.9 to the equation

$$y(x) = \int_0^x (x + \xi) y(\xi)\, d\xi + 1,$$

taking $y^{(0)}(x) = 1$, and obtaining the results of three successive substitutions.

(b) Show that the problem considered is equivalent to that specified by the differential equation $y'' - 2x\,y' - 3y = 0$ and the initial conditions $y(0) = 1$, $y'(0) = 0$.

Section 4.10.

47. Determine the resolvent kernel associated with $K(x, \xi) = x\xi$ in the interval (0, 1) in the form of a power series in λ, obtaining the first three terms.

48. Proceed as in Problem 47, for the kernel $K(x, \xi) = x + \xi$.

49. Determine the coefficient of λ in the expansion of the resolvent kernel associated with $K(x, \xi) = e^{-|x-\xi|}$ in the interval $(0, a)$.

50. Suppose that the resolvent kernel in (189) is assumed, as the *ratio* of two power series in λ, in the form

$$\Gamma(x, \xi; \lambda) = \frac{\sum_{n=0}^{\infty} N_n(x, \xi)\lambda^n}{\sum_{n=0}^{\infty} A_n \lambda^n}$$

(a) By replacing $y(x)$ by $F(x) + \lambda \int_a^b \Gamma(x, \xi; \lambda)F(\xi)\,d\xi$ in the basic equation (186), and clearing fractions, deduce formally that the relation

$$\sum_{n=0}^{\infty} \lambda^n \int_a^b N_n(x, \xi)F(\xi)\,d\xi - \left(\sum_{n=0}^{\infty} A_n\lambda^n\right) \int_a^b K(x, \xi)F(\xi)\,d\xi$$

$$- \sum_{n=0}^{\infty} \lambda^{n+1} \int_a^b \int_a^b K(x, \xi_1)N_n(\xi_1, \xi)F(\xi)\,d\xi_1\,d\xi = 0$$

must be satisfied identically in λ and F. [Assume that the order of integration and summation can be interchanged.]

(b) Show that this requirement implies the relation

$$N_n(x, \xi) = A_nK(x, \xi) + \int_a^b K(x, \xi_1)N_{n-1}(\xi_1, \xi)\,d\xi_1 \qquad (n = 1, 2, \cdots),$$

together with the initial condition

$$N_0(x, \xi) = A_0K(x, \xi).$$

(c) Verify that the preceding determination leads to the resolvent kernel defined by equation (188) if we take $A_0 = 1$ and $A_n = 0$ when $n \geqq 1$. [Whereas this series converges only for $|\lambda| < |\lambda_1|$, the constants A_n can be so chosen that the numerator and denominator series both converge *for all values of* λ. This result, which is discussed in Section 4.11, is of basic importance in the theory of linear integral equations.]

Section 4.11.

51. Verify that the Fredholm form (192) of the resolvent kernel satisfies the requirement derived in Problem 50(b). [Here $N_n(x, \xi) = (-1)^nD_n(x, \xi)/n!$ and $A_n = (-1)^nC_n/n!$.]

52. Obtain the resolvent kernel associated with $K(x, \xi) = x\xi$ in $(0, 1)$, by the method of Section 4.11.

53. Proceed as in Problem 52 for the kernel $K(x, \xi) = x + \xi$ in $(0, 1)$.

54. (a) Obtain the solution of the equation

$$y(x) = \lambda \int_a^b K(x, \xi)y(\xi)\,d\xi + F(x),$$

where

$$K(x, \xi) = u(x)v(\xi),$$

in the form

$$y(x) = F(x) + \lambda \frac{\displaystyle\int_a^b K(x, \xi)F(\xi)\,d\xi}{1 - \lambda \displaystyle\int_a^b K(x, x)\,dx},$$

by the methods of Section 4.6. [Write $y(x) = \lambda c\, u(x) + F(x)$, where $c = \int_a^b v(\xi)y(\xi)\, d\xi$, determine c, and identify the result with the form given.]

(b) Obtain the same result by the methods of Section 4.11. [Notice that the single characteristic number is $\lambda_1 = 1/\int_a^b K(x, x)\, dx$.]

Section 4.12.

55. Show formally that the Green's function $G(x, \xi)$ associated with the expression $\frac{d^2y}{dx^2} - y$ over the *infinite* interval $(-\infty, \infty)$, subject to the requirement that y remain finite as $x \to \pm\infty$, is of the form

$$G(x, \xi) = \tfrac{1}{2}e^{-|x-\xi|}.$$

56. (a) If $I(x) = \int_{-\infty}^{\infty} e^{-|x-\xi|}\Phi(\xi)\, d\xi$, verify that $I''(x) = I(x) - 2\Phi(x)$ for any continuous function $\Phi(x)$ which is dominated by $e^{-|x|}$ as $x \to \pm\infty$.

(b) Use this result to show that any continuous solution of the integral equation

$$y(x) = \lambda \int_{-\infty}^{\infty} e^{-|x-\xi|}y(\xi)\, d\xi + F(x)$$

must also satisfy the differential equation

$$y''(x) - (1 - 2\lambda)y(x) = F''(x) - F(x).$$

57. Suppose that $F(x) \equiv 0$ in Problem 56.

(a) In the case when $1 - 2\lambda$ is positive, and hence we may write $1 - 2\lambda = \alpha^2$ or $\lambda = (1 - \alpha^2)/2$, show that the general solution,

$$y(x) = c_1e^{\alpha x} + c_2e^{-\alpha x},$$

of the homogeneous differential equation, also satisfies the homogeneous integral equation when and only when $|\alpha| < 1$. Deduce that any *positive* value of λ of the form $\lambda = (1 - \alpha^2)/2$ (and hence any λ between 0 and $\frac{1}{2}$) is a characteristic number of multiplicity two, with the characteristic functions $e^{\pm\alpha x}$.

(b) In the case when $\lambda = \frac{1}{2}$, verify that the solution $y(x) = c_1 + c_2x$ satisfies the integral equation, so that $\lambda = \frac{1}{2}$ is also a double characteristic number, with characteristic functions 1 and x.

(c) In the case when $1 - 2\lambda$ is negative, and hence we may write $\lambda = (1 + \beta^2)/2$, show similarly that any such value of λ (and hence any $\lambda > \frac{1}{2}$) is again a double characteristic number, corresponding to $y = \cos\beta x$ and $y = \sin\beta x$. Thus deduce that all positive values of λ, and only those values, are characteristic numbers.

(d) If *only functions which remain finite as* $x \to \pm\infty$ *are accepted as characteristic functions*, deduce that only those values of λ for which $\lambda \geqq \frac{1}{2}$ are then characteristic numbers, that the number $\lambda = \frac{1}{2}$ is then of multiplicity one, and that all values of λ greater than $\frac{1}{2}$ are of multiplicity two.

58. Suppose that $F(x) = \sin \mu x$ in Problem 56.

(a) Show that the differential equation admits a *bounded* solution if and only if $\lambda \neq (1 + \mu^2)/2$, where $\mu > 0$, and that this solution must be of the *unique* form

$$y(x) = \frac{1 + \mu^2}{1 + \mu^2 - 2\lambda} \sin \mu x$$

when $\lambda < \frac{1}{2}$. Verify that this expression also satisfies the integral equation for all values of λ such that $\lambda \neq (1 + \mu^2)/2$.

(b) Deduce that the integral equation

$$y(x) = \lambda \int_{-\infty}^{\infty} e^{-|x-\xi|} y(\xi)\, d\xi + \sin \mu x$$

possesses a unique bounded continuous solution

$$y(x) = \frac{1 + \mu^2}{1 + \mu^2 - 2\lambda} \sin \mu x$$

when $\lambda < \frac{1}{2}$, a single infinity of such solutions,

$$y(x) = \frac{1 + \mu^2}{\mu^2} \sin \mu x + c_1,$$

when $\lambda = \frac{1}{2}$, and a double infinity of such solutions,

$$y(x) = \frac{1 + \mu^2}{1 + \mu^2 - 2\lambda} \sin \mu x + c_1 \cos (\sqrt{2\lambda - 1}\, x) + c_2 \sin (\sqrt{2\lambda - 1}\, x),$$

when $\lambda > \frac{1}{2}$ but $\lambda \neq (1 + \mu^2)/2$, and that it possesses *no* continuous solution (bounded or otherwise) when $\lambda = (1 + \mu^2)/2$. [Notice that, whereas all values of $\lambda \geqq \frac{1}{2}$ are characteristic numbers, so that the *homogeneous* equation is then solvable, the *nonhomogeneous* equation here is *also* solvable unless λ coincides with a *particular* characteristic number which depends upon the right-hand member, rather than on the kernel itself. This is in contrast with the situation relevant to nonsingular equations, where (for a symmetric kernel) solvability of the *homogeneous* equation precludes solvability of the *nonhomogeneous* equation unless $F(x)$ is orthogonal to the corresponding homogeneous solutions.]

59. (a) Show that the equation

$$y(x) = \lambda \int_{-\infty}^{\infty} K(|\, x - \xi\, |) y(\xi)\, d\xi,$$

in which the kernel is a function only of $|x - \xi|$, can be written in the form

$$y(x) = \lambda \int_0^\infty K(u)[y(x + u) + y(x - u)]\, du.$$

(b) Show that this equation is satisfied by

$$y(x) = c_1 \cos \beta x + c_2 \sin \beta x.$$

if λ is of the form

$$\lambda = \frac{1}{2\int_0^\infty K(u) \cos \beta u\, du}.$$

Deduce that all values of β for which the integral involved exists give rise to characteristic values of λ, for which solutions exist which are bounded as $x \to \pm\infty$.

(c) In the special case $K(|x - \xi|) = e^{-|x-\xi|}$, considered in Problems 56 to 58, show that the preceding result is in accordance with the result of Problem 57(d).

Section 4.13.

60. Obtain the solution of the generalized Abel integral equation

$$F(x) = \int_0^x \frac{y(\xi)}{(x - \xi)^\alpha} d\xi \qquad (0 < \alpha < 1)$$

in the form

$$y(x) = \frac{\sin \alpha\pi}{\pi} \frac{d}{dx} \int_0^x \frac{F(\xi)\, d\xi}{(x - \xi)^{1-\alpha}},$$

by dividing both sides of the given equation by $(s - x)^{1-\alpha}$ and proceeding as in Section 4.13(d). [Notice that $\int_0^1 t^{-\alpha}(1 - t)^{\alpha-1}\, dt = \Gamma(1 - \alpha)\Gamma(\alpha) = \pi/\sin \alpha\pi$ when $0 < \alpha < 1$.]

