Eastern Economy Edition

FIFTH EDITION

Introductory Methods of Numerical Analysis

S.S. Sastry

PHI

Introductory Methods of Numerical Analysis

Introductory Methods of Numerical Analysis

Fifth Edition

S.S. SASTRY

Formerly, Scientist/Engineer SF
Vikram Sarabhai Space Centre
Trivandrum

PHI Learning Private Limited

New Delhi-110001

2012

INTRODUCTORY METHODS OF NUMERICAL ANALYSIS, Fifth Edition
S.S. Sastry

ISBN-978-81-203-4592-8

Forty-fifth Printing (Fifth Edition) ... ... June, 2012

Published by Asoke K. Ghosh, PHI Learning Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Rajkamal Electric Press, Plot No. 2, Phase IV, HSIDC, Kundli-131028, Sonepat, Haryana.

To

My Grandsons

Venkata Bala Nagendra

Venkata Badrinath

Contents

Preface

This fifth edition of *Introductory Methods of Numerical Analysis* contains eleven chapters on numerical methods which could be used by scientists and engineers to solve problems arising in research and industry. It also covers the syllabus prescribed for engineering and science students at undergraduate and graduate levels in Indian Universities. The present edition includes the following features:

1. Most of the illustrative examples and problems in exercises have been modified and a number of new problems are included in every chapter. This edition contains 511 problems including the illustrative examples and exercises for homework.
2. Many minor changes and refinements are made in the presentation of material in some chapters of the text. Because of their increasing importance in applications, the spline functions are discussed in a separate chapter. Their applications are considered in the appropriate chapters of the text.
3. Algorithms, computational steps or flow charts are provided for some of the numerical methods and these can easily be transformed into a computer program by including suitable input/output statements. Also, problems have been set to design algorithms or write flow-charts for their computation.
4. Answers have been provided for all the problems in exercises.
5. *Instructors Manual* is also available for teachers which provides relevant information concerning each chapter and also solutions to 317 problems in the exercises.
6. Four model question papers on numerical methods are provided at the end of the book.

All the essential features of the previous editions like the over-all easy-to-understand presentation and organization of the material and the choice of suitable illustrative examples are retained in this edition.

Although our primary objective has been to provide the student with an introduction to the methods of numerical analysis, we have also strived to make this book as student-friendly as possible. It is hoped that this edition will serve this purpose and meet the requirements of students and teachers in numerical analysis.

The author is indebted to Sri Asoke K. Ghosh, Chairman and Managing Director, PHI Learning, for his courteous cooperation in bringing out this new edition.

Any information concerning corrections or errors in this book will be gratefully received.

S.S. SASTRY

Chapter 1

Errors in Numerical Calculations

1.1 INTRODUCTION

In practical applications, an engineer would finally obtain results in a numerical form. For example, from a set of tabulated data derived from an experiment, inferences may have to be drawn; or, a system of linear algebraic equations is to be solved. The aim of numerical analysis is to provide efficient methods for obtaining numerical answers to such problems. This book deals with methods of numerical analysis rather than the analysis of numerical methods, because our main concern is to provide computer-oriented, efficient and reliable numerical methods for solving problems arising in different areas of higher mathematics. The areas of numerical mathematics, addressed in this book, are:

(a) *Algebraic and transcendental equations:* The problem of solving nonlinear equations of the type $f(x) = 0$ is frequently encountered in engineering. For example, the equation

$$\frac{M_0}{M_0 - u_f t} = e^{(u+gt)/u_0} \tag{1.1}$$

is a nonlinear equation for t when M_0, g, u, u_0 and u_f are given. Equations of this type occur in rocket studies.

(b) *Interpolation:* Given a set of data values (x_i, y_i), $i = 0, 1, 2, \ldots, n$, of a function $y = f(x)$, where the explicit nature of $f(x)$ is not known, it is often required to find the value of y for a given value of x, where $x_0 < x < x_n$. This process is called *interpolation*. If this process

is carried out for functions of several variables, it is called *multivariate* interpolation.

(c) *Curve fitting:* This is a special case where the data points are subject to errors, both round off and systematic. In such a case, interpolation formulae yield unsatisfactory solutions. Experimental results are often subject to errors and, in such cases, the method is to fit a curve which passes through the data points and then use the curve to predict the intermediate values. This problem is usually referred to as *data smoothing*.

(d) *Numerical differentiation and integration:* It is often required to determine the numerical values of

(i) $\dfrac{dy}{dx}, \dfrac{d^2y}{dx^2}, \ldots,$ for a certain value of x in $x_0 \le x \le x_n$, and

(ii) $I = \int_{x_0}^{x_n} y\,dx,$

where the set of data values (x_i, y_i), $i = 0, 1, \ldots, n$ is given, but the explicit nature of $y(x)$ is not known. For example, if the data consist of the angle θ (in radians) of a rotating rod for values of time t (in seconds), then its angular velocity and angular acceleration at any time can be computed by numerical differentiation formulae.

(e) *Matrices and linear systems:* The problem of solving systems of linear algebraic equations and the determination of eigenvalues and eigenvectors of matrices are major problems of disciplines such as differential equations, fluid mechanics, theory of structures, etc.

(f) *Ordinary and partial differential equations:* Engineering problems are often formulated in terms of an ordinary or a partial differential equation. For example, the mathematical formulation of a falling body involves an ordinary differential equation and the problem of determining the steady-state distribution of temperature on a heated plate is formulated in terms of a partial differential equation. In most cases, exact solutions are not possible and a numerical method has to be adopted. In addition to the finite difference methods, this book also presents a brief introduction to the cubic spline method for solving certain partial differential equations.

(g) *Integral equations:* An equation in which the unknown function appears under the integral sign is known as an *integral equation*. Equations of this type occur in several areas of higher mathematics such as aerodynamics, elasticity, electrostatics, etc. A short account of some well-known methods is given.

In the numerical solution of problems, we usually start with some initial data and then compute, after some intermediate steps, the final results. The given numerical data are only approximate because they may be true to two, three or more figures. In addition, the

methods used may also be approximate and therefore the error in a computed result may be due to the errors in the data, or the errors in the method, or both. In Section 1.3, we discuss some basic ideas concerning errors and their analyses, since such an understanding is essential for an effective use of numerical methods. Before discussing about errors in computations, we shall first look into some important computer languages and software.

1.1.1 Computer and Numerical Software

It is well known that computers and mathematics are two important tools of numerical methods. Prior to 1950, numerical methods could only be implemented by manual computations, but the rapid technological advances resulted in the production of computing machines which are faster, economical and smaller in size. Today's engineers have access to several types of computing systems, viz., mainframe computers, personal computers and super computers. Of these, the personal computer is a smaller machine which is useful, less expensive and, as the name implies, can easily be possessed and used by individuals. Nevertheless, mere possession of a computer is not of great consequence; it can be used effectively only by providing suitable instructions to it. These instructions are known as *software*. It is therefore imperative that we develop suitable software for an effective implementation of numerical methods on computers.

Essentially, there are three phases in the development of numerical software for solving a problem. In the first phase, the problem to be solved must be formulated mathematically indicating the input and outputs and also the checks to be made on the solution. The second phase consists of choosing an *algorithm*, i.e., a suitable numerical procedure to solve the mathematical problem. An algorithm is a set of instructions leading to the solution of the mathematical problem, and also contains information regarding the accuracy required and computation of error in the solution. In the final phase, the algorithm must be transformed into a *computer program* (called *code*) which is a set of step-by-step instructions to the computer written in a computer language. Usually, it may be preferable to prepare a *flowchart* first and then transform the flowchart into a computer program. The flowchart consists of the step-by-step procedures, in block form, which the computer will follow and which can easily be understood by others who wish to know about the program. It is easy to see that the flowchart enables a programmer to develop a quality computer program using one of the computer languages listed in the next section. However, experienced programmers often transform a detailed algorithm into an efficient computer program.

1.1.2 Computer Languages

Several computer languages have so far been developed and there are limitations on every language. The question of preferring a particular language over

others depends on the problem and its requirements. We list below some important problem-solving languages, which are currently in use:

(a) *FORTRAN:* Standing for FORmula TRANslation, FORTRAN was introduced by IBM in 1957. Since then, it has undergone many changes and the present version, called FORTRAN 90, is favoured by most scientists and engineers. It is readily available on almost all computers and one of its important features is that it allows a programmer to express the mathematical algorithm more precisely. It has special features like extended double precision, special mathematical functions and complex variables. Besides, FORTRAN is the language used in numerically oriented subprograms developed by many software libraries. For example, (IMSL) (International Mathematical and Statistical Library, Inc.) consists of FORTRAN subroutines and functions in applied mathematics, statistics and special functions. FORTRAN programs are also available in the book, *Numerical Recipes*, published by the Cambridge University Press, for most of the standard numerical methods.

(b) *C:* This is a high-level programming language developed by Bell Telephone Laboratories in 1972. Presently, it is being taught at several engineering colleges as the first computer language and is therefore used by a large number of engineers and scientists. Computer programs in C for standard numerical methods are available in the book, *Numerical Recipes in C*, published by the Cambridge University Press.

(c) *BASIC:* Originally developed by John Kemeny and Thomas Kurtz in 1960, BASIC was used in the first few years only for instruction purposes. Over the years, it has grown tremendously and the present version is called Visual Basic. One of its important applications is in the development of software on personal computers. It is easy to use.

1.1.3 Software Packages

It is well known that the programming effort is considerably reduced by using standard *functions* and *subroutines*. Several software packages for numerical methods are available in the form of 'functions' and these are being extensively used by engineering students. One such package is MATLAB, standing for MATrices LABoratory. It was developed by Cleve Moler and John N. Little. As the name implies, it was originally founded to develop a matrix package but now it incorporates several numerical methods such as root-finding of polynomials, cubic spline interpolation, discrete Fourier transforms, numerical differentiation and integration, ordinary differential equations and eigenvalue problems. Besides, MATLAB has excellent display capabilities which can be used in the case of two-dimensional problems. Using the

MATLAB functions, it is possible to implement most of the numerical methods on personal computers and hence it has become one of the most popular packages in most laboratories and technical colleges. MATLAB has its own programming language and this is described in detail in the text by Stephen J. Chapman.*

1.2 MATHEMATICAL PRELIMINARIES

In this section we state, without proof, certain mathematical results which would be useful in the sequel.

Theorem 1.1 If $f(x)$ is continuous in $a \le x \le b$, and if $f(a)$ and $f(b)$ are of opposite signs, then $f(\xi) = 0$ for at least one number ξ such that $a < \xi < b$.

Theorem 1.2 (*Rolle's theorem*) If $f(x)$ is continuous in $a \le x \le b$, $f'(x)$ exists in $a < x < b$ and $f(a) = f(b) = 0$, then, there exists at least one value of x, say ξ, such that $f'(\xi) = 0$, $a < \xi < b$.

Theorem 1.3 (*Generalized Rolle's theorem*) Let $f(x)$ be a function which is n times differentiable on $[a, b]$. If $f(x)$ vanishes at the $(n+1)$ distinct points $x_0, x_1, \ldots, x_n$ in (a, b), then there exists a number ξ in (a, b) such that $f^{(n)}(\xi) = 0$.

Theorem 1.4 (*Intermediate value theorem*) Let $f(x)$ be continuous in $[a, b]$ and let k be any number between $f(a)$ and $f(b)$. Then there exists a number ξ in (a, b) such that $f(\xi) = k$ (see Fig. 1.1).

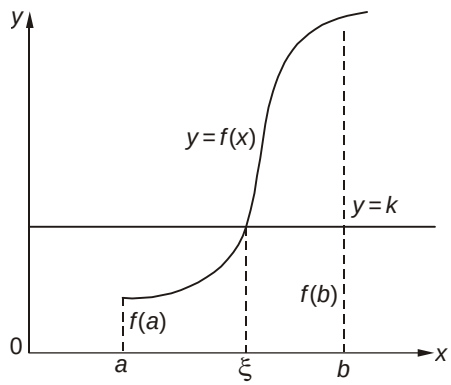


Figure 1.1 Intermediate value theorem.

Theorem 1.5 (*Mean-value theorem for derivatives*) If $f(x)$ is continuous in $[a, b]$ and $f'(x)$ exists in (a, b), then there exists at least one value of x, say ξ, between a and b such that

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}, \qquad a < \xi < b.$$

*Published by Thomson Asia Pte. Ltd., Singapore (2002).

Setting $b = a + h$, this theorem takes the form

$$f(a+h) = f(a) + hf'(a+\theta h), \qquad 0 < \theta < 1.$$

Theorem 1.6 (*Taylor's series for a function of one variable*) If $f(x)$ is continuous and possesses continuous derivatives of order n in an interval that includes $x = a$, then in that interval

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^{n-1}}{(n-1)!}f^{(n-1)}(a) + R_n(x),$$

where $R_n(x)$, the *remainder term*, can be expressed in the form

$$R_n(x) = \frac{(x-a)^n}{n!}f^{(n)}(\xi), \qquad a < \xi < x.$$

Theorem 1.7 (*Maclaurin's expansion*) It states

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + \cdots$$

Theorem 1.8 (*Taylor's series for a function of two variables*) It states

$$f(x_1 + \Delta x_1, x_2 + \Delta x_2) = f(x_1, x_2) + \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2$$

$$+\frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\Delta x_1)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}\Delta x_1 \Delta x_2 + \frac{\partial^2 f}{\partial x_2^2}(\Delta x_2)^2\right] + \cdots$$

This can easily be generalized.