61. Let the Laplace transform of a function $f(x)$ be considered as a function of a new independent variable p, so that

$$\mathcal{L}\, f(x) \equiv F(p) = \int_0^\infty e^{-px} f(x)\, dx.$$

(a) By making the substitution $p\,x = u$ in the integral defining the transform, show that

$$\mathcal{L}\, x^{-\alpha} = \Gamma(1 - \alpha)p^{\alpha-1}, \qquad p^{-\alpha} = \frac{1}{\Gamma(\alpha)} \mathcal{L}\, x^{\alpha-1},$$

when $0 < \alpha < 1$.

(b) Establish the property

$$\frac{1}{p} \mathcal{L}\, f(x) = \mathcal{L} \int_0^x f(\xi)\, d\xi.$$

62. By making use of the results of Problem 61, and of the convolution property

$$\mathfrak{L}\int_0^x \phi_1(x-\xi)\phi_2(\xi)\,d\xi = \mathfrak{L}\,\phi_1(x)\mathfrak{L}\,\phi_2(x),$$

verify that the solution of Abel's integral equation,

$$F(x) = \int_0^x (x-\xi)^{-\frac{1}{2}} y(\xi)\,d\xi,$$

can be obtained by the following steps:

$$\mathfrak{L}\,F(x) = \mathfrak{L}\,x^{-\frac{1}{2}}\mathfrak{L}\,y(x); \quad \mathfrak{L}\,y(x) = \frac{1}{\Gamma(\frac{1}{2})}\,p^{\frac{1}{2}}\mathfrak{L}\,F(x);$$

$$\frac{1}{p}\,\mathfrak{L}\,y(x) = \frac{1}{[\Gamma(\frac{1}{2})]^2}\,\mathfrak{L}\,x^{-\frac{1}{2}}\mathfrak{L}\,F(x);$$

$$\int_0^x y(\xi)\,d\xi = \frac{1}{\pi}\int_0^x (x-\xi)^{-\frac{1}{2}}F(\xi)\,d\xi;$$

$$y(x) = \frac{1}{\pi}\frac{d}{dx}\int_0^x \frac{F(\xi)}{\sqrt{x-\xi}}\,d\xi.$$

63. Obtain the solution of Problem 60 by the method of Problem 62.

64. (a) If y satisfies the equation

$$F(x) = \int_0^x H(x-\xi)y(\xi)\,d\xi,$$

show that y also satisfies the equation

$$H(0)y(x) = F'(x) - \int_0^x H'(x-\xi)y(\xi)\,d\xi,$$

and that $Y(x) \equiv \int_0^x y(\xi)\,d\xi$ satisfies the equation

$$H(0)Y(x) = F(x) - \int_0^x H'(x-\xi)Y(\xi)\,d\xi.$$

(b) By formally integrating the equal members of the first equation of part (a) over $(0, \infty)$ with respect to x, and formally interchanging the order of integration in the resultant right-hand member, obtain the relation

$$\int_0^\infty F(x)\,dx = \int_0^\infty H(u)\,du \int_0^\infty y(\xi)\,d\xi.$$

65. Suppose that $y(x)$ satisfies the equation

$$y(x) = F(x) + \lambda\int_a^b K(x,\xi)y(\xi)\,d\xi,$$

where $K(x, \xi)$ is continuous, but not necessarily symmetric.

(a) By multiplying the equal members of this equation by $K(x, t)$, and integrating the results over (a, b) with respect to x, deduce that there follows also

$$0 = \int_a^b K(\xi, x)y(\xi)\,d\xi - \int_a^b K(t, x)F(t)\,dt - \lambda \int_a^b \int_a^b K(t, x)K(t, \xi)y(\xi)\,dt\,d\xi,$$

after an appropriate relettering of variables.

(b) By multiplying the equal members of this equation by λ, and adding the results to the corresponding equal members of the original equation, show that any solution of that equation also satisfies the equation

$$y(x) = f(x) + \lambda \int_a^b K_s(x, \xi)y(\xi)\,d\xi,$$

where

$$f(x) = F(x) - \lambda \int_a^b K(t, x)F(t)\,dt,$$

and where $K_s(x, \xi)$ is the *symmetric* kernel

$$K_s(x, \xi) = K(x, \xi) + K(\xi, x) - \lambda \int_a^b K(t, x)K(t, \xi)\,dt.$$

66. Suppose that $y(x)$ satisfies the equation

$$F(x) = \int_a^b K(x, \xi)y(\xi)\,d\xi,$$

where $K(x, \xi)$ is not necessarily symmetric.

(a) By multiplying the equal members by $K(x, t)$, and integrating the results over (a, b) with respect to x, show that $y(x)$ also satisfies the equation

$$F_{\mathrm{I}}(x) = \int_a^b K_{\mathrm{I}}(x, \xi)y(\xi)\,d\xi,$$

where

$$F_{\mathrm{I}}(x) = \int_a^b K(\xi, x)F(\xi)\,d\xi$$

and where $K_{\mathrm{I}}(x, \xi)$ is the *symmetric* kernel

$$K_{\mathrm{I}}(x, \xi) = \int_a^b K(t, x)K(t, \xi)\,dt.$$

(b) In a similar way, show that $y(x)$ also satisfies the equation

$$F_{\mathrm{II}}(x) = \int_a^b K_{\mathrm{II}}(x, \xi)y(\xi)\,d\xi,$$

where

$$F_{\mathrm{II}}(x) = \int_a^b K(x, \xi)F(\xi)\,d\xi$$

and

$$K_{\mathrm{II}}(x, \xi) = \int_a^b K(x, t)K(\xi, t)\,dt.$$

[Notice that the methods of Section 4.8 (pages 418–421) are applicable to the new equations.]

67. Let λ_n and $u_n(x)$ denote corresponding characteristic quantities for the symmetric kernel

$$K_1(x, \xi) = \int_a^b K(t, x)K(t, \xi)\, dt$$

over (a, b), so that

$$u_n(x) = \lambda_n \int_a^b K_1(x, \xi)u_n(\xi)\, d\xi.$$

(a) By replacing K_1 by its definition, and interchanging the order of integration in the resultant double integral, show that there follows

$$u_n(x) = \lambda_n \int_a^b K(t, x)\, V_n(t)\, dt,$$

where

$$V_n(x) = \int_a^b K(x, \xi)u_n(\xi)\, d\xi.$$

(b) By multiplying both members of the integral equation by $u_n(x)$, and integrating the results over (a, b), show that

$$\int_a^b [u_n(x)]^2\, dx = \lambda_n \int_a^b [V_n(x)]^2\, dx,$$

and hence deduce that *the characteristic numbers* λ_n *of* $K_1(x, \xi)$ *are positive.*

(c) By writing $v_n(x) = \sqrt{\lambda_n}\, V_n(x)$, show that the equations of part (a) can be expressed in the symmetrical form

$$u_n(x) = \sqrt{\lambda_n} \int_a^b K(\xi, x)v_n(\xi)\, d\xi,$$

$$v_n(x) = \sqrt{\lambda_n} \int_a^b K(x, \xi)u_n(\xi)\, d\xi,$$

and that there then follows also

$$\int_a^b u_n{}^2\, dx = \int_a^b v_n{}^2\, dx.$$

(d) Show that the result of introducing the first relation of part (c) into the right-hand member of the second one is of the form

$$v_n(x) = \lambda_n \int_a^b K_{II}(x, \xi)v_n(\xi)\, d\xi,$$

where

$$K_{II}(x, \xi) = \int_a^b K(x, t)K(\xi, t)\, dt.$$

Hence deduce that *the auxiliary kernels* K_I *and* K_{II}, *associated with a non-symmetric kernel* $K(x, \xi)$, *possess the same characteristic numbers*, that *these*

numbers are all positive, and that *the respective characteristic functions* $u_n(x)$ *and* $v_n(x)$ *corresponding to* λ_n *satisfy the simultaneous equations of part* (c). [Notice that the two sets of characteristic functions thus associated with a nonsymmetric kernel are *orthogonal* sets, and that, by putting the equations of part (c) in a symmetrical form, we have ensured the fact that one of the sets is *normalized* when the other set has this property. Notice that the characteristic numbers λ_n are *not*, in general, characteristic numbers of the kernel $K(x, \xi)$ itself.]

68. The *Cauchy principal value* of an integral, $\oint_a^b f(x)\,dx$, in which $f(x)$ becomes infinite at a point $x = c$ inside the range of integration, is defined as the limit

$$\oint_a^b f(x)\,dx = \lim_{\epsilon \to 0} \left[\int_a^{c-\epsilon} f(x)\,dx + \int_{c+\epsilon}^b f(x)\,dx \right],$$

when that limit exists. When the *separate* limits on the right exist, the integral is convergent in the strict sense and the symbol $\oint$ may be replaced by the usual symbol $\int$.

(a) Verify that

$$\oint_{-1}^2 \frac{dx}{x} = \log 2,$$

but that $\int_{-1}^2 \frac{dx}{x}$ does not exist. [Recall that $\int \frac{dx}{x} = \log |x| + C$.]

(b) Verify that neither $\int_{-1}^1 \frac{dx}{x^2}$ nor $\oint_{-1}^1 \frac{dx}{x^2}$ exists.

69. The Hilbert integral representation of a suitably regular function $g(\theta)$, over the interval $0 < \theta < \pi$, is of the form

$$g(\theta) = \frac{1}{\pi^2} \oint_0^\pi \oint_0^\pi K(\theta, \phi_1) K(\phi, \phi_1) g(\phi)\,d\phi_1\,d\phi + \frac{1}{\pi} \int_0^\pi g(\phi)\,d\phi$$

where

$$K(\theta, \phi) = \frac{\sin \phi}{\cos \phi - \cos \theta}.$$

[The Cauchy principal values are necessary because of the strong singularity of $K(\theta, \phi)$ when $\phi = \theta$.] Show that, if we write

$$f(\phi_1) = \frac{1}{\pi} \oint_0^\pi K(\phi, \phi_1) g(\phi)\,d\phi,$$

there follows

$$g(\theta) = \frac{1}{\pi} \oint_0^\pi K(\theta, \phi_1) f(\phi_1)\,d\phi_1 + \frac{1}{\pi} \int_0^\pi g(\phi)\,d\phi,$$

and hence deduce the validity of the simultaneous equations

$$f(\theta) = \frac{1}{\pi} \oint_0^\pi g(\phi) \frac{\sin \theta}{\cos \theta - \cos \phi} d\phi,$$

$$g(\theta) = -\frac{1}{\pi} \oint_0^\pi f(\phi) \frac{\sin \phi}{\cos \phi - \cos \theta} d\phi + \frac{1}{\pi} \int_0^\pi g(\phi)\, d\phi.$$

[The function f is often called the *Hilbert transform* of g over $(0, \pi)$.]