Theorem 1.9 (*Taylor's series for a function of several variables*)

$$f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n)$$

$$= f(x_1, x_2, \ldots, x_n) + \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n$$

$$+\frac{1}{2}\left[\frac{\partial^2 f}{\partial x_1^2}(\Delta x_1)^2 + \cdots + \frac{\partial^2 f}{\partial x_n^2}(\Delta x_n)^2 + 2\frac{\partial^2 f}{\partial x_1 \partial x_2}\Delta x_1 \Delta x_2 + \cdots\right.$$

$$\left. + 2\frac{\partial^2 f}{\partial x_{n-1} \partial x_n}\Delta x_{n-1} \Delta x_n\right] + \cdots$$

1.3 ERRORS AND THEIR COMPUTATIONS

There are two kinds of numbers, *exact* and *approximate* numbers. Examples of exact numbers are 1, 2, 3, …, 1/2, 3/2, …, $\sqrt{2}, \pi$, e, etc., written in this manner. Approximate numbers are those that represent the numbers to a certain degree of accuracy. Thus, an approximate value of π is 3.1416, or if we desire a better approximation, it is 3.14159265. But we cannot write the *exact* value of π.

The digits that are used to express a number are called *significant digits* or *significant figures*. Thus, the numbers 3.1416, 0.66667 and 4.0687 contain five significant digits each. The number 0.00023 has, however, only two significant digits, viz., 2 and 3, since the zeros serve only to fix the position of the decimal point. Similarly, the numbers 0.00145, 0.000145 and 0.0000145 all have three significant digits. In case of ambiguity, the scientific notation should be used. For example, in the number 25,600, the number of significant figures is uncertain, whereas the numbers 2.56×10^4, 2.560×10^4 and 2.5600×10^4 have three, four and five significant digits, respectively.

In numerical computations, we come across numbers which have large number of digits and it will be necessary to cut them to a usuable number of figures. This process is called *rounding off*. It is usual to round-off numbers according to the following rule:

To round-off a number to n significant digits, discard all digits to the right of the nth digit, and if this discarded number is

(a) less than half a unit in the nth place, leave the nth digit unaltered;

(b) greater than half a unit in the nth place, increase the nth digit by unity;

(c) exactly half a unit in the nth place, increase the nth digit by unity if it is odd; otherwise, leave it unchanged.

The number thus rounded-off is said to be correct to n significant figures.

Example 1.1 The numbers given below are rounded-off to four significant figures:

1.6583	to	1.658
30.0567	to	30.06
0.859378	to	0.8594
3.14159	to	3.142

In hand computations, the round-off error can be reduced by carrying out the computations to more significant figures at each step of the computation. A useful rule is: at each step of the computation, retain at least one more significant figure than that given in the data, perform the last operation and then round-off. However, most computers allow more number

of significant figures than are usually required in engineering computations. Thus, there are computers which allow a precision of seven significant figures in the range of about 10^{-38} to 10^{39}. Arithmetic carried out with this precision is called *single precision* arithmetic, and several computers implement *double precision* arithmetic, which could be used in problems requiring greater accuracy. Usually, the double precision arithmetic is carried out to 15 decimals with a range of about 10^{-308} to 10^{308}. In MATLAB, there is a *provision* to use double precision arithmetic.

In addition to the round-off error discussed above, there is another type of error which can be caused by using approximate formulae in computations, —such as the one that arises when a *truncated* infinite series is used. This type of error is called *truncation error* and its study is naturally associated with the problem of convergence. Truncation error in a problem can be evaluated and we are often required to make it as small as possible. Sections 1.4 and 1.5 will be devoted to a discussion of these errors.

Absolute, relative and percentage errors

Absolute error is the numerical difference between the true value of a quantity and its approximate value. Thus, if X is the true value of a quantity and X_1 is its approximate value, then the absolute error E_A is given by

$$E_A = X - X_1 = \delta X. \tag{1.2}$$

The relative error E_R is defined by

$$E_R = \frac{E_A}{X} = \frac{\delta X}{X}, \tag{1.3}$$

and the percentage error (E_P) by

$$E_P = 100\, E_R. \tag{1.4}$$

Let ΔX be a number such that

$$\left| X_1 - X \right| \le \Delta X. \tag{1.5}$$

Then ΔX is an upper limit on the magnitude of the absolute error and is said to measure *absolute accuracy.* Similarly, the quantity

$$\frac{\Delta X}{|X|} \approx \frac{\Delta X}{|X_1|}$$

measures the *relative accuracy*.

It is easy to deduce that if two numbers are added or subtracted, then the magnitude of the absolute error in the result is the sum of the magnitudes of the absolute errors in the two numbers. More generally, if $E_A^1, E_A^2, \ldots, E_A^n$ are the absolute errors in n numbers, then the magnitude of the absolute error in their sum is given by

$$|E_A^1| + |E_A^2| + \cdots + |E_A^n|.$$

Note: While adding up several numbers of different absolute accuracies, the following procedure may be adopted:

(i) Isolate the number with the greatest absolute error,

(ii) Round-off all other numbers retaining in them one digit more than in the isolated number,

(iii) Add up, and

(iv) Round-off the sum by discarding one digit.

To find the absolute error, E_A, in a product of two numbers a and b, we write $E_A = (a + E_A^1)(b + E_A^2) - ab$, where E_A^1 and E_A^2 are the absolute errors in a and b respectively. Thus,

$$\begin{aligned} E_A &= aE_A^2 + bE_A^1 + E_A^1 E_A^2 \\ &= bE_A^1 + aE_A^2, \text{ approximately} \end{aligned} \tag{1.6}$$

Similarly, the absolute error in the quotient a/b is given by

$$\begin{aligned} \frac{a + E_A^1}{b + E_A^2} - \frac{a}{b} &= \frac{bE_A^1 - aE_A^2}{b(b + E_A^2)} \\ &= \frac{bE_A^1 - aE_A^2}{b^2(1 + E_A^2/b)} \\ &= \frac{bE_A^1 - aE_A^2}{b^2}, \text{ assuming that } E_A^2/b \text{ is small in comparison with 1} \\ &= \frac{a}{b}\left(\frac{E_A^1}{a} - \frac{E_A^2}{b}\right). \end{aligned} \tag{1.7}$$

Example 1.2 If the number X is rounded to N decimal places, then

$$\Delta X = \frac{1}{2}(10^{-N}).$$

If X = 0.51 and is correct to 2 decimal places, then ΔX = 0.005, and the percentage accuracy is given by $\frac{0.005}{0.51} \times 100 = 0.98\%$.

Example 1.3 An approximate value of π is given by $X_1 = 22/7 = 3.1428571$ and its true value is $X = 3.1415926$. Find the absolute and relative errors. We have

$$E_A = X - X_1 = -0.0012645$$

and

$$E_R = \frac{-0.0012645}{3.1415926} = -0.000402.$$

Example 1.4 Three approximate values of the number 1/3 are given as 0.30, 0.33 and 0.34. Which of these three values is the best approximation?

We have

$$\left|\frac{1}{3} - 0.30\right| = \frac{1}{30}.$$

$$\left|\frac{1}{3} - 0.33\right| = \frac{0.01}{3} = \frac{1}{300}.$$

$$\left|\frac{1}{3} - 0.34\right| = \frac{0.02}{3} = \frac{1}{150}.$$

It follows that 0.33 is the best approximation for 1/3.

Example 1.5 Find the relative error of the number 8.6 if both of its digits are correct.

Here

$$E_A = 0.05$$

Hence

$$E_R = \frac{0.05}{8.6} = 0.0058.$$

Example 1.6 Evaluate the sum $S = \sqrt{3} + \sqrt{5} + \sqrt{7}$ to 4 significant digits and find its absolute and relative errors.

We have

$$\sqrt{3} = 1.732,\ \sqrt{5} = 2.236 \text{ and } \sqrt{7} = 2.646$$

Hence $S = 6.614$. Then

$$E_A = 0.0005 + 0.0005 + 0.0005 = 0.0015$$

The total absolute error shows that the sum is correct to 3 significant figures only. Hence we take $S = 6.61$ and then

$$E_R = \frac{0.0015}{6.61} = 0.0002.$$

Example 1.7 Sum the following numbers:

0.1532, 15.45, 0.000354, 305.1, 8.12, 143.3, 0.0212, 0.643 and 0.1734,

where in each of which all the given digits are correct.

Here we have two numbers which have the greatest absolute error. These are 305.1 and 143.3, and the absolute error in both these numbers is 0.05. Hence, we round-off all the other numbers to two decimal digits. These are:

0.15, 15.45, 0.00, 8.12, 0.02, 0.64 and 0.17.

The sum S is given by

$$\begin{aligned} S &= 305.1 + 143.3 + 0.15 + 15.45 + 0.00 + 8.12 + 0.02 + 0.64 + 0.17 \\ &= 472.95 \\ &= 473 \end{aligned}$$

To determine the absolute error, we note that the first-two numbers have each an absolute error of 0.05 and the remaining seven numbers have an absolute error of 0.005 each. Thus, the absolute error in all the 9 numbers is

$$\begin{aligned} E_A &= 2(0.05) + 7(0.005) \\ &= 0.1 + 0.035 \\ &= 0.135 \\ &= 0.14 \end{aligned}$$

In addition to the above absolute error, we have to take into account the rounding error in the above and this is 0.01. Hence the total absolute error is $S = 0.14 + 0.01 = 0.15$. Thus,

$$S = 472.95 \pm 0.15.$$

Example 1.8 Find the difference

$$\sqrt{6.37} - \sqrt{6.36}$$

to three significant figures.

We have

$$\sqrt{6.37} = 2.523885893$$

and

$$\sqrt{6.36} = 2.521904043$$

Therefore, $\sqrt{6.37} - \sqrt{6.36} = 0.001981850$

$= 0.00198$, correct to three significant figures.

Alternatively, we have

$$\begin{aligned} \sqrt{6.37} - \sqrt{6.36} &= \frac{6.37 - 6.36}{\sqrt{6.37} + \sqrt{6.36}} \\ &= \frac{0.01}{2.524 + 2.522} \\ &= 0.198 \times 10^{-2}, \text{ which is the same result as obtained before.} \end{aligned}$$

Example 1.9 Two numbers are given as 2.5 and 48.289, both of which being correct to the significant figures given. Find their product.

Here the number 2.5 is the one with the greatest absolute error. Hence, we round-off the second number to three significant digits, i.e., 48.3. The required product is given by

$$P = 48.3 \times 2.5$$
$$= 1.2 \times 10^2$$

In the product, we retained only two significant digits, since one of the given numbers, viz. 2.5, contained only two significant digits.

1.4 A GENERAL ERROR FORMULA

We now derive a general formula for the error committed in using a certain formula or a functional relation. Let

$$u = f(x, y, z) \tag{1.8}$$

and let the errors in x, y, z be Δx, Δy and Δz, respectively. Then the error Δu in u is given by

$$u + \Delta u = f(x + \Delta x, y + \Delta y, z + \Delta z) \tag{1.9}$$

Expanding the right-side of Eq. (1.9) by Taylor's series, we obtain

$$u + \Delta u = f(x, y, z) + \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z$$
$$+ \text{ terms involving higher powers of } \Delta x, \Delta y \text{ and } \Delta z \tag{1.10}$$

Assuming that the errors Δx, Δy, Δz are small, their higher powers can be neglected and Eq. (1.10) becomes

$$\Delta u = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z \tag{1.11}$$

The relative error in u is then given by

$$E_R = \frac{\Delta u}{u} = \frac{\partial u}{\partial x}\frac{\Delta x}{u} + \frac{\partial u}{\partial y}\frac{\Delta y}{u} + \frac{\partial u}{\partial z}\frac{\Delta z}{u} \tag{1.12}$$

Example 1.10 Find the value of

$$s = \frac{a^2\sqrt{b}}{c^3},$$

where $a = 6.54 \pm 0.01$, $b = 48.64 \pm 0.02$, and $c = 13.5 \pm 0.03$.

Also, find the relative error in the result.

We have

$$a^2 = 42.7716, \sqrt{b} = 6.9742 \text{ and } c^3 = 2460.375$$

Therefore,

$$s = \frac{42.7716 \times 6.9742}{2460.375} = 0.12124\ldots$$
$$= 0.121$$

Also,

$$\log s = 2\log a + \frac{1}{2}\log b - 3\log c$$

$$\Rightarrow \left|\frac{\Delta s}{s}\right| \le 2\left|\frac{\Delta a}{a}\right| + \frac{1}{2}\left|\frac{\Delta b}{b}\right| + 3\left|\frac{\Delta c}{c}\right| = 2\left(\frac{0.01}{6.54}\right) + \frac{1}{2}\left(\frac{0.02}{48.64}\right) + 3\left(\frac{0.03}{13.5}\right)$$

$$= 0.009931$$

Example 1.11 Given that

$$u = \frac{5xy^2}{z^3},$$

find the relative error at $x = y = z = 1$ when the errors in each of x, y, z is 0.001.

We have

$$\frac{\partial u}{\partial x} = \frac{5y^2}{z^3},\ \frac{\partial u}{\partial y} = \frac{10xy}{z^3} \text{ and } \frac{\partial u}{\partial z} = -\frac{15xy^2}{z^4}$$

Then

$$\Delta u = \frac{5y^2}{z^3}\Delta x + \frac{10xy}{z^3}\Delta y - \frac{15xy^2}{z^4}\Delta z.$$

In general, the errors Δx, Δy and Δz may be positive or negative. Hence, we take the absolute values of the terms on the right side. We then obtain

$$(\Delta u)_{\max} = \left|\frac{5y^2}{z^3}\Delta x\right| + \left|\frac{10xy}{z^3}\Delta y\right| + \left|\frac{15xy^2}{z^4}\Delta z\right|$$

but $\Delta x = \Delta y = \Delta z = 0.001$ and $x = y = z = 1$. Then, the relative maximum error $(E_R)_{\max}$ is given by

$$(E_R)_{\max} = \frac{(\Delta u)_{\max}}{u}$$
$$= \frac{0.03}{5} = 0.006.$$

1.5 ERROR IN A SERIES APPROXIMATION

The truncated error committed in a series approximation can be evaluated by using Taylor's series stated in Theorem 1.6. If x_i and x_{i+1} are two successive values of x, then we have

$$f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i) f'(x_i) + \cdots + \frac{(x_{i+1} - x_i)^n}{n!} f^{(n)}(x_i) + R_{n+1}(x_{i+1}), \quad (1.13)$$

where

$$R_{n+1}(x_{i+1}) = \frac{(x_{i+1} - x_i)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \qquad x_i < \xi < x_{i+1} \quad (1.14)$$

In Eq. (1.13), the last term, $R_{n+1}(x_{i+1})$, is called the *remainder term* which, for a convergent series, tends to zero as $n \to \infty$. Thus, if $f(x_{i+1})$ is approximated by the first-n terms of the series given in Eq. (1.13), then the maximum error committed by using this approximation (called the *nth order approximation*) is given by the remainder term $R_{n+1}(x_{i+1})$. Conversely, if the accuracy required is specified in advance, then it would be possible to find n, the number of terms, such that the finite series yields the required accuracy.

Defining the interval length,

$$x_{i+1} - x_i = h, \quad (1.15)$$

Equation (1.13) may be written as

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2!} f''(x_i) + \cdots + \frac{h^n}{n!} f^{(n)}(x_i) + \mathrm{O}(h^{n+1}), \quad (1.16)$$

where $\mathrm{O}(h^{n+1})$ means that the truncation error is of the order of h^{n+1}, i.e., it is proportional to h^{n+1}. The meaning of this statement will be made clearer now.

Let the series be truncated after the first term. This gives the *zero-order* approximation:

$$f(x_{i+1}) = f(x_i) + \mathrm{O}(h), \quad (1.17)$$

which means that halving the interval length h will also halve the error in the approximate solution. Similarly, the *first-order* Taylor series approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \mathrm{O}(h^2), \quad (1.18)$$

which means that halving *the interval length*, h will quarter the error in the approximation. In such a case we say that approximation has a

second-order of convergence. We illustrate these facts through numerical examples.

Example 1.12 Evaluate $f(1)$ using Taylor's series for $f(x)$, where

$$f(x) = x^3 - 3x^2 + 5x - 10.$$

It is easily seen that $f(1) = -7$ but it will be instructive to see how the Taylor series approximations of orders 0 to 3 improve the accuracy of $f(1)$ gradually.

Let $h = 1$, $x_i = 0$ and $x_{i+1} = 1$. We then require $f(x_{i+1})$. The derivatives of $f(x)$ are given by

$$f'(x) = 3x^2 - 6x + 5, \quad f''(x) = 6x - 6, \quad f'''(x) = 6,$$

$f^{\text{iv}}(x)$ and higher derivatives being all zero. Hence

$$f'(x_i) = f'(0) = 5, \qquad f''(x_i) = f''(0) = -6, \qquad f'''(0) = 6.$$

Also,

$$f(x_i) = f(0) = -10.$$

Hence, Taylor's series gives

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f'''(x_i). \tag{i}$$

From Eq. (i), the zero-order approximation is given by

$$f(x_{i+1}) = f(x_i) + \mathrm{O}(h), \tag{ii}$$

and, therefore,

$$f(1) = f(0) + \mathrm{O}(h) \approx -10,$$

the error in which is $-7 + 10$, i.e., 3 units.

For the first approximation, we have

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \mathrm{O}(h^2), \tag{iii}$$

and, therefore,

$$f(1) = -10 + 5 + \mathrm{O}(h^2) \approx -5,$$

the error in which is $-7 + 5$, i.e., -2 units.

Again, the second-order Taylor approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \mathrm{O}(h^3), \tag{iv}$$

and, therefore,

$$f(1) = -10 + 5 + \frac{1}{2}(-6) + \mathrm{O}(h^3) \approx -8,$$

in which the error is $-7 + 8$, i.e., 1 unit.

Finally, the third-order Taylor series approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f'''(x_i), \tag{v}$$

and, therefore,

$$\begin{aligned} f(1) &= f(0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(x_0) \\ &\approx -10 + 5 + \frac{1}{2}(-6) + \frac{1}{6}(6) \\ &= -7, \end{aligned}$$

which is the exact value of $f(1)$.

This example demonstrates that if the given function is a polynomial of third degree, then its third-order Taylor series approximation gives exact results.

Example 1.13 Given $f(x) = \sin x$, construct the Taylor series approximations of orders 0 to 7 at $x = \pi/3$ and state their absolute errors.

Let $x_{i+1} = \pi/3$ and $x_i = \pi/6$ so that $h = \pi/3 - \pi/6 = \pi/6$. We then have

$$\begin{aligned} f\left(\frac{\pi}{3}\right) = f\left(\frac{\pi}{6}\right) + hf'\left(\frac{\pi}{6}\right) + \frac{h^2}{2} f''\left(\frac{\pi}{6}\right) + \frac{h^3}{6} f'''\left(\frac{\pi}{6}\right) + \frac{h^4}{24} f^{\text{iv}}\left(\frac{\pi}{6}\right) \\ + \frac{h^5}{120} f^{\text{v}}\left(\frac{\pi}{6}\right) + \frac{h^6}{720} f^{\text{vi}}\left(\frac{\pi}{6}\right) + \frac{h^7}{5040} f^{\text{vii}}\left(\frac{\pi}{6}\right) + O(h^8) \end{aligned} \tag{i}$$

Since $f(x) = \sin x$, Eq. (i) becomes:

$$\begin{aligned} \sin\left(\frac{\pi}{3}\right) &\approx \sin\left(\frac{\pi}{6}\right) + \frac{\pi}{6}\cos\left(\frac{\pi}{6}\right) + \frac{1}{2}\left(\frac{\pi}{6}\right)^2\left(-\sin\frac{\pi}{6}\right) + \frac{1}{6}\left(\frac{\pi}{6}\right)^3\left(-\cos\frac{\pi}{6}\right) \\ &\quad + \frac{1}{24}\left(\frac{\pi}{6}\right)^4\left(\sin\frac{\pi}{6}\right) + \frac{1}{120}\left(\frac{\pi}{6}\right)^5\left(\cos\frac{\pi}{6}\right) + \frac{1}{720}\left(\frac{\pi}{6}\right)^6\left(-\sin\frac{\pi}{6}\right) \\ &\quad + \frac{1}{5040}\left(\frac{\pi}{6}\right)^7\left(-\cos\frac{\pi}{6}\right). \\ &= 0.5 + \frac{\pi}{12}\sqrt{3} - \frac{1}{4}\frac{\pi^2}{36} - \frac{\sqrt{3}}{12}\left(\frac{\pi}{6}\right)^3 + \frac{1}{48}\left(\frac{\pi}{6}\right)^4 + \frac{\sqrt{3}}{240}\left(\frac{\pi}{6}\right)^5 - \frac{1}{1440}\left(\frac{\pi}{6}\right)^6 \\ &\quad - \frac{\sqrt{3}}{10080}\left(\frac{\pi}{6}\right)^7. \end{aligned}$$

The different orders of approximation can now be evaluated successively. Thus, the zero-order approximation is 0.5; the first-order approximation is $0.5+\pi\sqrt{3}/12$, i.e., 0.953449841; and the second-order approximation is

$$0.5+\frac{\pi\sqrt{3}}{12}-\frac{\pi^2}{144},$$

which simplifies to 0.884910921. Similarly, the successive approximations are evaluated and the respective absolute errors can be calculated since the exact value of sin $(\pi/3)$ is 0.866025403. Table 1.1 gives the approximate values of sin $(\pi/3)$ for the orders 0 to 7 as also the absolute errors in these approximations. The results show that the error decreases with an increase in the order of approximation.

Table 1.1 Taylor's Series Approximations of $f(x) = \sin x$

Order of approximation	Computed value of $\sin \pi/3$	Absolute error
0	0.5	0.366025403
1	0.953449841	0.087424438
2	0.884910921	0.018885518
3	0.864191613	0.00183379
4	0.865757474	0.000267929
5	0.86604149	0.000016087
6	0.86602718	0.000001777
7	0.866025326	0.000000077

We next demonstrate the effect of halving the interval length on any approximate value. For this, we consider the first-order approximation in the form:

$$f(x+h) = f(x)+hf'(x)+E(h), \tag{ii}$$

where $E(h)$ is the absolute error of the first-order approximation with interval h. Taking $f(x)=\sin x$ and $x=\pi/6$, we obtain

$$\sin\left(\frac{\pi}{6}+h\right)=\sin\frac{\pi}{6}+h\cos\frac{\pi}{6}+E(h). \tag{iii}$$

Putting $h=\pi/6$ in (iii), we get

$$\sin\frac{\pi}{3}=0.5+\frac{\pi\sqrt{3}}{12}+E(h)=0.953449841+E(h).$$

Since $\sin(\pi/3)=0.866025403$, the above equation gives

$$E(h)=-0.087424438.$$

Now, let the interval be halved so that we now take $h = \pi/12$. Then, (iii) gives:

$$\sin\left(\frac{\pi}{6}+\frac{\pi}{12}\right) = 0.5 + \frac{\pi}{12}\frac{\sqrt{3}}{2} + E\left(\frac{h}{2}\right), \tag{iv}$$

where $E(h/2)$ is the absolute error with interval length $h/2$. Since

$$\sin\left(\frac{\pi}{6}+\frac{\pi}{12}\right) = \sin\frac{\pi}{4} = \frac{1}{\sqrt{2}}$$

Equation (iv) gives

$$E\left(\frac{h}{2}\right) = \frac{1}{\sqrt{2}} - 0.5 - \frac{\pi\sqrt{3}}{24} = -0.019618139,$$

and then

$$\frac{E(h)}{E(h/2)} = 4.45630633.$$

In a similar way, we obtain the values

$$\frac{E(h/2)}{E(h/4)} = 4.263856931$$

and

$$\frac{E(h/4)}{E(h/8)} = 4.141353027.$$

The h^2-order of convergence is quite revealing in the above results.

Example 1.14 The Maclaurin expansion for e^x is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^{n-1}}{(n-1)!} + \frac{x^n}{n!}e^{\xi}, \qquad 0 < \xi < x.$$

We shall find n, the number of terms, such that their sum yields the value of e^x correct to 8 decimal places at $x = 1$.

Clearly, the error term (i.e. the remainder term) is $(x^n/n!)\,e^{\xi}$, so that at $\xi = x$, this gives the *maximum* absolute error, and hence the maximum relative error is $x^n/n!$ For an 8 decimal accuracy at $x = 1$, we must have

$$\frac{1}{n!} < \frac{1}{2}(10^{-8})$$

which gives $n = 12$. Thus, we need to take 12 terms of the exponential series in order that its sum is correct to 8 decimal places.

Example 1.15 Derive the series

$$\log_e \frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right)$$

and use it to compute the value of $\log_e(1.2)$, correct to seven decimal places. If, instead, the series for $\log_e(1 + x)$ is used, how many terms must be taken to obtain the same accuracy for $\log_e(1.2)$?

We have

$$\log_e (1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \cdots \tag{i}$$

and

$$\log_e (1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \frac{x^5}{5} - \cdots \tag{ii}$$

Therefore,

$$\log_e \frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right) \tag{iii}$$

Putting $x = \frac{1}{11}$ in Eq. (iii), we obtain

$$\log_e 1.2 = 2\left[\frac{1}{11} + \frac{1}{3(11)^3} + \frac{1}{5(11)^5} + \cdots\right]$$

$$= 2\left[\frac{1}{11} + \frac{1}{3(11)^3} + \frac{1}{5(11)^5}\right], \text{ since } \frac{1}{7(11)^7} = 7.33 \times 10^{-9}.$$

Hence we obtain

$$\log_e 1.2 = 2[0.09090909 + 0.00025044 + 0.00000124]$$

$$= 0.1823216$$

On the other hand, if we use series (i), we have

$$\left|\frac{x^n}{n}\right| < 2 \times 10^{-7}$$

$$\Rightarrow \qquad \frac{(1.2)^n}{n} < 2 \times 10^{-7}$$

$$\Rightarrow \qquad n > 9.$$

Thus, 9 terms of the series (i) have to be taken in order to obtain a seven decimal accuracy.

EXERCISES

1.1 Explain the term 'round-off error' and round-off the following numbers to two decimal places:

48.21416, 2.3742, 52.275, 2.375, 2.385, 81.255

1.2 Round-off the following numbers to four significant figures:

38.46235, 0.70029, 0.0022218, 19.235101, 2.36425

1.3 Calculate the value of $\sqrt{102}-\sqrt{101}$ correct to four significant figures.

1.4 If $p = 3c^6 - 6c^2$, find the percentage error in p at $c = 1$, if the error in c is 0.05.

1.5 Find the absolute error in the sum of the numbers 105.6, 27.28, 5.63, 0.1467, 0.000523, 208.5, 0.0235, 0.432 and 0.0467, where each number is correct to the digits given.

1.6 If $z = \frac{1}{8}xy^3$, find the percentage error in z when $x = 3.14 \pm 0.0016$ and $y = 4.5 \pm 0.05$.

1.7 Find the absolute error in the product uv if $u = 56.54 \pm 0.005$ and $v = 12.4 \pm 0.05$.

1.8 Prove that the relative error in a product of three nonzero numbers does not exceed the sum of the relative errors of the given numbers.