70. (a) Verify, by direct calculation, that

$$\oint_0^\pi \frac{d\phi}{\cos \theta - \cos \phi} = 0.$$

(b) Deduce from equation (51) of Chapter 3 that

$$\oint_0^\pi \frac{\cos n\phi}{\cos \theta - \cos \phi} d\phi = -\pi \frac{\sin n\theta}{\sin \theta} \qquad (n = 0, 1, 2, \cdots).$$

(c) Let a function $g(\theta)$ be defined by the Fourier cosine series

$$g(\theta) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\theta \qquad (0 < \theta < \pi).$$

Assuming the validity of a certain interchange of order of summation and integration, deduce that the Hilbert transform of $g(\theta)$ is of the form

$$f(\theta) = -\sum_{n=1}^{\infty} a_n \sin n\theta \qquad (0 < \theta < \pi).$$

[In particular, notice that the Hilbert transform of a constant is zero.]

(d) If $f(\theta)$ is defined by the preceding equation, verify that $g(\theta)$ is given correctly by the last equation of Problem 69. [Express the product of two sines as the difference between two cosines.]

71. (a) By making the change in variables

$$\cos \theta = x, \quad \cos \phi = \xi,$$

and writing

$$\frac{f(\cos^{-1} x)}{\sqrt{1 - x^2}} = F(x), \quad \frac{g(\cos^{-1} x)}{\sqrt{1 - x^2}} = y(x),$$

in the relations of Problem 69, show that the solution of the singular integral equation

$$F(x) = \frac{1}{\pi} \oint_{-1}^{1} y(\xi) \frac{d\xi}{\xi - x} \qquad (-1 < x < 1)$$

is defined by the relation

$$\sqrt{1-x^2}\,y(x) = \frac{1}{\pi}\oint_{-1}^{1}\sqrt{1-\xi^2}\,F(\xi)\,\frac{d\xi}{x-\xi} + \frac{1}{\pi}\int_{-1}^{1} y(\xi)\,d\xi.$$

[Notice that the solution is not unique, since the value of the integral $\int_{-1}^{1} y(\xi)\,d\xi$ may be prescribed.]

(b) If it is required that $y(-1)$ be finite, show that there must follow

$$\int_{-1}^{1} y(\xi)\,d\xi = \int_{-1}^{1}\sqrt{1-\xi^2}\,F(\xi)\,\frac{d\xi}{1+\xi}$$

and verify that the solution of part (a) can then be written in the form

$$y(x) = \frac{1}{\pi}\sqrt{\frac{1+x}{1-x}}\oint_{-1}^{1}\sqrt{\frac{1-\xi}{1+\xi}}\,F(\xi)\,\frac{d\xi}{x-\xi}.$$

[These results are of importance, for example, in the aerodynamic theory of airfoils.]

Section 4.14.

72. Obtain two successive approximations to the smallest characteristic number and the corresponding characteristic function of the problem

$$y(x) = \lambda\int_0^1 K(x,\,\xi)y(\xi)\,d\xi \quad \text{where } K(x,\,\xi) = \begin{cases} x, & x < \xi, \\ \xi, & x > \xi, \end{cases}$$

starting with the initial approximation $y^{(1)} = 1$. [Show that $f^{(1)} = \frac{1}{2}(2x - x^2)$, and that equations (241a,b,c,) give the estimates $\lambda_1^{(1)} = 3$, 3, and 2.5, respectively. With $y^{(2)} = 2x - x^2$, show that there follows $f^{(2)} = \frac{1}{12}(8x - 4x^3 + x^4)$, and that (241a,b) give the respective estimates $\lambda_1^{(2)} = 2.5$ and 2.471. (Equation (241c) gives the estimate 2.4677.) By plotting the functions $y^{(2)}(x)$ and $\lambda_1^{(2)}f^{(2)}(x)$, show also that the difference between the input and output in the second cycle is less than 3 per cent of the maximum value.]

73. Show that the integral equation of Problem 72 is equivalent to the problem $y'' + \lambda y = 0$, $y(0) = y'(1) = 0$, and deduce that the exact characteristic numbers are of the form $\frac{1}{4}(2n-1)^2\pi^2$, with corresponding characteristic functions proportional to $\sin\frac{1}{2}(2n-1)\pi x$. [Notice that $\lambda_1 = \pi^2/4 = 2.4674$.]

74. Obtain an approximation to the second characteristic number and corresponding characteristic function of Problem 72, assuming (for simplicity) that the fundamental characteristic function is given with sufficient accuracy by $y_1(x) = 2x - x^2$, and taking $F(x) = x$ in equation (242). [Show that, neglecting an arbitrary multiplicative factor in $y^{(1)}$,

we may take $y^{(1)} = 25x^2 - 18x$, and so obtain $f^{(1)} = \frac{1}{12}(-8x + 36x^3 - 25x^4)$. Show also that equation (241a) here fails to estimate $\lambda_2^{(1)}$, whereas (241b) gives $\lambda_2^{(1)} = 23.3$. (Equation (241c) would give a more nearly accurate result.) Notice that the true value is $\lambda_2 = 9\pi^2/4 = 22.2$, in accordance with the result of Problem 73.]

75. Let λ_n denote a characteristic number associated with $K(x, \xi)$ over (a, b), and denote by $\phi_n(x)$ a corresponding *normalized* characteristic function, so that

$$\phi_n(x) = \lambda_n \int_a^b K(x, \xi)\phi_n(\xi)\, d\xi$$

and

$$\int_a^b [\phi_n(x)]^2\, dx = 1.$$

(a) By multiplying both members of the first relation by $\phi_n(x)$, integrating over (a, b), and using the normalizing condition, deduce that

$$\lambda_n = \frac{1}{\int_a^b \int_a^b K(x, \xi)\phi_n(x)\phi_n(\xi)\, dx\, d\xi}.$$

(b) By making use of the Schwarz inequality (Problem 87 of Chapter 1), deduce that

$$\frac{1}{|\lambda_n|} \leqq \sqrt{\int_a^b \int_a^b [K(x, \xi)]^2\, dx\, d\xi} \equiv A,$$

so that $|\lambda_n| \leqq 1/A$. [This result establishes equation (166).]

76. If $K(x, \xi)$ is a *symmetric* kernel, and $\mathcal{K}$ is the integral operator such that

$$\mathcal{K}f(x) \equiv \int_a^b K(x, \xi)f(\xi)\, d\xi,$$

the form

$$J(y) \equiv \int_a^b y(x)\, \mathcal{K}\, y(x)\, dx \equiv \int_a^b \int_a^b K(x, \xi)y(x)y(\xi)\, dx\, d\xi$$

is known as the *quadratic integral form* associated with the operator $\mathcal{K}$. [Notice that this definition is in complete analogy with the definition of the quadratic form associated with a symmetric *matrix* $\mathbf{a}$, as the scalar product of a vector $\mathbf{x}$ and its transform $\mathbf{a}\,\mathbf{x}$ (see Section 1.13).]

(a) Verify that the variation of the functional $J(y)$ can be expressed in the form

$$\delta J = 2 \int_a^b \left[\int_a^b K(x, \xi)y(\xi)\, d\xi \right] \delta y(x)\, dx,$$

with the notation of Section 2.4.

(b) Deduce that the requirement that the expression

$$I(y) \equiv \lambda \int_a^b \int_a^b K(x, \xi)y(x)y(\xi)\, dx\, d\xi - \int_a^b [y(x)]^2\, dx + 2 \int_a^b y(x)F(x)\, dx$$

be stationary, for arbitrary small continuous variations in the function $y(x)$, leads to the requirement that y satisfy the integral equation

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi + F(x)$$
$$\equiv \lambda \mathcal{K}\, y(x) + F(x).$$

77. Deduce from Problem 76 that the characteristic numbers of the operator $\mathcal{K}$ are stationary values of the ratio

$$\lambda = \frac{\int_a^b [y(x)]^2\, dx}{\int_a^b \int_a^b K(x, \xi) y(x) y(\xi)\, dx\, d\xi} \equiv \frac{\int_a^b y^2\, dx}{J(y)},$$

or of the ratio

$$\lambda = \frac{1}{\int_a^b \int_a^b K(x, \xi) \phi(x) \phi(\xi) dx\, d\xi} \equiv \frac{1}{J(\phi)},$$

where ϕ is subject to the constraint (normalizing condition)

$$\int_a^b \phi^2\, dx = 1,$$

and that the extremals are the corresponding characteristic functions. [Compare equation (69) of Chapter 2, and Problems 33 and 75 of that chapter. The *smallest* stationary value can be shown to be a *minimum*. Hence, of *all* continuous functions ϕ for which $\int_a^b \phi^2\, dx = 1$, that function ϕ_1 which *maximizes* $J(\phi)$ is the characteristic function corresponding to λ_1, where $1/\lambda_1 \geqq 1/\lambda_2 \geqq \ldots \geqq 1/\lambda_k \geqq \ldots$, and $\lambda_1 = 1/J(\phi_1)$. Of all functions ϕ which are normalized and which are *also* orthogonal to ϕ_1, that function ϕ_2 which maximizes $J(\phi)$ is the *second* characteristic function, corresponding to $\lambda_2 = 1/J(\phi_2)$, and so forth. (Compare Problem 77 of Chapter 1, noticing that λ here is analogous to $1/\lambda$ in Chapter 1.)]

78. (a) Noticing that $f^{(n)}(x) = \int_a^b K(x, \xi) y^{(n)}(\xi)\, d\xi$, verify that equation (241b) can be obtained by replacing $y(x)$ in the ratio of Problem 77 by an approximation $y^{(n)}(x)$ to a characteristic function.

(b) Recalling also that a constant multiple of the output $f^{(n)}(x)$ is taken as the input $y^{(n+1)}(x)$ in the following cycle, verify that equation (241c) can be written in the form

$$\lambda_1 \approx \frac{\int_a^b y^{(n+1)}(x) y^{(n)}(x)\, dx}{\int_a^b \int_a^b K(x, \xi)\, y^{(n+1)}(x)\, y^{(n)}(\xi)\, dx\, d\xi},$$

and so verify that the estimate afforded by (241c) corresponds to that obtained by replacing y^2 by $y^{(n)} y^{(n+1)}$ in the ratio of Problem 77. [Notice

that the estimates of (241b,c) are obtained without the necessity of explicitly evaluating a double integral. A still better approximation (but one requiring such an evaluation) would be obtained by replacing y by $f^{(n)}$ in the ratio of Problem 77, to give $\lambda_1 \approx \dfrac{1}{J(f_n)} \displaystyle\int_a^b |f^{(n)}(x)|^2\,dx$.

79. An operator $\mathcal{K}$, associated with a (real) *symmetric* kernel $K(x, \xi)$, whose quadratic integral form

$$J(y) \equiv \int_a^b y(x)\,\mathcal{K}\,y(x)\,dx \equiv \int_a^b \int_a^b K(x, \xi)y(x)y(\xi)\,dx\,d\xi$$

is *nonnegative* for any (real) function y, is called a *positive* integral operator. If, in addition, $J(y)$ is zero only when $y(x)$ is the zero function, then $\mathcal{K}$ is said to be *positive definite*.