1.9 Find the relative error in the quotient 4.536/1.32, the numbers being correct to the digits given.

1.10 The exponential series is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Find the number of terms of the above series such that their sum gives the value of e correct to five decimal places.

1.11 Compute the value of ln 3 correct to five decimal places.

1.12 Write down the Taylor's series expansion of $f(x) = \cos x$ at $x = \frac{\pi}{3}$ in terms of $f(x)$, and its derivatives at $x = \frac{\pi}{4}$. Compute the approximations from the zeroth order to the fifth order and also state the absolute error in each case.

1.13 The Maclaurin expansion of $\sin x$ is given by

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

where x is in radians. Use the series to compute the value of $\sin 25°$ to an accuracy of 0.001.

Answers to Exercises

1.1 Required values are:

48.21, 2.37, 52.28, 2.38, 2.38 and 81.26

1.2 38.46, 0.7003, 0.002222, 19.24, 2.364

1.3 0.04963

1.4 10%

1.5 $S = 347.7 \pm 0.15$

1.6 3%

1.7 2.9

1.9 0.004

1.10 $n = 9$

1.11 $\ln 3 = 1.09861$

1.12 Successive approximations are

0.707106781, 0521986658, 0.497754491, 0.499869146, 0.500007551, …

1.13 0.423

Chapter 2

Solution of Algebraic and Transcendental Equations

2.1 INTRODUCTION

In scientific and engineering studies, a frequently occurring problem is to find the roots of equations of the form

$$f(x) = 0 \tag{2.1}$$

If $f(x)$ is a quadratic, cubic or a biquadratic expression, then algebraic formulae are available for expressing the roots in terms of the coefficients. On the other hand, when $f(x)$ is a polynomial of higher degree or an expression involving transcendental functions, algebraic methods are not available, and recourse must be taken to find the roots by approximate methods.

This chapter is concerned with the description of several numerical methods for the solution of equations of the form given in Eq. (2.1), where $f(x)$ is algebraic or transcendental or a combination of both. Now, algebraic functions of the form

$$f_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n, \tag{2.2}$$

are called *polynomials* and we discuss some special methods for determining their roots. A non-algebraic function is called a *transcendental* function, e.g., $f(x) = \ln x^3 - 0.7$ $\phi(x) = e^{-0.5x} - 5x$, $\psi(x) = \sin^2 x - x^2 - 2$, etc. The roots of Eq. (2.1) may be either real or complex. We discuss methods of finding a real root of algebraic or transcendental equations and also methods of determining all real and complex roots of polynomials. Solution of systems of nonlinear equations will be considered at the end of the chapter.

If $f(x)$ is a polynomial of the form Eq. (2.2), the following results, from the theory of equations would be useful in locating its roots.

(i) Every polynomial equation of the nth degree has n and only n roots.

(ii) If n is odd, the polynomial equation has atleast one real root whose sign is opposite to that of the last term.

(iii) If n is even and the constant term is negative, then the equation has atleast one positive root and atleast one negative root.

(iv) If the polynomial equation has (a) real coefficients, then imaginary roots occur in pairs and (b) rational coefficients, then irrational roots occur in pairs.

(v) *Descartes' Rule of Signs*

(a) A polynomial equation $f(x) = 0$ cannot have more number of positive real roots than the number of changes of sign in the coefficients of $f(x)$.

(b) In (a) above, $f(x) = 0$ cannot have more number of negative real roots than the number of changes of sign in the coefficients of $f(-x)$.

2.2 BISECTION METHOD

This method is based on Theorem 1.1 which states that if a function $f(x)$ is continuous between a and b, and $f(a)$ and $f(b)$ are of opposite signs, then there exists at least one root between a and b. For definiteness, let $f(a)$ be negative and $f(b)$ be positive. Then the root lies between a and b and let its approximate value be given by $x_0 = (a + b)/2$. If $f(x_0) = 0$, we conclude that x_0 is a root of the equation $f(x) = 0$. Otherwise, the root lies either between x_0 and b, or between x_0 and a depending on whether $f(x_0)$ is negative or positive. We designate this new interval as $[a_1, b_1]$ whose length is $|b - a|/2$. As before, this is bisected at x_1 and the new interval will be exactly half the length of the previous one. We repeat this process until the latest interval (which contains the root) is as small as desired, say ε. It is clear that the interval width is reduced by a factor of one-half at each step and at the end of the nth step, the new interval will be $[a_n, b_n]$ of length $|b - a|/2^n$. We then have

$$\frac{|b-a|}{2^n} \le \varepsilon,$$

which gives on simplification

$$n \ge \frac{\log_e(|b-a|/\varepsilon)}{\log_e 2} \tag{2.3}$$

Equation (2.3) gives the number of iterations required to achieve an accuracy ε. For example, if $|b - a| = 1$ and $\varepsilon = 0.001$, then it can be seen that

$$n \ge 10 \tag{2.4}$$

The method is shown graphically in Fig. 2.1.

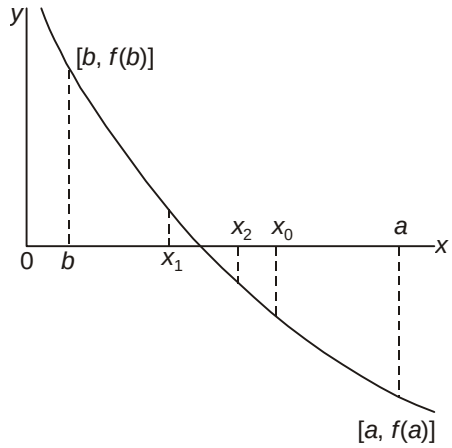


Figure 2.1 Graphical representation of the bisection method.

It should be noted that this method always succeeds. If there are more roots than one in the interval, bisection method finds one of the roots. It can be easily programmed using the following computational steps:

1. Choose two real numbers a and b such that $f(a)\, f(b) < 0$.
2. Set $x_r = (a + b)/2$.
3. (a) If $f(a)\, f(x_r) < 0$, the root lies in the interval (a, x_r). Then, set $b = x_r$ and go to step 2 above.
   (b) If $f(a)\, f(x_r) > 0$, the root lies in the interval (x_r, b). Then, set $a = x_r$ and go to step 2.
   (c) If $f(a)\, f(x_r) = 0$, it means that x_r is a root of the equation $f(x) = 0$ and the computation may be terminated.

In practical problems, the roots may not be exact so that condition (c) above is never satisfied. In such a case, we need to adopt a criterion for deciding when to terminate the computations.

A convenient criterion is to compute the percentage error ε_r defined by

$$\varepsilon_r = \left| \frac{x_r' - x_r}{x_r'} \right| \times 100\%. \tag{2.5}$$

where x_r' is the new value of x_r. The computations can be terminated when ε_r becomes less than a prescribed tolerance, say ε_p. In addition, the maximum number of iterations may also be specified in advance.

Example 2.1 Find a real root of the equation $f(x) = x^3 - x - 1 = 0$.

Since $f(1)$ is negative and $f(2)$ positive, a root lies between 1 and 2 and, therefore, we take $x_0 = 3/2$. Then

$$f(x_0) = \frac{27}{8} - \frac{3}{2} = \frac{15}{8}, \text{ which is positive.}$$

Hence the root lies between 1 and 1.5 and we obtain

$$x_1 = \frac{1+1.5}{2} = 1.25$$

We find $f(x_1) = -19/64$, which is negative. We, therefore, conclude that the root lies between 1.25 and 1.5. If follows that

$$x_2 = \frac{1.25+1.5}{2} = 1.375$$

The procedure is repeated and the successive approximations are

$$x_3 = 1.3125, \qquad x_4 = 1.34375, \qquad x_5 = 1.328125, \text{ etc.}$$

Example 2.2 Find a real root of the equation $x^3 - 2x - 5 = 0$.

Let $f(x) = x^3 - 2x - 5$. Then

$$f(2) = -1 \quad \text{and} \quad f(3) = 16.$$

Hence a root lies between 2 and 3 and we take

$$x_1 = \frac{2+3}{2} = 2.5$$

Since $f(x_1) = f(2.5) = 5.6250$, the root lies between 2 and 2.25. Hence

$$x_2 = \frac{2+2.5}{2} = 2.25$$

Now, $f(x_2) = 1.890625$, the root lies between 2 and 2.25. Therefore,

$$x_3 = \frac{2+2.25}{2} = 2.125$$

Since $f(x_3) = 0.3457$, the root lies between 2 and 2.125. Therefore,

$$x_4 = \frac{2+2.125}{2} = 2.0625$$

Proceeding in this way, we obtain the successive approximations:

$$x_5 = 2.09375, \quad x_6 = 2.10938, \quad x_7 = 2.10156,$$
$$x_8 = 2.09766, \quad x_9 = 2.09570, \quad x_{10} = 2.09473,$$
$$x_{11} = 2.09424, \ \ldots$$

We find

$$x_{11} - x_{10} = -0.0005,$$

and

$$\left|\frac{x_{11} - x_{10}}{x_{11}}\right| \times 100 = \frac{0.0005}{2.09424} \times 100 = 0.02\%$$

Hence a root, correct to three decimal places, is 2.094.

Example 2.3 Find a real root of $f(x) = x^3 + x^2 + x + 7 = 0$ correct to three decimal places.

The given equation is a cubic and the last term is positive. Hence, $f(x) = 0$ will have a negative real root. We find that

$$f(-1) = 6, \quad f(-2) = 1 \text{ and } f(-3) = -14.$$

Therefore, a real root lies between –3 and –2.

We take

$$x_1 = \frac{-2-3}{2} = -2.5$$

Since $f(-2.5) = -4.875$, the root lies between –2 and –2.5, and then

$$x_2 = \frac{-2-2.5}{2} = -2.25$$

Now $f(x_2) = -1.5781$, and, therefore, the root lies between –2 and –2.25.

It follows that

$$x_3 = \frac{-4.25}{2} = -2.125$$

Successive approximations are given by

$$\begin{aligned} x_4 &= -2.0625, & x_5 &= -2.0938, & x_6 &= -2.1094, \\ x_7 &= -2.1016, & x_8 &= -2.1055, & x_9 &= -2.1035, \\ x_{10} &= -2.1045, & x_{11} &= -2.1050, \ \ldots \end{aligned}$$

The difference between x_{10} and x_{11} is 0.0005. Hence, we conclude that the root is given by $x = -2.105$, correct to three decimal places.

Example 2.4 Find the positive root, between 0 and 1, of the equation $x = e^{-x}$ to a tolerance of 0.05%.

Let

$$f(x) = xe^x - 1 = 0$$

We have, $f(0) = -1$ and $f(1) = e - 1$, which is positive. Hence, a root exists between 0 and 1, and

$$x_1 = \frac{0+1}{2} = 0.5$$

Because, $f(x_1) = -0.1756$, the root lies between 0.5 and 1.0.

Then

$$x_2 = \frac{0.5+1.0}{2} = 0.75$$

Now, the tolerance ε_1 is given by

$$\varepsilon_1 = \left|\frac{x_2 - x_1}{x_2}\right| \times 100$$

$$= \frac{0.25}{0.75} \times 100 = 33.33\%$$

since $f(x_2) = 0.5878$, the root lies between 0.5 and 0.75.
Therefore,

$$x_3 = \frac{0.5 + 0.75}{2} = 0.625$$

also,

$$\varepsilon_2 = \left|\frac{0.625 - 0.75}{0.625}\right| \times 100 = 20\%.$$

Proceeding in this way, successive approximations and tolerances are obtained:

$x_4 = 0.5625$, $\varepsilon_3 = 11.11\%$; $x_5 = 0.5938$, $\varepsilon_4 = 5.26\%$;
$x_6 = 0.5781$, $\varepsilon_5 = 2.71\%$; $x_7 = 0.5703$, $\varepsilon_6 = 1.37\%$;
$x_8 = 0.5664$, $\varepsilon_7 = 0.69\%$; $x_9 = 0.5684$, $\varepsilon_8 = 0.35\%$;
$x_{10} = 0.5674$, $\varepsilon_9 = 0.18\%$; $x_{11} = 0.5669$, $\varepsilon_{10} = 0.09\%$;
$x_{12} = 0.5671$, $\varepsilon_{11} = 0.035\%$

Since $\varepsilon_{11} = 0.035\% < 0.05\%$, the required root is 0.567, correct to three decimal places.

Example 2.5 Find a root, correct to three decimal places and lying between 0 and 0.5, of the equation

$$4e^{-x} \sin x - 1 = 0$$

Let

$$f(x) = 4e^{-x} \sin x - 1$$

We have $f(0) = -1$ and $f(0.5) = 0.163145$
Therefore,

$$x_1 = 0.25$$

Since $f(0.25) = -0.22929$, it follows that the root lies between 0.25 and 0.5. Therefore,

$$x_2 = \frac{0.75}{2} = 0.375$$

The successive approximations are given by

$x_3 = 0.3125$, $x_4 = 0.3438$, $x_5 = 0.3594$,
$x_6 = 0.3672$, $x_7 = 0.3711$, $x_8 = 0.3692$,
$x_9 = 0.3702$, $x_{10} = 0.3706$, $x_{11} = 0.3704$
$x_{12} = 0.3705$, …

Hence the required root is 0.371, correct to three decimal places.