(a) Use the result of Problem 77 (or of Problem 75) to show that *the characteristic numbers corresponding to a positive integral operator are positive.*

(b) Show that the operators involving the kernels

$$K_{\mathrm{I}}(x, \xi) = \int_a^b K(t, x)K(t, \xi)\,dt, \qquad K_{\mathrm{II}}(x, \xi) = \int_a^b K(x, t)K(\xi, t)\,dt,$$

associated with a nonsymmetric kernel $K(x, \xi)$, are positive operators. [In the case of K_{I}, show that $J(y) = \displaystyle\int_a^b \left\{\int_a^b K(t, x)y(x)\,dx\right\}^2 dt$.]

80. A kernel which has the property

$$K(\xi, x) = -K(x, \xi)$$

is called a *skew-symmetric* kernel.

(a) If $K(x, \xi)$ is skew symmetric, show that the equation

$$y(x) = \lambda \int_a^b K(x, \xi)y(\xi)\,d\xi$$

implies the equation

$$y(x) = -\lambda^2 \int_a^b K_1(x, \xi)y(\xi)\,d\xi.$$

(b) Use the result of Problem 79(b) to deduce that a *skew-symmetric kernel possesses no real characteristic numbers.* [Notice that a nonhomogeneous equation with a continuous skew-symmetric kernel (over a finite interval) therefore possesses a solution for any continuous prescribed function $F(x)$.]

(c) Verify this conclusion in the special case of the kernel $K(x, \xi) = x - \xi$, associated with the interval $(0, 1)$.

Section 4.15.

81. Obtain approximate values of the solution of the equation

$$\frac{1}{2}(x - x^2) = \int_0^1 K(x, \xi)y(\xi)\,d\xi,$$

where $K(x, \xi)$ is defined by equation (258), at the points $x = 0$, 0.25, 0.5, 0.75, and 1, by the methods of Section 4.15. Use the weighting coefficients of the trapezoidal rule.

82. With the notation of Section 4.15, show that approximate characteristic numbers relevant to the problem

$$y(x) = \lambda \int_a^b K(x, \xi) y(\xi)\, d\xi$$

are afforded by *reciprocals* of the characteristic numbers relevant to the corresponding matrix $\mathbf{K D}$.

83. Determine approximately the smallest characteristic value of λ for the problem

$$y(x) = \lambda \int_0^1 e^{-x\xi}\, y(\xi)\, d\xi,$$

by the method of Problem 82, using the ordinates at the points $x = 0$, 0.5, and 1, and the weighting coefficients of Simpson's rule. Determine also the approximate ratios of the ordinates of the corresponding characteristic function at those points. [Write $\kappa = 6/\lambda$, and use the iterative methods of Chapter 1, retaining three significant figures.]

Section 4.16.

84. (a) If $K(x, \xi)$ is of the form

$$K(x, \xi) = \begin{cases} x & \text{when } x < \xi, \\ \xi & \text{when } x > \xi, \end{cases}$$

show that the assumption $y(x) \approx A_1 + A_2 x + A_3 x^2$ reduces the equation

$$y(x) = F(x) + \lambda \int_0^1 K(x, \xi) y(\xi)\, d\xi \qquad (0 < x < 1)$$

to the requirement

$$A_1[1 - \lambda(x - \tfrac{1}{2}x^2)] + A_2[x - \tfrac{1}{2}\lambda(x - \tfrac{1}{3}x^3)] + A_3[x^2 - \tfrac{1}{3}\lambda(x - \tfrac{1}{4}x^4)] \approx F(x) \quad (0 < x < 1).$$

(b) Show that the problem of part (a) can be reduced to the problem

$$y''(x) + \lambda\, y(x) = F''(x), \quad y(0) = F(0), \quad y'(1) = F'(1).$$

85. (a) Show that the assumption

$$y(\theta) \approx \sum_k A_k \sin k\theta \qquad (0 < \theta < \pi)$$

reduces the *integro-differential* equation

$$y(\theta) = F(\theta) + \lambda f(\theta) \oint_0^\pi \frac{dy(\phi)}{d\phi} \frac{d\phi}{\cos\theta - \cos\phi} \qquad (0 < \theta < \pi)$$

to the requirement

$$\sum_k A_k \left[1 + \lambda \pi k \frac{f(\theta)}{\sin \theta} \right] \sin k\theta \approx F(\theta) \qquad (0 < \theta < \pi).$$

[The symbol $\mathscr{f}$ denotes the Cauchy principal value. See Problems 68 and 70(b). With a suitable interpretation of the symbols, this equation is one form of the basic equation of the *Prandtl lifting-line* theory of aerodynamics.]

(b) In the special case when $f(\theta) = \sin \theta$, show that the formal solution of the equation of part (a) is given by

$$y(\theta) = \sum_{k=1}^{\infty} \frac{a_k}{1 + \lambda \pi k} \sin k\theta \qquad (0 < \theta < \pi),$$

where a_k is the kth coefficient in the Fourier sine-series development of $F(\theta)$ over $(0, \pi)$. [Notice also that the numbers $-1/\pi$, $-1/2\pi$, . . . are hence characteristic values of λ when $f(\theta) = \sin \theta$.]

Section 4.17.

86. (a) Obtain an approximate solution of the special case of Problem 84(a) in which $F(x) = x$ and $\lambda = 1$, using the method of collocation at the points $x = 0$, 0.5, and 1.

(b) Compare the approximate solution with the exact solution, obtained from Problem 84(b), at the points $x = 0$, 0.25, 0.5, 0.75, and 1.

87. (a) Obtain an approximate solution of the equation

$$\sin \frac{\pi x}{2} = \int_0^1 K(x, \xi) y(\xi)\, d\xi,$$

where $K(x, \xi)$ is defined in Problem 84(a), by assuming $y(x) \approx A_1 + A_2 x + A_3 x^2$, and using the method of collocation at the points $x = 0$, 0.5, and 1.

(b) Compare the result with the exact solution $\frac{1}{4}\pi^2 \sin \frac{1}{2}\pi x$ at points $x = 0$, 0.25, 0.5, 0.75, and 1.

88. Obtain an approximate solution of the special case of Problem 85(a) in which $f(\theta) = F(\theta) = 1$ and $\lambda = 1/\pi$, assuming a three-term approximation of the form

$$y(\theta) \approx A_1 \sin \theta + A_3 \sin 3\theta + A_5 \sin 5\theta,$$

and collocating at $\theta = 0$, $\pi/4$, and $\pi/2$. [The exact solution is not known. Notice that, if $f(\theta)$ and $F(\theta)$ are symmetric with respect to $\theta = \pi/2$, only harmonics of odd order need be considered, and recall that

$$\lim_{\theta \to 0} \frac{\sin k\theta}{\sin \theta} = k.]$$

89. (a) Suppose that the quantities

$$\Phi_i(x_i) \equiv \int_a^b K(x_i, \xi)\phi_j(\xi)\, d\xi \qquad (i = 1, 2, \cdots, n)$$

cannot be conveniently evaluated by direct integration, but are to be determined approximately, as weighted sums of the ordinates at a set of N points $\xi_1, \xi_2, \ldots, \xi_N$. Show that the matrix $\mathbf{\Phi}$, for which

$$\Phi_{ij} = \Phi_j(x_i) \qquad (i = 1, 2, \cdots, n; j = 1, 2, \cdots, n)$$

can be obtained by matrix multiplication, in the form

$$\mathbf{\Phi} = \mathbf{K}\,\mathbf{D}\,\boldsymbol{\phi}_\xi,$$

where $\mathbf{K}$ is the $n \times N$-matrix

$$\mathbf{K} = \begin{bmatrix} K(x_1, \xi_1) & \cdots & K(x_1, \xi_N) \\ \cdots & \cdots & \cdots \\ K(x_n, \xi_1) & \cdots & K(x_n, \xi_N) \end{bmatrix}$$

and $\boldsymbol{\phi}_\xi$ is the $N \times n$-matrix

$$\boldsymbol{\phi}_\xi = \begin{bmatrix} \phi_1(\xi_1) & \cdots & \phi_n(\xi_1) \\ \cdots & \cdots & \cdots \\ \phi_1(\xi_N) & \cdots & \phi_n(\xi_N) \end{bmatrix},$$

and where $\mathbf{D}$ is the diagonal $N \times N$-matrix such that the diagonal element $D_i \equiv D_{ii}$ is the weighting coefficient associated with the point ξ_i in the numerical integrations. [Notice that the matrix $\mathbf{K}\,\mathbf{D}$ is hence obtained by multiplying all elements of the jth column of $\mathbf{K}$ by D_j.]

(b) Deduce that the matrix of coefficients of the A's in equation (272) can be expressed in the form

$$\mathbf{f} = \boldsymbol{\phi}_x - \lambda\,\mathbf{K}\,\mathbf{D}\,\boldsymbol{\phi}_\xi,$$

where, in addition to the matrices defined in part (a), the matrix $\boldsymbol{\phi}_x$ is the square $n \times n$-matrix

$$\boldsymbol{\phi}_x = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \cdots & \cdots & \cdots \\ \phi_1(x_n) & \cdots & \phi_n(x_n) \end{bmatrix}.$$

90. (a) Verify that the application of the procedure of Problem 89 to the approximate solution of the equation

$$y(x) = x + \frac{1}{4}\int_0^1 \frac{\sin^2(x - \xi)}{(x - \xi)^2}\, y(\xi)\, d\xi,$$

with the assumption $y(x) \approx A_1 + A_2 x$, with approximate integration involving the three ordinates $\xi = 0$, 0.5, and 1 according to Simpson's rule, and with collocation at $x = 0$ and $x = 1$, leads to the equations

specified by the matrix relation $\mathbf{f}\,\mathbf{A} = \mathbf{F}$, where $\mathbf{F} = \{0, 1\}$, and where $\mathbf{f}$ is evaluated as follows:

$$\mathbf{f} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} - \frac{1}{24}\begin{bmatrix} 1 & 0.919 & 0.708 \\ 0.708 & 0.919 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.817 & -0.106 \\ 0.817 & 0.882 \end{bmatrix}.$$

(b) Obtain the corresponding approximate solution.

Section 4.18.

91. (a) Obtain an approximate solution of the integral equation considered in Problem 86(a), determining the parameters by use of the weighting functions 1, x, and x^2.

(b) Compare the results so obtained with the data of Problem 86(b).

92. (a) Apply the method of weighting functions to the equation treated in Problem 88, first multiplying both sides of the approximate relation by $\sin\theta$, and then using the weighting functions $\sin\theta$, $\sin 3\theta$, and $\sin 5\theta$.

(b) Compare the results with those of Problem 88.

Section 4.19.

93. (a) Apply the modified method of least squares to the integral equation treated in Problem 86(a), using the points $x = 0$, 0.25, 0.5, 0.75, and 1 as the points x_i, and using weighting coefficients corresponding to Simpson's rule.

(b) Compare the results with those of Problems 86(b) and (91).

94. (a) Proceed as in Problem 93 in dealing with Problem 88, introducing the two additional points $\theta = \pi/8$ and $3\pi/8$.

(b) Compare the results with those of Problems 88 and 92.

95. (a) Apply the modified method of least squares to the treatment of Problem 90, introducing the one additional point $x = 0.5$. [First obtain the three equations corresponding to (298) by the method of Problem 89, taking $n = N = 3$.]

(b) Compare the results with those of Problem 90.

Section 4.20.

96. (a) If $K(x, \xi) = x$ when $x < \xi$ and $K(x, \xi) = \xi$ when $x > \xi$, determine the coefficients in the approximation

$$K(x, \xi) \approx A_1 + A_2 x + A_3 x^2 \qquad (0 \leqq x \leqq 1),$$

as functions of ξ, in such a way that the two members are equal at $x = 0$, and at $x = 1$, and that they possess equal integrals over (0, 1). [Thus obtain the approximation $K(x, \xi) \approx 4x\xi - 3x\xi^2 - 3x^2\xi + 3x^2\xi^2$.]

(b) Use this approximation to obtain an approximate solution of the problem $y(x) = x + \int_0^1 K(x, \xi)y(\xi)\,d\xi$, and compare the result with the exact solution. [See Problem 84(b).]

97. By replacing $K(x, \xi) = \dfrac{\sin^2(x - \xi)}{(x - \xi)^2}$ by the first two terms of its expansion in powers of $(x - \xi)$,

$$K(x, \xi) = 1 - \tfrac{1}{3}(x - \xi)^2 + \tfrac{2}{45}(x - \xi)^4 + \cdots,$$

obtain an approximate solution of the problem

$$y(x) = x + \frac{1}{4}\int_0^1 K(x, \xi)y(\xi)\,d\xi,$$

and compare the result with those of Problems 90 and 95.

98. (a) Determine the constants A_1 and A_2 in such a way that the integral of the squared difference between the two members of the relation

$$\frac{\sin^2 u}{u^2} \approx A_1 + A_2u^2 \qquad (0 \leqq u \leqq 1)$$

is as small as possible. [Evaluate $\int_0^1 \dfrac{\sin^2 u}{u^2}\,du$ numerically, by use of series or otherwise.]

(b) Treat Problem 97 by replacing $K(x, \xi)$ by the corresponding approximation $A_1 + A_2(x - \xi)^2$, and compare the result with the results of Problem 97.

99. Obtain an estimate of the smallest characteristic value of λ for the problem $y(x) = \lambda \int_0^1 K(x, \xi)y(\xi)\,d\xi$, where $K(x, \xi)$ is defined in Problem 97.

100. Suppose that $y(x)$ is the required solution of the integral equation

$$y(x) = F(x) + \lambda \int_a^b K(x, \xi)y(\xi)\,d\xi,$$

and that $\bar{y}(x)$ is the solution of the equation obtained by approximating $K(x, \xi)$ by a separable kernel $\bar{K}(x, \xi)$. If the error $y(x) - \bar{y}(x)$ in the solution is denoted by $\varepsilon(x)$, and the error $K(x, \xi) - \bar{K}(x, \xi)$ in the approximation to the kernel is denoted by $\Delta(x, \xi)$, show that the true solution is of the form

$$y(x) = \bar{y}(x) + \varepsilon(x),$$

where $\varepsilon(x)$ satisfies the equation

$$\varepsilon(x) = \Phi(x) + \lambda \int_a^b K(x, \xi)\varepsilon(\xi)\,d\xi,$$

in which

$$\Phi(x) = \lambda \int_a^b \Delta(x, \xi)\bar{y}(\xi)\, d\xi.$$

[Notice that this equation is of the same form as the original equation, with the prescribed function $F(x)$ replaced by the calculable function $\Phi(x)$. Thus, if $K(x, \xi)$ is again approximated by $\bar{K}(x, \xi)$ in the integral equation for the correction $\varepsilon(x)$, the linear algebraic equations which are then to be solved differ only in the right-hand members from those already determined in obtaining the first approximation. The function $\Phi(x)$, and the relevant integrals involving that function, may be evaluated by numerical methods, extreme accuracy usually being unnecessary if the kernel error $\Delta(x, \xi)$ is small over (a, b).]

APPENDIX

The Crout Method for Solving Sets of Linear Algebraic Equations

The method of Crout, for solving a set of n linear algebraic equations in n unknowns, is basically equivalent to the method of Gauss (see footnote to page 4). However, the calculations are systematized in such a way that they are conveniently carried out on a desk calculator, with a minimum number of separate machine operations. Furthermore, a very considerable saving of time and labor results from the fact that the recording of auxiliary data is minimized and compactly arranged.

Only a description and illustration of the method is given here; an analytic justification is included in the original paper.*

The calculation proceeds from the *augmented matrix* of the system,

$$\mathbf{M} \equiv \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & c_n \end{array}\right] \equiv [\mathbf{a}|\mathbf{c}], \tag{1}$$

which may be considered as partitioned into the coefficient matrix $\mathbf{a}$ and the column vector $\mathbf{c}$, to an *auxiliary matrix*

$$\mathbf{M}' \equiv \left[\begin{array}{cccc|c} a'_{11} & a'_{12} & \cdots & a'_{1n} & c'_1 \\ a'_{21} & a'_{22} & \cdots & a'_{2n} & c'_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a'_{n1} & a'_{n2} & \cdots & a'_{nn} & c'_n \end{array}\right] \equiv [\mathbf{a}'|\mathbf{c}'], \tag{2}$$

of the same dimensions, and thence to the required *solution vector*

$$\mathbf{x} \equiv \left\{\begin{array}{c} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{array}\right\}. \tag{3}$$

It is convenient to define the *diagonal element* of any element to the *right* of the principal diagonal of a matrix as that element of the principal

* See Reference 7 to Chapter 1.

diagonal which lies in the same *row* as the given element. The diagonal element of any element *below* the principal diagonal is defined as that element of the principal diagonal which lies in the same *column* as the given element.

With this definition, the procedure for obtaining the elements of $\mathbf{M}'$ from those of the given matrix $\mathbf{M}$ may be described by the four rules which follow:

1. The elements of $\mathbf{M}'$ are determined in the following order: elements of the first column, then elements of the first row to the right of the first column; elements of the second column below the first row, then elements of the second row to the right of the second column; and so on, until all elements are determined.

2. The first column of $\mathbf{M}'$ is identical with the first column of $\mathbf{M}$. Each element of the first row of $\mathbf{M}'$ except the first is obtained by dividing the corresponding element of $\mathbf{M}$ by the leading element a_{11}.

3. Each element a'_{ij} on or below the principal diagonal of $\mathbf{M}'$ is obtained by subtracting from the corresponding element a_{ij} of $\mathbf{M}$ the sum of the products of elements in the ith row and corresponding elements in the jth column of $\mathbf{M}'$, all uncalculated elements being imagined to be zeros. In symbols, we thus have

$$a'_{ij} = a_{ij} - \sum_{k=1}^{j-1} a'_{ik}a'_{kj} \qquad (i \geqq j). \tag{4}$$

4. Each element a'_{ij} to the right of the principal diagonal is calculated by the procedure of Rule 3 *followed by a division by the diagonal element* a'_{ii} in $\mathbf{M}'$. Thus there follows

$$a'_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} a'_{ik}a'_{kj}}{a'_{ii}} \qquad (i < j). \tag{5}$$

In the important cases when the coefficient matrix $\mathbf{a}$ is *symmetric* ($a_{ji} = a_{ij}$), it can be shown that *any element* a'_{ij} *to the right of the principal diagonal is equal to the result of dividing the symmetrically placed element* a'_{ji} (below the diagonal) *by its diagonal element* a'_{ii}. This fact reduces the labor involved in the formation of $\mathbf{M}'$, in such a case, by a factor of nearly two, since each element *below* the diagonal may be recorded as a by-product of the calculation of the symmetrically placed element, before the final division is effected.

The procedure for obtaining the final solution vector $\mathbf{x}$ from the matrix $\mathbf{a}'$ and the vector $\mathbf{c}'$, into which $\mathbf{M}'$ is partitioned, may be described by the three rules which follow:

1. The elements of $\mathbf{x}$ are determined in the reverse order $x_n, x_{n-1}, x_{n-2}, \ldots, x_1$, from the last element to the first.

2. The last element x_n is identical with the last element c'_n of $\mathbf{c}'$.

3. Each element x_i of $\mathbf{x}$ is obtained by subtracting from the corresponding element c_i' of $\mathbf{c}'$ the sum of the products of elements in the ith row of $\mathbf{a}'$ by corresponding elements of the column $\mathbf{x}$, all uncalculated elements of $\mathbf{x}$ being imagined to be zeros. Thus there follows

$$x_i = c_i' - \sum_{k=i+1}^{n} a'_{ik}x_k. \tag{6}$$

The solution may, of course, be checked completely by substitution into the n basic linear equations. However, if desired, a *check column* may be carried along in the calculation to provide a *continuous* check on the work. Each element of the initial check column, corresponding to the augmented matrix $\mathbf{M}$, is the sum of the elements of the corresponding row of $\mathbf{M}$. If this column is recorded to the right of the augmented matrix, *and is treated in the same manner as the column* $\mathbf{c}$, corresponding check columns are thus obtained for the auxiliary matrix $\mathbf{M}'$, and for the solution vector $\mathbf{x}$. Continuous checks on the calculation are then afforded by the two rules which follow:

1. In the auxiliary matrix, any element of the check column should *exceed by unity* the sum of the other elements in its row which lie *to the right of the principal diagonal.*

2. Each element of the check column associated with the solution vector should *exceed by unity* the corresponding element of the solution vector.

The preceding checks serve, not only to display numerical errors, but also to give an estimate of the effect of round-off errors resulting from the retention of insufficiently many significant figures.*

The simplicity and efficiency of this procedure can be appreciated only when it is applied to specific problems and compared with other procedures.