2.3 METHOD OF FALSE POSITION

This is the oldest method for finding the real root of a nonlinear equation $f(x) = 0$ and closely resembles the bisection method. In this method, also known as *regula–falsi* or the *method of chords*, we choose two points a and b such that $f(a)$ and $f(b)$ are of opposite signs. Hence, a root must lie in between these points. Now, the equation of the chord joining the two points $[a, f(a)]$ and $[b, f(b)]$ is given by

$$\frac{y - f(a)}{x - a} = \frac{f(b) - f(a)}{b - a}. \tag{2.6}$$

The method consists in replacing the part of the curve between the points $[a, f(a)]$ and $[b, f(b)]$ by means of the *chord* joining these points, and taking the point of intersection of the chord with the x-axis as an *approximation* to the root. The point of intersection in the present case is obtained by putting $y = 0$ in Eq. (2.6). Thus, we obtain

$$x_1 = a - \frac{f(a)}{f(b) - f(a)}(b - a) = \frac{af(b) - bf(a)}{f(b) - f(a)}, \tag{2.7}$$

which is the *first approximation* to the root of $f(x) = 0$. If now $f(x_1)$ and $f(a)$ are of opposite signs, then the root lies between a and x_1, and we replace b by x_1 in Eq. (2.7), and obtain the *next* approximation. Otherwise, we replace a by x_1 and generate the next approximation. The procedure is repeated till the root is obtained to the desired accuracy. Figure 2.2 gives a graphical representation of the method. The error criterion Eq. (2.5) can be used in this case also.

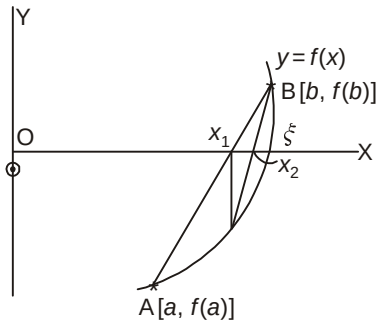


Figure 2.2 Method of false position.

Example 2.6 Find a real root of the equation:

$$f(x) = x^3 - 2x - 5 = 0.$$

We find $f(2) = -1$ and $f(3) = 16$. Hence $a = 2$, $b = 3$, and a root lies between 2 and 3. Equation (2.7) gives

$$x_1 = \frac{2(16) - 3(-1)}{16 - (-1)} = \frac{35}{17} = 2.058823529.$$

Now, $f(x_1) = -0.390799917$ and hence the root lies between 2.058823529 and 3.0. Using formula (2.7), we obtain

$$x_2 = \frac{2.058823529(16) - 3(-0.390799917)}{16.390799917} = 2.08126366.$$

Since $f(x_2) = -0.147204057$, it follows that the root lies between 2.08126366 and 3.0. Hence, we have

$$x_3 = \frac{2.08126366(16) - 3(-0.147204057)}{16.147204057} = 2.089639211.$$

Proceeding in this way, we obtain successively:

$$x_4 = 2.092739575, \qquad x_5 = 2.09388371,$$

$$x_6 = 2.094305452, \qquad x_7 = 2.094460846,\ldots$$

The correct value is 2.0945…, so that x_7 is correct to five significant figures.

Example 2.7 Given that the equation $x^{2.2} = 69$ has a root between 5 and 8. Use the method of regula–falsi to determine it.

Let $f(x) = x^{2.2} - 69$. We find

$$f(5) = -34.50675846 \quad \text{and} \quad f(8) = 28.00586026.$$

Hence

$$x_1 = \frac{5(28.00586026) - 8(-34.50675846)}{28.00586026 + 34.50675846} = 6.655990062.$$

Now, $f(x_1) = -4.275625415$ and therefore, the root lies between 6.655990062 and 8.0. We obtain

$$x_2 = 6.83400179, \qquad x_3 = 6.850669653.$$

The correct root is 6.8523651…, so that x_3 is correct to three significant figures.

Example 2.8 The equation $2x = \log_{10}x + 7$ has a root between 3 and 4. Find this root, correct to three decimal places, by regula–falsi method.

Let

$$f(x) = 2x - \log_{10}x - 7,\ a = 3 \text{ and } b = 4.$$

Then we find

$$f(3) = -1.4771 \text{ and } f(4) = 0.3979.$$

Hence

$$x_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}$$
$$= \frac{3(0.3979) - 4(-1.4771)}{0.3979 + 1.4771}$$
$$= \frac{7.1021}{1.8750} = 3.7878.$$

Therefore, the root lies between 3 and 3.7878. Now, we take $a = 3$ and $b = 3.7878$ Then,

$$f(b) = 2(3.7878) - \log_{10} 3.7878 - 7 = -0.002787$$

Hence,

$$x_2 = \frac{3(-0.002787) - 3.7878(-1.4771)}{-0.002787 + 1.4771}$$
$$= 3.7893,$$

and

$$f(x_2) = 2(3.7893) - \log_{10}(3.7893) - 7$$
$$= 0.000041,$$

which shows that $x = 3.789$ is the root correct to three decimal places.

Example 2.9 Find a root of the equation $4e^{-x} \sin x - 1 = 0$ by regular–falsi method given that the root lies between 0 and 0.5.

Let

$$f(x) = 4e^{-x} \sin x - 1, \; a = 0, \; b = 0.5.$$

We have

$$f(a) = -1 \text{ and } f(b) = 4e^{-0.5} \sin 0.5 - 1 = 0.163145$$

Therefore,

$$x_1 = \frac{0(0.163145) - 0.5(-1)}{1.163145}$$
$$= \frac{0.5}{1.163145} = 0.4298690$$

Now, we take

$$a = 0 \quad \text{and} \quad b = 0.4298690$$

Then

$$f(x) = 0.08454$$

Therefore,

$$x_2 = \frac{0(0.08454) - 0.42987(-1)}{1.08454}$$
$$= 0.39636$$

Now,

$$a = 0, \quad b = 0.39636 \quad \text{and} \quad f(b) = 0.038919$$

Hence

$$x_3 = \frac{0(0.038919) - 0.39636(-1)}{1.038919}$$
$$= 0.381512,$$

and

$$f(x_3) = 0.016934$$

Taking $a = 0$ and $b = 0.381512$, we obtain

$$x_4 = \frac{0(0.016934) - 0.381512(-1)}{1.016934}$$
$$= 0.375159,$$

and

$$f(x_4) = 0.0071873$$

Proceeding as above, we obtain

$$x_5 = 0.37248, \qquad x_6 = 0.37136,$$
$$x_7 = 0.37089, \qquad x_8 = 0.370697$$

It follows that the required root is 0.371, correct to three decimal places.

2.4 ITERATION METHOD

We have so far discussed root-finding methods which require an interval in which the root lies. We now describe methods which require one or more approximate values to start the solution and these values need not necessarily bracket the root. The first is the *iteration* method which requires one starting value of x.

To describe this method for finding a root of the equation

$$f(x) = 0, \tag{2.1}$$

we rewrite this equation in the form

$$x = \phi(x) \tag{2.8}$$

There are many ways of doing this. For example, the equation

$$x^3 + x^2 - 2 = 0$$

can be expressed in different forms

$$x = \sqrt{\frac{2}{1+x}}, \quad x = \sqrt{2 - x^3}, \quad x = (2 - x^2)^{1/3}, \text{etc.}$$

Now, let x_0 be an approximate root of Eq. (2.8). Then, substituting in Eq. (2.8), we get the first approximation as

$$x_1 = \phi(x_0)$$

Successive substitutions give the approximations

$$x_2 = \phi(x_1), \qquad x_3 = \phi(x_2), \ldots, \qquad x_n = \phi(x_{n-1}).$$

The preceding sequence may not converge to a definite number. But if the sequence converges to a definite number ξ, then ξ will be a root of the equation $x = \phi(x)$. To show this, let

$$x_{n+1} = \phi(x_n) \tag{2.9}$$

be the relation between the nth and $(n + 1)$th approximations. As n increases, $x_{n+1} \to \xi$ and if $\phi(x)$ is a continuous function, then $\phi(x_n) \to \phi(\xi)$. Hence, in the limit, we obtain

$$\xi = \phi(\xi), \tag{2.10}$$

which shows that ξ is a root of the equation $x = \phi(x)$.

To establish the condition of convergence of Eq. (2.8), we proceed in the following way:

From Eq. (2.9), we have

$$x_1 = \phi(x_0) \tag{2.11}$$

From Eqs. (2.10) and (2.11), we get

$$\begin{aligned} \xi - x_1 &= \phi(\xi) - \phi(x_0) \\ &= (\xi - x_0)\,\phi'(\xi_0), \quad x_0 < \xi_0 < \xi, \end{aligned} \tag{2.12}$$

on using Theorem 1.5. Similarly, we obtain

$$\xi - x_2 = (\xi - x_1)\ \phi'(\xi_1), \quad x_1 < \xi_1 < \xi \tag{2.13}$$

$$\xi - x_3 = (\xi - x_2)\ \phi'(\xi_2), \quad x_2 < \xi_2 < \xi \tag{2.14}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\xi - x_{n+1} = (\xi - x_n)\ \phi'(\xi_n), \quad x_n < \xi_n < \xi \tag{2.15}$$

If we assume

$$|\phi'(\xi_i)| \le k \text{ (for all } i), \tag{2.16}$$

then Eqs. (2.12) to (2.15) give

$$\left.\begin{aligned} |\xi - x_1| &\le k\,|\xi - x_0| \\ |\xi - x_2| &\le k\,|\xi - x_1| \\ |\xi - x_3| &\le k\,|\xi - x_2| \\ &\vdots \\ |\xi - x_{n+1}| &\le k\,|\xi - x_n| \end{aligned}\right\} \tag{2.17}$$

Multiplying the corresponding sides of the above equations, we obtain

$$|\xi - x_{n+1}| \le k^{n+1}\,|\xi - x_0| \tag{2.18}$$

If $k < 1$, i.e., if $|\phi'(\xi_i)| < 1$, then the right side of Eq. (2.18) tends to zero and the sequence of approximations $x_0, x_1, x_2, \ldots$ converges to the root ξ. Thus, when we express the equation $f(x) = 0$ in the form $x = \phi(x)$, then $\phi(x)$ must be such that

$$|\,\phi'(x)\,| < 1$$

in an immediate neighbourhood of the root. It follows that if *the initial approximation x_0 is chosen in an interval containing the root ξ, then the sequence of approximations converges to the root ξ.*

Now, we show that the root so obtained is *unique*. To prove this, let ξ_1 and ξ_2 be two roots of the equation $x = \phi(x)$. Then, we must have

$$\xi_1 = \phi(\xi_1) \quad \text{and} \quad \xi_2 = \phi(\xi_2).$$

Therefore,

$$\begin{aligned} |\xi_1 - \xi_2| &= |\phi(\xi_1) - \phi(\xi_2)| \\ &= |\xi_1 - \xi_2|\,\phi'(\eta),\ \eta \in (\xi_1, \xi_2) \end{aligned}$$

Hence,

$$|\xi_1 - \xi_2|\,[1 - \phi'(\eta)] = 0 \tag{2.19}$$

Since $|\phi'(\eta)| < 1$, it follows that $\xi_1 = \xi_2$, which proves that the root obtained is unique.

Finally, we shall find the error in the root obtained. We have

$$\begin{aligned} |\xi - x_n| &\le k|\xi - x_{n-1}| \\ &= k|\xi - x_n + x_n - x_{n-1}| \\ &\le k\big[|\xi - x_n| + |x_n - x_{n-1}|\big] \end{aligned}$$

$$\Rightarrow \qquad |\xi - x_n| \le \frac{k}{1-k}|x_n - x_{n-1}| = \frac{k}{1-k}k^{n-1}|x_1 - x_0|$$

$$\le \frac{k^n}{1-k}|x_1 - x_0|, \tag{2.20}$$

which shows that the convergence would be faster for smaller values of k.

Now, let ε be the specified accuracy so that

$$|\xi - x_n| \le \varepsilon$$

Then, Eq. (2.20) gives

$$|x_n - x_{n-1}| \le \frac{1-k}{k}\varepsilon, \tag{2.21}$$

which can be used to find the difference between two successive approximations (or *iterations*) to achieve a prescribed accuracy.

Example 2.10 Find a real root of the equation $x^3 = 1 - x^2$ on the interval [0, 1] with an accuracy of 10^{-4}.

We rewrite the equation as

$$x = \frac{1}{\sqrt{x+1}} \tag{i}$$

Here

$$\phi(x) = \frac{1}{\sqrt{x+1}} = (x+1)^{-1/2}$$

Therefore,

$$\phi'(x) = -\frac{1}{2}(x+1)^{-3/2} = -\frac{1}{2\sqrt{(x+1)^3}} < 1 \text{ in } [0, 1].$$

Also,

$$\max|\phi'(x)| = \frac{1}{2\sqrt{8}} = \frac{1}{4\sqrt{2}} = k < 0.2$$

Therefore, Eq. (2.21) gives

$$|x_n - x_{n-1}| \le \frac{1-0.2}{0.2}\varepsilon = 4\times 10^{-4} = 0.0004.$$

Taking $x_0 = 0.75$, we find

$$x_1 = \frac{1}{\sqrt{1.75}} = 0.75593,$$

$$x_2 = \frac{1}{\sqrt{1.75593}} = 0.75465,$$

$$x_3 = \frac{1}{\sqrt{1.75465}} = 0.75493.$$

Now, $|x_3 - x_2| = 0.00028 < 0.0004$. Hence, the required root is 0.7549, correct to four decimal places.

Example 2.11 Find a real root, correct to three decimal places, of the equation

$$2x - 3 = \cos x$$

lying in the interval $\left[\frac{3}{2}, \frac{\pi}{2}\right]$.