It is of some interest to notice that (as is shown in the reference cited) the set of equations which would be obtained by the direct use of the Gauss reduction would possess the augmented matrix

$$\tilde{\mathbf{M}} = \left[\begin{array}{cccc|c} 1 & a'_{12} & \cdots & a'_{1n} & c_1' \\ 0 & 1 & \cdots & a'_{2n} & c_2' \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & c_n' \end{array}\right] \equiv [\tilde{\mathbf{a}}|\mathbf{c}'], \tag{7}$$

which differs from (2) only in the substitution of ones in the principal diagonal and zeros below it. The transition from the set of equations represented by (7) to the solution (6) is seen to correspond to the "back solution" of the Gauss procedure. *The compactness of the Crout procedure*

* Appreciable loss of accuracy may be encountered if a *small diagonal element* appears in $\mathbf{M}'$ at an early stage of the calculation. Such a situation may often be remedied by reordering the equations and/or unknowns.

is achieved by recording auxiliary data in the spaces of $\tilde{\mathbf{M}}$ *which would otherwise be occupied by ones and zeros.*

The ith diagonal element of $\mathbf{a}'$ happens to be the coefficient by which the ith equation is divided, before that equation is used to eliminate the ith unknown from succeeding equations in the Gauss reduction. Since all other steps in the reduction do not affect the *determinant* of the coefficient matrix, it follows that *the determinant of* $\mathbf{a}$ *is equal to the product of the diagonal elements of* $\mathbf{a}'$,

$$|\mathbf{a}| = a'_{11}a'_{22} \cdots a'_{nn}. \tag{8}$$

Thus the Crout procedure is useful also in evaluating determinants, the columns $\mathbf{c}$ and $\mathbf{c}'$, as well as the final vector $\mathbf{x}$, then being omitted.

In the reference cited, it is shown that the method can be extended to the convenient treatment of equations with *complex* coefficients, and to the case of m equations in n unknowns.

In order to illustrate the procedure numerically, we apply it to the system

$$\begin{aligned} 554.11\ x_1 - 281.91\ x_2 - 34.240x_3 &= 273.02, \\ -281.91\ x_1 + 226.81\ x_2 + 38.100x_3 &= -63.965, \\ -34.240x_1 + 38.100x_2 + 80.221x_3 &= 34.717, \end{aligned}$$

with the *augmented matrix* (and associated check column)

$$\mathbf{M} \equiv \begin{bmatrix} \underline{554.11} & -281.91 & -34.240 & 273.02 \\ -281.91 & \underline{226.81} & 38.100 & -63.965 \\ -34.240 & 38.100 & \underline{80.221} & 34.717 \end{bmatrix} \quad \begin{matrix} \textit{Check} \\ 510.98 \\ -80.965 \\ 118.80 \end{matrix}$$

The *auxiliary matrix* (and associated check column) are obtained in the form

$$\mathbf{M}' \equiv \begin{bmatrix} \underline{554.11} & -0.50876 & -0.061793 & 0.49272 \\ -281.91 & \underline{83.385} & 0.24801 & 0.89870 \\ -34.240 & 20.680 & \underline{72.976} & 0.45224 \end{bmatrix} \quad \begin{matrix} \textit{Check} \\ 0.92216 \\ 2.14668 \\ 1.45228 \end{matrix}$$

and the *solution vector* (and final check column) are found to be

$$\mathbf{x} \equiv \begin{Bmatrix} 0.92083 \\ 0.78654 \\ 0.45224 \end{Bmatrix}, \quad \begin{matrix} \textit{Check} \\ 1.92080 \\ 1.78650 \\ 1.45228 \end{matrix}$$

if all calculated values are rounded off to five significant figures throughout the calculation. Thus there follows

$$x_1 = 0.92083, \quad x_2 = 0.78654, \quad x_3 = 0.45224,$$

where reference to the final check column indicates the possible presence of round-off errors of the order of four units in the last place retained. Such errors would be decreased by retaining additional significant figures in the formation of $\mathbf{M}'$ and $\mathbf{x}$.

The elements of the first column of $\mathbf{M}'$ are identical with the corresponding elements of $\mathbf{M}$; the elements of the first row of $\mathbf{M}'$ following the first element are obtained by dividing the corresponding elements of $\mathbf{M}$ by 554.11. The remaining elements of $\mathbf{M}'$ are determined as follows:

$$a'_{22} = 226.81 - (-281.91)(-0.50876) = 83.385.$$

$$a'_{32} = 38.100 - (-0.50876)(-34.240) = 20.680.$$

$$a'_{23} = \frac{38.100 - (-0.061793)(-281.91)}{83.385} = \frac{20.680}{83.385} = 0.24801.$$

$$c'_2 = \frac{-63.965 - (0.49272)(-281.91)}{83.385} = 0.89870.$$

$$a'_{33} = 80.221 - (-0.061793)(-34.240) - (0.24801)(20.680) = 72.976.$$

$$c'_3 = \frac{34.717 - (0.49272)(-34.240) - (0.89870)(20.680)}{72.976} = 0.45224.$$

The last element of $\mathbf{x}$ is identical with c'_3. The remaining elements of $\mathbf{x}$ are determined as follows:

$$x_2 = 0.89870 - (0.24801)(0.45224) = 0.78654.$$

$$x_1 = 0.49272 - (-0.50876)(0.78654) - (-0.061793)(0.45224) = 0.92083.$$

It may be noticed that, because of the symmetry of the coefficient matrix, it is not actually necessary to calculate a'_{23} and a'_{32} independently. If a'_{23} is calculated, the numerator (20.680) may be recorded as a'_{32} before the final division is effected.

The advantages of the procedure (and the additional simplifications introduced by symmetry of the matrix a) increase with the number of equations involved.

It is particularly important to notice that the calculation of each element of either $\mathbf{M}'$ or $\mathbf{x}$ involves only a *single continuous machine operation* (a sum of products, with or without a final division), without the necessity of intermediate tabulation or transfer of auxiliary data.

If the *determinant* of the coefficient matrix were required, it would be obtained as the product of the diagonal elements of $\mathbf{a}'$:

$$|\mathbf{a}| = (554.11)(83.385)(72.976) = 3.3718 \times 10^7.$$

Answers to Problems

Chapter 1

1. (a) $x_1 = 1, x_2 = -1, x_3 = 1.$
(b) $x_1 = 2 - \frac{1}{2}c, x_2 = -\frac{5}{2} + \frac{3}{4}c, x_3 = c.$

2. (a) $\begin{bmatrix} 3 & -2 & 3 \\ 0 & 1 & 0 \end{bmatrix}$. (b) $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. (c) $[a_1b_1 + a_2b_2 + \cdots + a_nb_n]$.

(d) $\begin{bmatrix} a_1b_1 & a_2b_1 & \cdots & a_nb_1 \\ a_1b_2 & a_2b_2 & \cdots & a_nb_2 \\ \cdots & \cdots & \cdots & \cdots \\ a_1b_n & a_2b_n & \cdots & a_nb_n \end{bmatrix}$. (e) $\begin{bmatrix} c_1a_{11} & c_1a_{12} \\ c_2a_{21} & c_2a_{22} \end{bmatrix}$.

(f) $\begin{bmatrix} c_1a_{11} & c_2a_{12} \\ c_1a_{21} & c_2a_{22} \end{bmatrix}$.

5. 0.319, 0.363, 0.462, 0.587, 0.724.

8. $\lambda = 1$: $\mathbf{x} = c\{1, -1, 2\}$; $\lambda = -9$: $\mathbf{x} = c\{3, 9, -2\}$.

10. $\begin{vmatrix} x_1^2 & x_1y_1 & y_1^2 & 1 \\ x_2^2 & x_2y_2 & y_2^2 & 1 \\ x_3^2 & x_3y_3 & y_3^2 & 1 \\ x_4^2 & x_4y_4 & y_4^2 & 1 \end{vmatrix} = 0.$

16. $\mathbf{a}^T = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, Adj $\mathbf{a} = \begin{bmatrix} 1 & -2 & -1 \\ -2 & 2 & 2 \\ 1 & -2 & -3 \end{bmatrix}$,

$\mathbf{a}^{-1} = \begin{bmatrix} -\frac{1}{2} & 1 & \frac{1}{2} \\ 1 & -1 & -1 \\ -\frac{1}{2} & 1 & \frac{3}{2} \end{bmatrix}$.

17. $|\mathbf{a}| \neq 0$.

21. (b) $\mathbf{x} = \{\frac{6}{5}, \frac{2}{5}, 0\} + c\{\frac{1}{3}, -\frac{3}{5}, 1\} = \frac{1}{5}\{6 + c, 2 - 3c, 5c\}$.

22. (b) $\lambda = 1$: $\mathbf{x} = c_1\{2, 1, 0\} + c_2\{-1, 0, 1\}$;
$\lambda = -3$: $\mathbf{x} = c\{1, 2, -1\}$.

25. (a) $r = 3$. (b) $r = 3$. (c) $r = 2$.

27. (b) The set $\mathbf{a}\,\mathbf{x} = \mathbf{c}$ possesses a solution for any $\mathbf{c}$.

29. (a) $\lambda_1 = 1$: $\mathbf{e}_1 = \pm\{1/\sqrt{5}, -2/\sqrt{5}\}$; $\lambda_2 = 6$: $\mathbf{e}_2 = \pm\{2/\sqrt{5}, 1/\sqrt{5}\}$.
(c) $\mathbf{x} = \{2/(6 - \lambda), 1/(6 - \lambda)\}$ if $\lambda \neq 1$ or 6. If $\lambda = 6$, there is no solution. If $\lambda = 1$, the solution is $\mathbf{x} = \{\frac{2}{5} + c, \frac{1}{5} - 2c\}$, where c is arbitrary.

31. If $\mathbf{e}_1$ is taken as a multiple of the first vector, and $\mathbf{e}_2$ a combination of the first two, then

$$\mathbf{e}_1 = \pm\frac{1}{3}\{1, 0, 2, 2\},\ \mathbf{e}_2 = \pm\frac{\sqrt{2}}{6}\{2, 3, -2, 1\},\ \mathbf{e}_3 = \pm\frac{\sqrt{2}}{6}\{2, 1, 2, -3\}.$$

33.
$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \end{bmatrix}, \qquad \mathbf{Q}^T\mathbf{a}\,\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

(Columns may be interchanged. Also, the signs of all elements in any column of Q may be changed.)

34. With Q as defined in answer to Problem 33, $F = x_1'^2 + 2x_2'^2 + 4x_3'^2$.

36.
$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{Q} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}.$$

37. $\lambda_1 = 1$: $\mathbf{e}_1 = \pm\frac{\sqrt{2}}{6}\{1 + i, -4\}$; $\lambda_2 = 10$: $\mathbf{e}_2 = \pm\frac{1}{3}\{2 + 2i, 1\}$.

41. F is not positive definite.

42. With $x_1 = \frac{1}{2}(\alpha_1 + \alpha_2)$, $x_2 = \frac{1}{2}(\alpha_1 - \alpha_2)$, there follows $A = \alpha_1^2 + 2\alpha_2^2$, $B = \alpha_1^2 + \alpha_2^2$. (Other sign combinations are possible in the definitions of x_1 and x_2.)

43. 7; 10.

44. **a** is not positive definite.

45. (a) All characteristic numbers must be negative.
(b) Discriminants with odd subscripts must be negative; those with even subscripts must be positive.

46. $\mathbf{x}' = \{\sqrt{2}, 0, 1\}$.

47.
$$\mathbf{y}' = \begin{bmatrix} \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} & 0 \\ \frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{Bmatrix} 3 \\ 2 \\ 1 \end{Bmatrix} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} & \sqrt{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{Bmatrix} \sqrt{2} \\ 0 \\ 1 \end{Bmatrix} = \begin{Bmatrix} \frac{5}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} \\ 1 \end{Bmatrix}.$$

52. (a) Characteristic numbers: -1 and 5; corresponding characteristic vectors: multiples of $\{1, -1\}$ and $\{1, 1\}$, respectively.
(b) **b** is not positive definite.

53.
$$\begin{bmatrix} \frac{3^{100} + 1}{2} & \frac{3^{100} - 1}{2} \\ \frac{3^{100} - 1}{2} & \frac{3^{100} + 1}{2} \end{bmatrix}.$$

57. $\lambda_3 = 8.12$: multiple of $\{0.229, 0.631, 1.000\}$.

58. The vector $\{0, 0\}$ is inevitably obtained after two iterations. The only characteristic number is $\lambda = 0$.

59. Successive approximations oscillate. The characteristic numbers are complex.

60. $\lambda_1 = 8.290$: multiple of $\{0.347, 0.653, 0.879, 1.000\}$;
$\lambda_2 = 1$: multiple of $\{1, 1, 0, -1\}$.

61. $\lambda_1 = 0$: multiple of $\{1, 1, 1, 1\}$; $\lambda_2 = 0.586$: multiple of $\{-1, -0.414, 0.414, 1\}$; $\lambda_3 = 2$: multiple of $\{1, -1, -1, 1\}$; $\lambda_4 = 3.414$: multiple of $\{-1, 2.414, -2.414, 1\}$.

65. $\lambda_1 = 1$: multiple of $\{1, -2\}$; $\lambda_2 = \frac{11}{2}$: multiple of $\{4, 1\}$.

66. $M = \begin{bmatrix} \frac{1}{3} & \frac{2}{3}\sqrt{2} \\ -\frac{2}{3} & \frac{1}{6}\sqrt{2} \end{bmatrix}$.
(Columns may be interchanged. Also, the signs of all elements in any column may be changed.)

67. With $\mathbf{M}$ as defined in answer to Problem 66, $\mathbf{x} = \mathbf{M}\alpha$ leads to desired forms with $\lambda_1 = 1$ and $\lambda_2 = \frac{11}{2}$.

69. $\lambda_1 = 0$: $\mathbf{e}_1 = c_1\{1, -2\}$, $\mathbf{e}_1' = c_1'\{1, 1\}$;
$\lambda_2 = 1$: $\mathbf{e}_2 = c_2\{1, -1\}$, $\mathbf{e}_2' = c_2'\{2, 1\}$.

72. $\omega_1 = 0.518\sqrt{k/M}$: $\mathbf{x} = c_1\{0.268, 0.732, 1\}$;
$\omega_2 = 1.41\sqrt{k/M}$: $\mathbf{x} = c_2\{1, 1, -1\}$;
$\omega_3 = 1.93\sqrt{k/M}$: $\mathbf{x} = c_3\{3.73, -2.73, 1\}$.

73. $\omega_1 = 0.404\sqrt{k/M}$: $\mathbf{x} = c_1\{0.238, 0.674, 1\}$;
$\omega_2 = 1.30\sqrt{k/M}$: $\mathbf{x} = c_2\{1.79, 2.36, -1\}$;
$\omega_3 = 1.91\sqrt{k/M}$: $\mathbf{x} = c_3\{9.53, -6.33, 1\}$.

74. $\omega_1 = 0$: $\mathbf{x} = c_1\{1, 1, 1\}$;
$\omega_2 = 1.00\sqrt{k/M}$: $\mathbf{x} = c_2\{1, 0, -1\}$;
$\omega_3 = 1.73\sqrt{k/M}$: $\mathbf{x} = c_3\{1, -2, 1\}$.

75. $\omega_1 = 0$: $\mathbf{x} = c_1\{1, 1, 1\}$;
$\omega_2 = 0.85\sqrt{k/M}$: $\mathbf{x} = c_2\{1.56, 0.44, -1\}$;
$\omega_3 = 1.67\sqrt{k/M}$: $\mathbf{x} = c_3\{2.56, -4.56, 1\}$.

96. $A_k = \dfrac{2}{\mu_k(1 + \cos \mu_k)}$. (This expression can also be written in various other forms.)

97. (b) $y(x) = \displaystyle\sum_{k=1}^{\infty} \frac{2 \sin \mu_k x}{\mu_k(\lambda - \mu_k^2)(1 + \cos \mu_k)}$ $\quad (0 < x < 1)$.

Chapter 2

2. Lengths $a/\sqrt{3}$ in the x-direction, $b/\sqrt{3}$ in the y-direction, and $c/\sqrt{3}$ in the z-direction.

3. Squares of semiaxes are $2\left[(A + C) \pm \sqrt{(A - C)^2 + 4B^2}\right]^{-1}$.

4. $y = \frac{1}{2}(5x^2 - 3x)$.

7. (b) $I(\epsilon) = 2 + \frac{1}{3}\epsilon^2$.

8. (a) $y'' - y = 0$. (b) $2(x\,y')' = 1$. (c) $2y'' + k^2 \sin y = 0$.
(d) $(a\,y')' + b\,y = 0$.

9. $Az + B\theta = C$.

10. $Ar = B \sec(\theta \sin\alpha + C)$.

11. $A \cot\phi = B\cos\theta + C\sin\theta$.

12. $A\theta = B\displaystyle\int \frac{\sqrt{1+f'^2}}{r\sqrt{r^2 - B^2}}\,dr + C$.

13. $I(x) = \sqrt{2}$; $I(\cosh x) = \sinh 1$.

14. (a) 2ϵ. (b) 3ϵ.

15. $\Delta I = \frac{3}{2}\epsilon + \frac{17}{15}\epsilon^2$; $\delta I = \frac{3}{2}\epsilon$.

18. Euler equation: $(a\,y'')'' + (b\,y')' + c\,y = 0$.
Natural boundary conditions: $\Big[\{(a\,y'')' + b\,y'\}\delta y\Big]_{x_1}^{x_2} = \Big[a\,y''\,\delta y'\Big]_{x_1}^{x_2} = 0$.

21. Euler equation: $(a\,u_x)_x + (b\,u_y)_y + c\,u = 0$.
Natural boundary condition:

$$\oint_C \left(a\frac{\partial u}{\partial x}\cos\nu + b\frac{\partial u}{\partial y}\sin\nu\right)\delta u\,ds = 0.$$

23. Euler equation: $u_{xxxx} + 2u_{xxyy} + u_{yyyy} = \nabla^4 u = 0$.
Natural boundary conditions:

$$\left[\left\{\frac{\partial}{\partial x}\nabla^2 u + (1-\alpha)u_{xyy}\right\}\delta u\right]_{x_1}^{x_2} = \left[(u_{xx} + \alpha\,u_{yy})\,\delta u_x\right]_{x_1}^{x_2} = 0,$$

$$\left[\left\{\frac{\partial}{\partial y}\nabla^2 u + (1-\alpha)u_{xxy}\right\}\delta u\right]_{y_1}^{y_2} = \left[(u_{yy} + \alpha\,u_{xx})\,\delta u_y\right]_{y_1}^{y_2} = 0.$$

26. $(x - \frac{1}{2})^2 + (y-k)^2 = k^2 + \frac{1}{4}$, where k satisfies the equation $(4k^2+1)\cot^{-1}2k = 4A + 2k$.

36. $\omega_n \approx \dfrac{n^2\pi^2}{L^2}\sqrt{\dfrac{\displaystyle\int_0^L E\,I\sin^2\frac{n\pi x}{L}\,dx}{\displaystyle\int_0^L \rho\sin^2\frac{n\pi x}{L}\,dx}}$.

42. $m_1\ddot{x}_1 = k_2(x_2 - x_1) - k_1x_1$, $m_2\ddot{x}_2 = k_3(x_3 - x_2) - k_2(x_2 - x_1)$, $m_3\ddot{x}_3 = -k_3(x_3 - x_2)$.

43. Equation of motion is $\ddot{q}_1 = \frac{1}{23}g$.

52. $H = \dfrac{1}{2m}\left(p_r^2 + \dfrac{1}{r^2}p_\theta^2\right) + V(r)$. $\dot{r} = \dfrac{p_r}{m}$, $\dot{\theta} = \dfrac{p_\theta}{mr^2}$;
$\dot{p}_r = p_\theta^2/mr^3 - V'(r)$, $\dot{p}_\theta = 0$.

56. (b) $m\ddot{x} + \sum_i k_i l_i (l_i x + m_i y + n_i z) = 0,$

$$m\ddot{y} + \sum_i k_i m_i (l_i x + m_i y + n_i z) = 0,$$

$$m\ddot{z} + \sum_i k_i n_i (l_i x + m_i y + n_i z) = 0.$$

61. $\omega^4 - g(A + C)\omega^2 + g^2(AC - B^2) = 0.$

64. (a) $V = \frac{Q^2}{2C} - EQ, \quad T = \frac{1}{2}L\dot{Q}^2, \quad F = \frac{1}{2}R\dot{Q}^2;$

$$L\ddot{Q} + R\dot{Q} + \frac{Q}{C} - E = 0; \quad \omega = \frac{1}{\sqrt{LC}}.$$

(b) $V = \frac{Q_1^2}{2C_1} + \frac{(Q_1 - Q_2)^2}{2C_{12}} - E_1 Q_1, \quad T = \frac{1}{2}(L_1\dot{Q}_1^2 + L_2\dot{Q}_2^2),$

$F = \frac{1}{2}R_{12}(\dot{Q}_1 - \dot{Q}_2)^2;$

$$L_1\ddot{Q}_1 + \frac{1}{C_1}Q_1 + \frac{1}{C_{12}}(Q_1 - Q_2) + R_{12}(\dot{Q}_1 - \dot{Q}_2) - E_1 = 0,$$

$$L_2\ddot{Q}_2 + \frac{1}{C_{12}}(Q_2 - Q_1) + R_{12}(\dot{Q}_2 - \dot{Q}_1) = 0;$$

$$\begin{vmatrix} \left(L_1\omega^2 - \frac{1}{C_1} - \frac{1}{C_{12}}\right) & \frac{1}{C_{12}} \\ \frac{1}{C_{12}} & \left(L_2\omega^2 - \frac{1}{C_{12}}\right) \end{vmatrix} = 0.$$

69. $\delta \int_a^b \left[\frac{1}{2}(x^2 y_1'^2 + x^4 y_2'^2 - x^2 y_1^2 + 2x^3 y_1 y_2 - x^4 y_2^2) + x\,\phi_1 y_1 + x^3 \phi_2 y_2\right] dx = 0.$

76. $c_1 = \frac{85}{26} \doteq 3.27, \quad c_2 = -\frac{35}{18} \doteq -2.69.$

77. $\lambda_1^{(1)} = 15; \quad \lambda_1^{(2)} \doteq 14.42, \quad \lambda_2^{(1)} \doteq 63.6.$

Chapter 3

1. (a) $y_2 = 2\cos\alpha, \; y_3 = 4\cos^2\alpha - 1, \; y_4 = 8\cos^3\alpha - 4\cos\alpha.$

2. $\lambda = 0, \pm\sqrt{2}.$

3. (b) $f_k = \frac{1}{2}k(k+1); \quad f_{100} = 5050.$

4. $A_k y_{k+2} + (B_k - 2A_k)y_{k+1} + (C_k - B_k + A_k)y_k = \phi_k.$

11. $T_2 = x^2 - \frac{1}{2}, \; T_3 = x^3 - \frac{3}{4}x, \; T_4 = x^4 - x^2 + \frac{1}{8}.$

13. (a) $y_k = c_1 2^k + c_2\left(\frac{3+\sqrt{17}}{4}\right)^k + c_3\left(\frac{3-\sqrt{17}}{4}\right)^k.$

(b) $y_k = c_1 + 2^k(c_2 + c_3 k)$.