We rewrite the given equation as

$$x = \frac{1}{2}(\cos x + 3)$$

Here

$$|\phi'(x)| = \frac{1}{2}|\sin x| < 1, \text{ in } \left[\frac{3}{2}, \frac{\pi}{2}\right].$$

Choosing $x_0 = 1.5$, we obtain successively:

$$x_1 = \frac{1}{2}(\cos 1.5 + 3) = 1.5354,$$

$$x_2 = \frac{1}{2}(\cos 1.5354 + 3) = 1.5177,$$

$$x_3 = \frac{1}{2}(\cos 1.5177 + 3) = 1.5265,$$

$$x_4 = \frac{1}{2}(\cos 1.5265 + 3) = 1.5221,$$

$$x_5 = \frac{1}{2}(\cos 1.5221 + 3) = 1.5243,$$

$$x_6 = \frac{1}{2}(\cos 1.5243 + 3) = 1.5232,$$

$$x_7 = \frac{1}{2}(\cos 1.5232 + 3) = 1.5238.$$

Now, $| x_7 - x_6 | = 0.0006 < 0.001$. Hence, the root, correct to three decimal places is 1.524.

Example 2.12 Use the method of iteration to find a positive root of the equation $xe^x = 1$, given that a root lies between 0 and 1.

Writing the equation in the form

$$x = e^{-x},$$

we find that

$$\phi'(x) = -e^{-x} = -\frac{1}{e} \qquad \text{for } x = 1$$

Therefore,

$$| \phi'(x) | < 1.$$

Choosing $x_0 = 0.5$, we find

$x_1 = e^{-0.5} = 0.60653,$	$x_2 = 0.54524,$
$x_3 = 0.57970,$	$x_4 = 0.56007,$
$x_5 = 0.57117,$	$x_6 = 0.56486,$
$x_7 = 0.56844,$	$x_8 = 0.56641,$
$x_9 = 0.56756,$	$x_{10} = 0.56691,$
$x_{11} = 0.56728,$	$x_{12} = 0.56706,$
$x_{13} = 0.56719,$	$x_{14} = 0.56712,$
$x_{15} = 0.56716,$	$x_{16} = 0.56713,$
$x_{17} = 0.56715,$	$x_{18} = 0.56714,$
$x_{19} = 0.56714.$	

It follows that the root, correct to four decimal places, is 0.5671.

Example 2.13 Use the iterative method to find a real root of the equation $\sin x = 10(x - 1)$. Give your answer correct to three decimal places.

Let

$$f(x) = \sin x - 10x + 10$$

We find, from graph, that a root lies between 1 and π (The student is advised to draw the graphs). Rewriting the equation as

$$x = 1 + \frac{\sin x}{10},$$

we have

$$\phi(x) = 1 + \frac{\sin x}{10},$$

and

$$|\phi'(x)| = \frac{\cos x}{10} < 1 \text{ in } 1 \le x \le \pi.$$

Taking $x_0 = 1$, we obtain the successive iterates as:

$$x_1 = 1 + \frac{\sin 1}{10} = 1.0841,$$

$$x_2 = 1 + \frac{\sin 1.0841}{10} = 1.0884,$$

$$x_3 = 1 + \frac{\sin 1.0884}{10} = 1.0886,$$

$$x_4 = 1 + \frac{\sin 1.0886}{10} = 1.0886.$$

Hence the required root is 1.089.

Acceleration of convergence: Aitken's Δ^2-process

From the relation

$$|\xi - x_{n+1}| = |\phi(\xi) - \phi(x_n)| \le k\,|\xi - x_n|, \qquad k < 1$$

it is clear that the iteration method is linearly convergent. This slow rate of convergence can be accelerated by using Aitken's method, which is described below.

Let x_{i-1}, x_i, x_{i+1} be three successive approximations to the desired root $x = \xi$ of the equation $x = \phi(x)$. From Eq. (2.17), we know that

$$\xi - x_i = k(\xi - x_{i-1}), \qquad \xi - x_{i+1} = k(\xi - x_i)$$

Dividing, we obtain

$$\frac{\xi - x_i}{\xi - x_{i+1}} = \frac{\xi - x_{i-1}}{\xi - x_i},$$

which gives on simplification

$$\xi = x_{i+1} - \frac{(x_{i+1} - x_i)^2}{x_{i+1} - 2x_i + x_{i-1}}. \tag{2.22}$$

If we now define Δx_i and $\Delta^2 x_i$ by the relations

$$\Delta x_i = x_{i+1} - x_i \quad \text{and} \quad \Delta^2 x_i = \Delta(\Delta x_i),$$

then

$$\begin{aligned}
\Delta^2 x_{i-1} &= \Delta(\Delta x_{i-1}) \\
&= \Delta(x_i - x_{i-1}) \\
&= \Delta x_i - \Delta x_{i-1} \\
&= x_{i+1} - x_i - (x_i - x_{i-1}) \\
&= x_{i+1} - 2x_i + x_{i-1}.
\end{aligned}$$

Hence Eq. (2.22) can be written in the simpler form

$$\xi = x_{i+1} - \frac{(\Delta x_i)^2}{\Delta^2 x_{i-1}}. \tag{2.23}$$

which explains the term Δ^2-process.

In any numerical application, the values of the following underlined quantities must be obtained.

x_{i-1}		
	Δx_{i-1}	
x_i		$\underline{\Delta^2 x_{i-1}}$
	$\underline{\Delta x_i}$	
$\underline{x_{i+1}}$		

Example 2.14 We consider again Example 2.11, viz., the equation

$$x = \frac{1}{2}(3 + \cos x)$$

As before,

$x_1 = 1.5$		
	0.035	
$x_2 = 1.535$		-0.052
	-0.017	
$x_3 = 1.518$		

Hence we obtain from Eq. (2.23)

$$x_4 = 1.518 - \frac{(-0.017)^2}{-0.052} = 1.524,$$

which corresponds to six normal iterations.

2.5 NEWTON–RAPHSON METHOD

This method is generally used to improve the result obtained by one of the previous methods. Let x_0 be an approximate root of $f(x) = 0$ and let $x_1 = x_0 + h$ be the correct root so that $f(x_1) = 0$. Expanding $f(x_0 + h)$ by Taylor's series, we obtain

$$f(x_0) + hf'(x_0) + \frac{h^2}{2!} f''(x_0) + \cdots = 0.$$

Neglecting the second and higher-order derivatives, we have

$$f(x_0) + hf'(x_0) = 0,$$

which gives

$$h = -\frac{f(x_0)}{f'(x_0)}.$$

A better approximation than x_0 is, therefore, given by x_1, where

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{2.24a}$$

Successive approximations are given by $x_2, x_3, \ldots, x_{n+1}$, where

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{2.24b}$$

which is the *Newton–Raphson formula*.

If we compare Eq. (2.24b) with the relation

$$x_{n+1} = \phi(x_n)$$

of the iterative method [see Eq. (2.9)], we obtain

$$\phi(x) = x - \frac{f(x)}{f'(x)},$$

which gives

$$\phi'(x) = \frac{f(x)\, f''(x)}{[f'(x)]^2}. \tag{2.25}$$

To examine the convergence we assume that $f(x)$, $f'(x)$ and $f''(x)$ are continuous and bounded on any interval containing the root $x = \xi$ of the equation $f(x) = 0$. If ξ is a simple root, then $f'(x) \neq 0$. Further since $f'(x)$ is continuous, $|f'(x)| \geq \varepsilon$ for some $\varepsilon > 0$ in a suitable neighbourhood of ξ. Within this neighbourhood we can select an interval such that $|f(x)\, f''(x)| < \varepsilon^2$ and this is possible since $f(\xi) = 0$ and since $f(x)$ is continuously twice differentiable. Hence, in this interval we have

$$|\phi'(x)| < 1. \tag{2.26}$$

Therefore, Newton–Raphson formula given in Eq. (2.24b) converges, provided that the initial approximation x_0 is chosen sufficiently close to ξ. When ξ is a multiple root, the Newton–Raphson method still converges but slowly. Convergence can, however, be made faster by modifying Eq. (2.24b). This will be discussed later.

To obtain the rate of convergence of the method, we note that $f(\xi) = 0$ so that Taylor's expansion gives

$$f(x_n) + (\xi - x_n)\, f'(x_n) + \frac{1}{2}(\xi - x_n)^2\, f''(x_n) + \cdots = 0,$$

from which we obtain

$$-\frac{f(x_n)}{f'(x_n)} = (\xi - x_n) + \frac{1}{2}(\xi - x_n)^2 \frac{f''(x_n)}{f'(x_n)} \tag{2.27}$$

From Eqs. (2.24b) and (2.27), we have

$$x_{n+1} - \xi = \frac{1}{2}(x_n - \xi)^2 \frac{f''(x_n)}{f'(x_n)} \tag{2.28}$$

Setting

$$\varepsilon_n = x_n - \xi, \tag{2.29}$$

Equation (2.28) gives

$$\varepsilon_{n+1} \approx \frac{1}{2}\varepsilon_n^2 \frac{f''(\xi)}{f'(\xi)}, \tag{2.30}$$

so that the Newton–Raphson process has a second-order or quadratic convergence.

Geometrically, the method consists in replacing the part of the curve between the point $[x_0, f(x_0)]$ and the x-axis by means of the tangent to the curve at the point, and is described graphically in Fig. 2.3. It can be used for solving both algebraic and transcendental equations and it can also be used when the roots are complex.

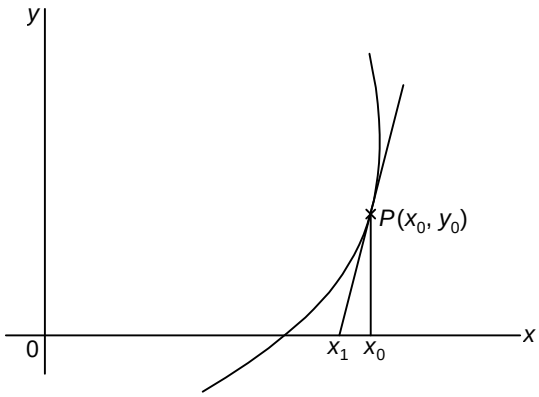


Figure 2.3 Newton–Raphson method.

Example 2.15 Use the Newton–Raphson method to find a root of the equation $x^3 - 2x - 5 = 0$.

Here $f(x) = x^3 - 2x - 5$ and $f'(x) = 3x^2 - 2$. Hence Eq. (2.24b) gives:

$$x_{n+1} = x_n - \frac{x_n^3 - 2x_n - 5}{3x_n^2 - 2} \tag{i}$$

Choosing $x_0 = 2$, we obtain $f(x_0) = -1$ and $f'(x_0) = 10$. Putting $n = 0$ in Eq. (i), we obtain

$$x_1 = 2 - \left(-\frac{1}{10}\right) = 2.1$$

Now,

$$f(x_1) = (2.1)^3 - 2(2.1) - 5 = 0.061,$$

and

$$f'(x_1) = 3(2.1)^2 - 2 = 11.23.$$

Hence

$$x_2 = 2.1 - \frac{0.061}{11.23} = 2.094568.$$

This example demonstrates that Newton–Raphson method converges more rapidly than the methods described in the previous sections, since this requires fewer iterations to obtain a specified accuracy. But since two function evaluations are required for each iteration, Newton–Raphson method requires more computing time.

Example 2.16 Find a root of the equation $x \sin x + \cos x = 0$.

We have

$$f(x) = x \sin x + \cos x \quad \text{and} \quad f'(x) = x \cos x.$$

The iteration formula is, therefore,

$$x_{n+1} = x_n - \frac{x_n \sin x_n + \cos x_n}{x_n \cos x_n}.$$

With $x_0 = \pi$, the successive iterates are given below

n	x_n	$f(x_n)$	x_{n+1}
0	3.1416	–1.0	2.8233
1	2.8233	–0.0662	2.7986
2	2.7986	–0.0006	2.7984
3	2.7984	0.0	2.7984

Example 2.17 Find a real root of the equation $x = e^{-x}$, using the Newton–Raphson method.

We write the equation in the form

$$f(x) = xe^x - 1 = 0 \tag{i}$$

Let $x_0 = 1$. Then

$$x_1 = 1 - \frac{e-1}{2e} = \frac{1}{2}\left(1 + \frac{1}{e}\right) = 0.6839397$$

Now

$$f(x_1) = 0.3553424, \quad \text{and} \quad f'(x_1) = 3.337012,$$

so that

$$x_2 = 0.6839397 - \frac{0.3553424}{3.337012} = 0.5774545$$

Proceeding in this way, we obtain

$$x_3 = 0.5672297 \quad \text{and} \quad x_4 = 0.5671433.$$

Example 2.18 Using Newton–Raphson method, find a real root, correct to 3 decimal places, of the equation $\sin x = x/2$ given that the root lies between $\pi/2$ and π.

Let

$$f(x) = \sin x - \frac{x}{2}$$

Then

$$f'(x) = \cos x - \frac{1}{2}$$

Choosing $x_0 = \dfrac{\pi}{2}$, we obtain

$$x_1 = \frac{\pi}{2} - \frac{\sin\frac{\pi}{2} - \frac{\pi}{4}}{-\frac{1}{2}} = 2,$$

$$x_2 = 2 - \frac{\sin 2 - 1}{\cos 2 - \frac{1}{2}} = 1.9010,$$

$$x_3 = 1.9010 - \frac{\sin 1.9010 - 0.9505}{\cos 1.9010 - 0.5} = 1.8955$$

Similarly, $x_4 = 1.8954$, $x_5 = 1.8955$, … . Hence the required root is $x = 1.896$.

Example 2.19 Given the equation $4e^{-x} \sin x - 1 = 0$, find the root between 0 and 0.5 correct to three decimal places.

Let $\quad f(x) = 4e^{-x} \sin x - 1$ and $x_0 = 0.2$.

Then

$$f(x_0) = -0.349373236,$$

and

$$f'(x_0) = 2.559015826.$$

Therefore,

$$x_1 = 0.2 + \frac{0.349373236}{2.559015826}$$
$$= 0.336526406 = 0.33653.$$

Now,

$$f(x_1) = -\ 0.056587$$

and

$$f'(x_1) = 1.753305735$$
$$= 1.75330$$

Therefore,

$$x_2 = 0.33653 + \frac{0.056587}{1.75330} = 0.36880$$

For the next approximation, we find $f(x_2) = -0.00277514755$ and $f'(x_2) = 1.583028705$. This gives

$$x_3 = 0.36880 - 0.00175 = 0.37055$$

since $f(x_3) = -0.00001274$, it follows that the required root is given by $x = 0.370$.

Generalized Newton's method

If ξ is a root of $f(x) = 0$ with multiplicity p, then the iteration formula corresponding to Eq. (2.24) is taken as

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)}, \tag{2.31}$$

which means that $(1/p)f'(x_n)$ is the slope of the straight line passing through (x_n, y_n) and intersecting the x-axis at the point $(x_{n+1}, 0)$.

Equation (2.31) is called the *generalized Newton's formula* and reduces to Eq. (2.24) for $p = 1$. Since ξ is a root of $f(x) = 0$ with multiplicity p, it follows that ξ is also a root of $f'(x) = 0$ with multiplicity $(p - 1)$, of $f''(x) = 0$ with multiplicity $(p - 2)$, and so on. Hence the expressions

$$x_0 - p\frac{f(x_0)}{f'(x_0)}, \quad x_0 - (p-1)\frac{f'(x_0)}{f''(x_0)}, \quad x_0 - (p-2)\frac{f''(x_0)}{f'''(x_0)}$$

must have the same value if there is a root with multiplicity p, provided that the initial approximation x_0 is chosen sufficiently close to the root.

Example 2.20 Find a double root of the equation $f(x) = x^3 - x^2 - x + 1 = 0$.

Choosing $x_0 = 0.8$, we have

$$f'(x) = 3x^3 - 2x - 1, \quad \text{and} \quad f''(x) = 6x - 2.$$

With $x_0 = 0.8$, we obtain

$$x_0 - 2\frac{f(x_0)}{f'(x_0)} = 0.8 - 2\frac{0.072}{-(0.68)} = 1.012,$$

and

$$x_0 - \frac{f'(x_0)}{f''(x_0)} = 0.8 - \frac{(-0.68)}{2.8} = 1.043.$$

The closeness of these values indicates that there is a double root near to unity. For the next approximation, we choose $x_1 = 1.01$ and obtain

$$x_1 - 2\frac{f(x_1)}{f'(x_1)} = 1.01 - 0.0099 = 1.0001,$$

and

$$x_1 - \frac{f'(x_1)}{f''(x_1)} = 1.01 - 0.0099 = 1.0001.$$

We conclude, therefore, that there is a double root at $x = 1.0001$ which is sufficiently close to the actual root unity.

On the other hand, if we apply Newton–Raphson method with $x_0 = 0.8$, we obtain

$$x_1 = 0.8 + 0.106 \approx 0.91, \quad \text{and} \quad x_2 = 0.91 + 0.046 \approx 0.96.$$

It is clear that the generalized Newton's method converges more rapidly than the Newton–Raphson procedure.

2.6 RAMANUJAN'S METHOD

Srinivasa Ramanujan (1887–1920) described an iterative procedure* to determine the *smallest* root of the equation

$$f(x) = 0, \tag{2.1}$$

where $f(x)$ is of the form

$$f(x) = 1 - (a_1x + a_2x^2 + a_3x^3 + \cdots) \tag{2.32}$$

To explain the method of procedure, we consider the quadratic equation

$$f(x) = a_0x^2 + a_1x + a_2 = 0,$$

with the roots x_1 and x_2, such that $|x_1| < |x_2|$. Then the equation defined by

$$\phi(x) = a_2x^2 + a_1x + a_0 = 0$$

$$\Rightarrow \qquad 1 + \frac{a_1}{a_0}x + \frac{a_2}{a_0}x^2 = 0$$

will have roots $\dfrac{1}{x_1}$ and $\dfrac{1}{x_2}$ such that $\dfrac{1}{|x_1|} > \dfrac{1}{|x_2|}$.

Now,

$$\frac{1}{\phi(x)} = \left(1 + \frac{a_1}{a_0}x + \frac{a_2}{a_0}x^2\right)^{-1}$$

*See Berndt [1985], p. 41.

can be written as

$$\left(1+\frac{a_1}{a_0}x+\frac{a_2}{a_0}x^2\right)^{-1} = \frac{k_1}{x-\dfrac{1}{x_1}}+\frac{k_2}{x-\dfrac{1}{x_2}}$$

$$= \frac{-k_1x_1}{1-xx_1}+\frac{-k_2x_2}{1-xx_2}$$

$$= -k_1x_1(1-xx_1)^{-1}-k_2x_2(1-xx_2)^{-1}$$

$$= \sum_{i=0}^{\infty} b_i x^i, \quad \text{where } b_i = -\sum_{r=1}^{2} k_r x_r^{i+1}$$

Then,

$$\frac{b_{i-1}}{b_i} = \frac{k_1x_1^i+k_2x_2^i}{k_1x_1^{i+1}+k_2x_2^{i+1}} = \frac{\dfrac{k_1}{k_2}\left(\dfrac{x_1}{x_2}\right)^i+1}{\dfrac{k_1}{k_2}\left(\dfrac{x_1}{x_2}\right)^{i+1}+1}\cdot\frac{1}{x_2}$$

Since $\dfrac{x_1}{x_2}<1$, it follows that

$$\lim_{i\to\infty}\frac{b_{i-1}}{b_i} = \frac{1}{x_2},$$

which is the smallest root. This is the basis of Ramanujan's method which is outlined below.

To find the smallest root of $f(x) = 0$, we consider $f(x)$ in the form

$$f(x) = 1 - (a_1x + a_2x^2 + a_3x^3 + \cdots),$$

and then write

$$\left[1-(a_1x+a_2x^2+a_3x^3+\cdots)\right]^{-1} = b_1+b_2x+b_3x^2+\cdots \tag{2.33}$$

$$\Rightarrow \quad 1 + (a_1x + a_2x^2 + a_3x^3 + \cdots) + (a_1x + a_2x^2 + a_3x^3 + \cdots)^2 + \cdots$$
$$= b_1 + b_2x + b_3x^2 + \cdots \tag{2.34}$$

To find b_i, we equate coeffecients of like powers of x on both sides of Eq. (2.34). We then obtain

$$\left.\begin{aligned}
b_1 &= 1\\
b_2 &= a_1 = a_1b_1, \quad \text{since } b_1 = 1\\
b_3 &= a_2 + a_1^2 = a_2b_1 + a_1b_2, \quad \text{since } b_2 = a_1\\
&\vdots\\
b_k &= a_1b_{k-1} + a_2b_k + \cdots + a_{k-1}b_1\\
&= a_{k-1}b_1 + a_{k-2}b_2 + \cdots + a_1b_{k-1}
\end{aligned}\right\} \tag{2.35}$$

The ratios $\frac{b_{i-1}}{b_i}$, called the *convergents*, approach, in the limit, the *smallest* root of $f(x) = 0$. The method is demonstrated in the following examples.

Example 2.21 Find the smallest root of the equation

$$f(x) = x^3 - 9x^2 + 26x - 24 = 0$$

We have

$$f(x) = 1 - \frac{26}{24}x + \frac{9}{24}x^2 - \frac{1}{24}x^3$$

$$= 1 - \left(\frac{13}{12}x - \frac{3}{8}x^2 + \frac{1}{24}x^3\right)$$

Here

$$a_1 = \frac{13}{12},\ a_2 = -\frac{3}{8},\ a_3 = \frac{1}{24},\ a_4 = a_5 = \cdots = 0$$

Now,

$$b_1 = 1$$

$$b_2 = a_1 = \frac{13}{12} = 1.0833,$$

Therefore,

$$\frac{b_1}{b_2} = \frac{12}{13} = 0.923$$

$$b_3 = a_1 b_2 + a_2 b_1$$

$$= \frac{13}{12}(1.0833) - \frac{3}{8}(1) = 0.7986$$

Therefore,

$$\frac{b_2}{b_3} = \frac{1.0833}{0.7986} = 1.356$$

$$b_4 = a_1 b_3 + a_2 b_2 + a_3 b_1$$

$$= 1.0833\ (0.7986) + \left(-\frac{3}{8}\right)(1.0833) + \frac{1}{24}(1)$$

$$= 0.5007$$

Therefore,

$$\frac{b_3}{b_4} = 1.595$$

$$b_5 = a_1 b_4 + a_2 b_3 + a_3 b_2$$

$$= 1.0833(0.5007) + \left(-\frac{3}{8}\right)(0.7986) + \frac{1}{24}(1.0833)$$

$$= 0.2880$$

Therefore,

$$\frac{b_4}{b_5} = \frac{0.5007}{0.2880} = 1.7382.$$

$$b_6 = a_1 b_5 + a_2 b_4 + a_3 b_3$$
$$= 0.1575$$

Therefore,

$$\frac{b_5}{b_6} = 1.8286.$$

$$b_7 = 0.0835$$

Therefore,

$$\frac{b_6}{b_7} = 1.8862.$$

$$b_8 = 0.0434$$

Therefore,

$$\frac{b_7}{b_8} = 1.9240.$$

$$b_9 = 0.0223$$

Therefore,

$$\frac{b_8}{b_9} = 1.9462.$$

The roots of the given equation are 2, 3 and 4 and it can be seen that the successive convergents approach the value 2.

Example 2.22 Find a root of the equation $xe^x = 1$.

Let

$$xe^x = 1 \tag{i}$$

Expanding e^x in ascending powers of x and simplifying, we can rewrite Eq. (i) as

$$1 = x + x^2 + \frac{x^3}{2} + \frac{x^4}{6} + \frac{x^5}{24} + \cdots \tag{ii}$$

which is of the form of the right side of Eq. (2.32).

Here,

$$a_1 = 1, \quad a_2 = 1, \quad a_3 = \frac{1}{2}, \quad a_4 = \frac{1}{6}, \quad a_5 = \frac{1}{24}, \ \ldots$$

We then have

$$b_1 = 1,$$
$$b_2 = a_2 = 1,$$
$$b_3 = a_1 b_2 + a_2 b_1 = 1 + 1 = 2$$
$$b_4 = a_1 b_3 + a_2 b_2 + a_3 b_1 = 2 + 1 + \frac{1}{2} = \frac{7}{2},$$

$$b_5 = a_1b_4 + a_2b_3 + a_3b_2 + a_4b_1$$
$$= \frac{7}{2} + 2 + \frac{1}{2} + \frac{1}{6}$$
$$= \frac{37}{6} = 6.1667,$$
$$b_6 = \frac{261}{24} = 10.8750;$$

Hence, we obtain

$$b_1/b_2 = 1,$$
$$b_2/b_3 = 0.5,$$
$$b_3/b_4 = 0.5714,$$
$$b_4/b_5 = 0.5676,$$
$$b_5/b_6 = 0.5670.$$

It can be seen that Newton's method (see Example 2.17) gives the value 0.5671433 to this root.

Example 2.23 Find the smallest root, correct to 4 decimal places, of the equation

$$f(x) = 3x - \cos x - 1 = 0.$$

We have

$$f(x) = 1 - 3x + \cos x$$
$$= 1 - 3x + 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$
$$= 2 - 3x - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$
$$= 2\left[1 - \frac{3}{2}x - \frac{x^2}{4} + \frac{x^4}{48} - \frac{x^6}{1440} + \frac{x^8}{80640} - \cdots\right]$$

Now, let

$$1 - \left(\frac{3}{2}x + \frac{x^2}{4} - \frac{x^4}{48} + \frac{x^6}{1440} - \frac{x^8}{80640}\right)^{-1} = b_1 + b_2x + b_3x^2 + \cdots$$

Here

$$a_1 = \frac{3}{2},\ a_2 = \frac{1}{4},\ a_3 = 0,\ a_4 = -\frac{1}{48},$$
$$a_5 = 0,\ a_6 = \frac{1}{1440},\ a_7 = 0,\ a_8 = -\frac{1}{80640}, \ldots$$

we then obtain,

$$b_1 = 1, \qquad b_2 = 1.5,$$
$$b_3 = 2.5, \qquad b_4 = 4.125,$$
$$b_5 = 6.79167, \qquad b_6 = 11.18750,$$
$$b_7 = 18.42778, \qquad b_8 = 30.35365$$

The successive convergents are

$$\frac{b_1}{b_2} = 0.66667; \qquad \frac{b_2}{b_3} = 0.60000,$$
$$\frac{b_3}{b_4} = 0.60606; \qquad \frac{b_4}{b_5} = 0.60736,$$
$$\frac{b_5}{b_6} = 0.60708; \qquad \frac{b_6}{b_7} = 0.60710,$$
$$\frac{b_7}{b_8} = 0.607102$$

Hence the required root, correct to four decimal places, is 0.6071.