(c) $y_k = c_1 \cos \frac{\pi k}{4} + c_2 \sin \frac{\pi k}{4} + c_3 \cos \frac{3\pi k}{4} + c_4 \sin \frac{3\pi k}{4}$.

(d) $y_k = (c_1 + c_2 k) \cos \frac{\pi k}{2} + (c_3 + c_4 k) \sin \frac{\pi k}{2}$.

18. (a) $y_k = c_1 + c_2 k + \frac{a^{k+1}}{(a-1)^2}$ $(a \neq 1)$.

(b) $y_k = c_1 + c_2 k + \frac{e^{b(k+1)}}{(e^b - 1)^2}$ $(b \neq 0)$.

(c) $y_k = c_1 + c_2 k - \frac{\sin ck}{2(1 - \cos c)}$ $(c \neq 2n\pi)$.

(d) $y_k = c_1 + c_2 k + \frac{1}{2}k^2$.

(e) $y_k = c_1 + c_2 k + \frac{1}{6}k^3$.

(f) $y_k = c_1 + c_2 k + \frac{e^{b(k+1)}}{(e^b - 1)^3}[(e^b - 1)k - (e^b + 1)]$ $(b \neq 0)$.

20. $y_k = c_1 + c_2 k + \begin{cases} 0 & \text{if } k \leq r, \\ k - r & \text{if } k \geq r, \end{cases}$ when $f_k = \delta_{kr}$.

24. $p_m = m/(m+n)$.

26. (a) $\omega_n = 2\sqrt{\frac{K}{M}} \sin \frac{n\pi}{2(N+1)}$ $(n = 1, 2, \ldots, N)$;

$x_{n,k} = C_n \sin \frac{n\pi k}{N+1} \cos(\omega_n t + \beta_n)$.

45. (a) $S_n = \frac{1}{4}n(n+1)(n+2)(n+3)$.

(b) $S_n = \frac{1}{4} - \frac{1}{2(n+1)(n+2)}$.

(c) $S_n = \frac{1}{2} - \frac{1}{2(2n+3)}$.

(d) $S_n = \frac{1}{6}n(n+1)(2n+7)$.

56. $y_k = \sum_{n=1}^{N} \sin\left(\frac{2n-1}{2N+1}\pi k\right)[A_n \cos \omega_n t + B_n \sin \omega_n t]$, where

$$A_n = \frac{4}{2N+1} \sum_{k=1}^{N} d_k \sin\left(\frac{2n-1}{2N+1}\pi k\right),$$

$$B_n = \frac{4}{\omega_n(2N+1)} \sum_{k=1}^{N} v_k \sin\left(\frac{2n-1}{2N+1}\pi k\right),$$

$$\omega_n = 2\sqrt{\frac{T}{Mh}} \sin\left(\frac{2n-1}{2N+1}\frac{\pi}{2}\right).$$

59. (a) $T_k(t) = \sin\left(\frac{r\pi k}{N+1}\right) e^{-\gamma^2\mu_r^2 t}$.

(b) $T_k(t) = \frac{2\phi_r}{N+1} \sum_{n=1}^{N} \sin\frac{n\pi r}{N+1} \sin\frac{n\pi k}{N+1} e^{-\gamma^2\mu_n^2 t}$.

(c) $T_k(t) = \frac{2}{N+1} \sum_{\substack{n=1 \\ (n \text{ odd})}}^{N} \cot\frac{n\pi}{2(N+1)} \sin\frac{n\pi k}{N+1} e^{-\gamma^2\mu_n^2 t}$.

66. (c)

x	0	0.2	0.4	0.6	0.8	1.0
y	0	0.200	0.398	0.590	0.768	0.921

(d)

x	0	0.2	0.4	0.6	0.8	1.0
y	0	0.217	0.432	0.641	0.834	1.000

67. (c)

x	0	0.2	0.4	0.6	0.8	1.0
y	0	0.200	0.398	0.590	0.767	0.919

(d)

x	0	0.2	0.4	0.6	0.8	1.0
y	0	0.217	0.433	0.641	0.834	1.000

68. $\lambda_1^{(1)} = 16$; $\lambda_1^{(2)} \doteq 17.1$, $\lambda_2^{(1)} \doteq 63.9$.

69. 1.208, 1.211, 1.209.

70. After 10 hours:

x	0	0.2	0.4	0.6	0.8	1.0
T	100	128	154	174	189	200

71. After 1 hour:

x	0	0.2	0.4	0.6	0.8	1.0
T	180	198	200	200	200	200

72. After 10 hours:

x	0	0.2	0.4	0.6	0.8	1.0
T	122	144	163	178	190	200

73. Temperatures at points A, B, C of Figure 3.18 (page 297):

Time (hr)	0	1	2	3	4	5
Point A	100	100.0	96.9	93.0	91.8	91.3
Point B	100	87.5	71.9	67.2	65.2	64.6
Point C	100	87.5	71.9	68.0	66.8	66.3

74. Temperatures at points A, B, C of Figure 3.18 (page 297):

Time (hr)	0	1	2	3	4	5
Point A	100	100.0	96.9	91.4	86.3	82.2
Point B	100	87.5	71.9	65.6	61.0	57.8
Point C	100	87.5	65.6	52.4	44.9	39.8

75.

0	50	100	150	200
25	69	112	156	200
50	88	125	162	200
75	106	138	169	200
100	125	150	175	200

76.

0	50	100	150	200
25	66	104	137	156
50	83	114	138	149
75	103	130	152	163
100	125	150	175	200

79. (b)

	0	0	0	
0	0.072	0.099	0.072	0
0	0.188	0.250	0.188	0
0	0.429	0.527	0.429	0
	I	I	I	

80.

100	100	100		
100	110	124	148	
100	118	138	164	200
100	122	145	171	200
100	125	150	175	200

81.

100	100	100		
120	122	130	149	
135	138	148	168	200
145	148	158	176	200
148	151	161	178	200

82. (b) Values of $u = 1600\phi/\beta a^2$:

0	0	0	0	0
0	68.7	87.4	68.7	0
0	87.4	112.4	87.4	0
0	68.7	87.4	68.7	0
0	0	0	0	0

(c) Repeated use of Simpson's rule gives $\iint_A \phi \, dA \approx 0.033\beta a^2$.

83. Values of $w/w_{\max.}$:

0	0	0	0	0
0	0.04	0.07	0.04	0
0	0.17	0.27	0.17	0
0	0.43	0.63	0.43	0
0	0.75	1.00	0.75	0

84. (c)

	200	170	135	
	200	172	136	
	200	184	172	
	200	190	185	
A'	200	192	188	B'

Chapter 4

10. (b) $y(x) = \lambda \int_0^1 G(x, \xi)\xi\, y(\xi)\, d\xi$.

36. (b) $F(x) = x$: $y = \dfrac{2\pi^2\lambda^2}{\pi^2\lambda^2 - 1} \sin x + \dfrac{2\pi\lambda}{\pi^2\lambda^2 - 1} \cos x + x$.

$F(x) = 1$: $y = 1$.

(c) $F(x) = 1$: $y = 1 + c(\cos x + \sin x)$.

(d) $F(x) = A \cos x + B \sin x$.

37. $y \approx 0.363x + x^2 - 0.039x^3$.

42. (c) $1 - 3x\xi = \dfrac{\sqrt{3}\,(1 - x)\,\sqrt{3}\,(1 - \xi)}{2} + \dfrac{(1 - 3x)(1 - 3\xi)}{-2}$.

44. $y = 1 + \dfrac{12\lambda x + 6\lambda + \lambda^2}{12 - 12\lambda - \lambda^2}$ $\quad (\lambda \neq -6 \pm 4\sqrt{3})$.

Estimated convergence limit: $|\lambda| < \sqrt{42}/7 \doteq 0.926$.

True convergence limit: $|\lambda| < 4\sqrt{3} - 6 \doteq 0.928$.

46. (a) $y^{(3)}(x) = 1 + \frac{3}{2}x^2 + \frac{7}{8}x^4 + \frac{7\,7}{2\,4\,0}x^6$.

47. $\Gamma(x, \xi; \lambda) = x\xi(1 + \frac{1}{3}\lambda + \frac{1}{9}\lambda^2 + \cdots)$.

48. $\Gamma(x, \xi; \lambda) = (x + \xi) + \lambda[\frac{1}{3} + \frac{1}{2}(x + \xi) + x\xi]$
$+ \lambda^2[\frac{1}{5} + \frac{7}{12}(x + \xi) + x\xi] + \cdots$.

49. $K_2(x, \xi) = \begin{cases} (1 + \xi - x)e^{x-\xi} - \frac{1}{2}[e^{-(x+\xi)} + e^{x+\xi-2a}] & (x < \xi), \\ (1 + x - \xi)e^{\xi-x} - \frac{1}{2}[e^{-(x+\xi)} + e^{x+\xi-2a}] & (x > \xi). \end{cases}$

52. $\Gamma(x, \xi; \lambda) = \dfrac{3x\xi}{3 - \lambda}$.

53. $\Gamma(x, \xi; \lambda) = \dfrac{12(x + \xi) + \lambda[4 - 6(x + \xi) + 12x\xi]}{12 - 12\lambda - \lambda^2}$.

81. Exact solution is $y(x) = 1$.

83. $\lambda_1 \approx 1.24$; $y(\frac{1}{2})/y(0) \approx 0.801$, $y(1)/y(0) \approx 0.656$.

88. $y(\theta) \approx 0.541 \sin\theta + 0.031 \sin 3\theta + 0.007 \sin 5\theta$.

90. (b) $y(x) \approx 0.131 + 1.012x$.

99. $\lambda_1 \approx 1.06$.

Index

(*Italicized* figures in parentheses refer to problem numbers)