Example 2.24 Using Ramanujan's method, find a real root of the equation

$$1 - x + \frac{x^2}{(2!)^2} - \frac{x^3}{(3!)^2} + \frac{x^4}{(4!)^2} - \cdots = 0$$

Let

$$1 - \left[x - \frac{x^2}{(2!)^2} + \frac{x^3}{(3!)^2} - \frac{x^4}{(4!)^2} + \cdots \right] = 0 \tag{i}$$

we have

$$a_1 = 1, \qquad a_2 = -\frac{1}{(2!)^2}, \qquad a_3 = \frac{1}{(3!)^2},$$
$$a_4 = -\frac{1}{(4!)^2}, \qquad a_5 = \frac{1}{(5!)^2}, \qquad a_6 = -\frac{1}{(6!)^2}, \ldots$$

Then we obtain

$$b_1 = 1; \qquad b_2 = a_1 = 1;$$
$$b_3 = a_1 b_2 + a_2 b_1 = 1 - \frac{1}{(2!)^2} = \frac{3}{4};$$
$$b_4 = a_1 b_3 + a_2 b_2 + a_3 b_1$$
$$= \frac{3}{4} - \frac{1}{4} + \frac{1}{36} = \frac{19}{36};$$
$$b_5 = \frac{211}{576}, \ldots$$

The successive convergents are:

$$\frac{b_1}{b_2} = 1; \qquad \frac{b_2}{b_3} = \frac{4}{3} = 1.3333\ldots$$

$$\frac{b_3}{b_4} = \frac{27}{19} = 1.4210\ldots$$

$$\frac{b_4}{b_5} = 1.4408,$$

where the last result is correct to three significant figures.

2.7 SECANT METHOD

We have seen that the Newton–Raphson method requires the evaluation of derivatives of the function and this is not always possible, particularly in the case of functions arising in practical problems. In the secant method, the derivative at x_i is approximated by the formula

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}},$$

which can be written as

$$f_i' = \frac{f_i - f_{i-1}}{x_i - x_{i-1}}, \tag{2.36}$$

where $f_i = f(x_i)$. Hence, the Newton–Raphson formula becomes

$$x_{i+1} = x_i - \frac{f_i(x_i - x_{i-1})}{f_i - f_{i-1}} = \frac{x_{i-1} f_i - x_i f_{i-1}}{f_i - f_{i-1}}. \tag{2.37}$$

It should be noted that this formula requires two initial approximations to the root.

Example 2.25 Find a real root of the equation $x^3 - 2x - 5 = 0$ using secant method.

Let the two initial approximations be given by

$$x_{-1} = 2 \quad \text{and} \quad x_0 = 3$$

We have

$$f(x_{-1}) = f_1 = 8 - 9 = -1, \quad \text{and} \quad f(x_0) = f_0 = 27 - 11 = 16.$$

Putting $i = 0$ in Eq. (2.37), we obtain

$$x_1 = \frac{2(16) - 3(-1)}{17} = \frac{35}{17} = 2.058823529.$$

Also,

$$f(x_1) = f_1 = -0.390799923.$$

Putting $i = 1$ in Eq. (2.37), we obtain

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{3(-0.390799923) - 2.058823529(16)}{-16.390799923} = 2.08126366.$$

Again

$$f(x_2) = f_2 = -0.147204057.$$

Setting $i = 2$ in Eq. (2.37), and simplifying, we get $x_3 = 2.094824145$, which is correct to three significant figures.

Example 2.26 Using the secant method, find a real root of the equation

$$f(x) = xe^x - 1 = 0$$

We have

$$f(0) = -1 \quad \text{and} \quad f(1) = e - 1 = 1.71828 = f_1$$

Therefore, a root lies between 0 and 1.

Let

$$x_0 = 0 \quad \text{and} \quad x_1 = 1.$$

Therefore,

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{1}{2.71828} = 0.36788.$$

and

$$\begin{aligned} f_2 &= 0.36788e^{0.36788} - 1 \\ &= -0.46854. \end{aligned}$$

Hence

$$\begin{aligned} x_3 &= \frac{x_1 f_2 - x_2 f_1}{f_2 - f_1} \\ &= \frac{1(-0.46854) - 0.36788(1.71828)}{-0.46854 - 1.71828} \\ &= 0.50332 \end{aligned}$$

and

$$f_3 = -0.16740$$

Hence

$$x_4 = \frac{x_2 f_3 - x_3 f_2}{f_3 - f_2} = 0.57861$$

and

$$f_4 = 0.03198$$

Hence

$$x_5 = \frac{x_3 f_4 - x_4 f_3}{f_4 - f_3} = 0.56653.$$

and

$$f_5 = -0.00169$$

Therefore,

$$x_6 = \frac{x_4 f_5 - x_5 f_4}{f_5 - f_4} = 0.56714.$$

We also find

$$f(x_6) = -0.0001196.$$

It follows that the required root is 0.5671, correct to four decimal places.

2.8 MULLER'S METHOD

In this method, the given function $f(x)$ is approximated by a second degree curve in the vicinity of a root. The roots of the quadratic are then assumed to be the approximations to the roots of the equation $f(x) = 0$. The method is iterative and can be used to compute complex roots. It has quadratic convergence [see, Muller (1956)].

Let (x_{i-2}, y_{i-2}), (x_{i-1}, y_{i-1}) and (x_i, y_i) be three distinct points on the curve $y = f(x)$ where x_{i-2}, x_{i-1} and x_i are approximations to a root of $f(x) = 0$. Now, a second degree curve passing through the three points is given by Lagrange's formula (see Section 3.9.1)

$$L(x) = \frac{(x - x_{i-1})(x - x_i)}{(x_{i-2} - x_{i-1})(x_{i-2} - x_i)} y_{i-2} + \frac{(x - x_{i-2})(x - x_i)}{(x_{i-1} - x_{i-2})(x_{i-1} - x_i)} y_{i-1}$$

$$+ \frac{(x - x_{i-2})(x - x_{i-1})}{(x_i - x_{i-2})(x_i - x_{i-1})} y_i \tag{2.38}$$

Let

$$h_i = x_i - x_{i-1},\ h_{i-1} = x_{i-1} - x_{i-2} \tag{2.39}$$

Then

$$\left.\begin{aligned} &x - x_{i-1} = x - x_i + x_i - x_{i-1} = (x - x_i) + h_i, \\ &x - x_{i-2} = x - x_i + x_i - x_{i-2} = (x - x_i) + (h_{i-1} + h_i) \\ &x_{i-2} - x_{i-1} = -h_{i-1}, \\ &x_{i-2} - x_i = -(h_{i-1} + h_i) \text{ and } \Delta_i = y_i - y_{i-1}. \end{aligned}\right\} \tag{2.40}$$

Hence

$$L(x) = \frac{(x - x_i + h_i)(x - x_i)}{h_{i-1}(h_{i-1} + h_i)} y_{i-2} + \frac{(x - x_i + h_{i-1} + h_i)(x - x_i)}{-h_{i-1}h_i} y_{i-1}$$

$$+ \frac{(x - x_i + h_i + h_{i-1})(x - x_i + h_i)}{h_i\,(h_{i-1} + h_i)} y_i \tag{2.41}$$

After simplification, the preceding equation can be written as

$$L(x) = A(x - x_i)^2 + B(x - x_i) + y_i,$$

where

$$\left.\begin{aligned} A &= \frac{1}{(h_{i-1} + h_i)}\left(\frac{\Delta_i}{h_i} - \frac{\Delta_{i-1}}{h_{i-1}}\right) \\ \text{and} \quad B &= \frac{\Delta_i}{h_i} + Ah_i \end{aligned}\right\} \tag{2.42}$$

With these values of A and B, the quadratic Eq. (2.38) gives the next approximation x_{i+1}

$$x_{i+1} = x_i + \frac{-B \pm \sqrt{B^2 - 4Ay_i}}{2A} \tag{2.43}$$

Since Eq. (2.43) leads to inaccurate results, we take the equivalent form

$$x_{i+1} = x_i - \frac{2y_i}{B \pm \sqrt{B^2 - 4Ay_i}} \tag{2.44}$$

In Eq. (2.44), the sign in the denominator should be chosen so that the denominator will be largest in magnitude. With this choice, Eq. (2.44) gives the next approximation to the root.

Example 2.27 Using Muller's method, find the root of the equation

$$f(x) = x^3 - x - 1 = 0,$$

with the initial approximations

$$x_{i-2} = 0, \quad x_{i-1} = 1, \quad x_i = 2.$$

We have

$$y_{i-2} = -1, \quad y_{i-1} = -1, \quad y_i = 5.$$

Also,

$$h_i = 1, \quad h_{i-1} = 1,$$
$$\Delta_i = 6, \ \Delta_{i-1} = 0.$$

Hence, Eq. (2.42) gives $A = 3$ and $B = 9$.

Then

$$\sqrt{B^2 - 4Ay_i} = \sqrt{21}$$

therefore, Eq. (2.44) gives

$$x_{i+1} = 2 - \frac{2(5)}{9 + \sqrt{21}}, \quad \text{since the sign of } B \text{ is positive}$$
$$= 1.26376.$$

Error in the above result $= \left|\dfrac{1.26376 - 2}{1.26376}\right| 100 = 58\%.$

For the second approximation, we take

$$x_{i-2} = 1, \quad x_{i-1} = 2, \quad x_i = 1.26376.$$

The corresponding values of y are

$$y_{i-2} = -1, \quad y_{i-1} = 5, \quad y_i = -0.24542.$$

The computed values of A and B are

$$A = 4.26375 \quad \text{and} \quad B = 3.98546.$$

Then

$$x_{i+1} = 1.32174,$$

and the error in the above result = 4.39%.

For the third approximation, we take

$$x_{i-2} = 2, \quad x_{i-1} = 1.26376, \quad x_i = 1.32174.$$

$$y_{i-2} = 5, \quad y_{i-1} = -0.24542, \quad y_i = -0.01266.$$

Then $A = 4.58544$, $B = 4.28035$ and $x_{i+1} = 1.32469$.

Error in the result = 0.22%.

For the next approximation, we have

$$x_{i-2} = 1.26376, \quad x_{i-1} = 1.32174, \quad x_i = 1.32469$$

These values give

$$A = 3.87920, \quad B = 4.26229 \text{ and } x_{i+1} = 1.32472.$$

The error in this result = 0.002%.

Hence the required root is 1.3247, correct to 4 decimal places.

2.9 GRAEFFE'S ROOT-SQUARING METHOD

This is another method recommended for the numerical solution of polynomial equations. The method is outlined here by considering a cubic equation.

Let the cubic equation be

$$A_0x^3 + A_1x^2 + A_2x + A_3 = 0, \tag{2.45}$$

whose roots ξ_1, ξ_2 and ξ_3 are such that

$$|\xi_1| \gg |\xi_2| \gg |\xi_3|.$$

The symbol $\gg$ means "*much greater than*". In other words, the ratios $\frac{\xi_2}{\xi_1}, \frac{\xi_3}{\xi_2}$ are very small quantities compared to unity and can be neglected. In such a case, we say that *the roots are widely separated.*

We now consider the relations between the roots and coefficients of Eq. (2.45).

$$\left.\begin{aligned} \xi_1 + \xi_2 + \xi_3 &= -\frac{A_1}{A_0} \\ \xi_1\xi_2 + \xi_2\xi_3 + \xi_3\xi_1 &= \frac{A_2}{A_0} \\ \text{and } \xi_1\xi_2\xi_3 &= -\frac{A_3}{A_0} \end{aligned}\right\} \tag{2.46}$$

Since $\frac{\xi_2}{\xi_1}, \frac{\xi_3}{\xi_2}$ are negligible, the above relations give

$$\xi_1 = -\frac{A_1}{A_0},\ \xi_2 = -\frac{A_2}{A_1} \text{ and } \xi_3 = -\frac{A_3}{A_2} \tag{2.47}$$

Thus, the magnitudes of the roots will be known when once the roots are widely separated. Now, we shall show the way, the roots are separated by considering the cubic equation

$$p(x) = (x - 1)(x - 2)(x - 3) \tag{2.48}$$

then

$$\begin{aligned} p(-x) &= (-x - 1)(-x - 2)(-x - 3) \\ &= (-1)^3(x + 1)(x + 2)(x + 3) \end{aligned} \tag{2.49}$$

therefore,

$$p(x)p(-x) = (-1)^3(x^2 - 1)(x^2 - 4)(x^2 - 9) \tag{2.50}$$

letting

$$q(z) = (z - 1)(z - 4)(z - 9), \tag{2.51}$$

where $z = x^2$, we find that the roots of Eq. (2.51) are the *squares* of the roots of Eq. (2.48). By transforming Eq. (2.51) in the same way as above, we get another equation whose roots are the squares of the roots of Eq. (2.51). This is the principle underlying this method and due to this reason, this method is called *root-squaring* method.

Now, let the given cubic be

$$f(x) = a_0x_3 + a_1x_2 + a_2x + a_3 = 0 \tag{2.52}$$

with roots α_1, α_2 and α_3 such that

$$|\alpha_1| > |\alpha_2| > |\alpha_3|.$$

Suppose that Eq. (2.52) is transformed 'm' times by the root-squaring process described above, and that the transformed equation is

$$\phi(u) = a_0^{(m)}u^3 + a_1^{(m)}u^2 + a_2^{(m)}u + a_3^{(m)} = 0 \tag{2.53}$$

If u_i are the roots of Eq. (2.53), then we have

$$u_i = \alpha_i^m, \quad i = 1, 2, 3. \tag{2.54}$$

But u_i are given by the coefficients in Eq. (2.53). Hence we have the following formulae for the roots of Eq. (2.52)

$$\alpha_i = \left(\frac{a_i}{a_{i-1}}\right)^{1/m}, \qquad i = 1, 2, 3. \tag{2.55}$$

These results can easily be generalized to a nth degree polynomial.

It is clear that this method gives approximations to the magnitudes of the roots. To find the *sign* of any root, we substitute the root in the original polynomial and find the result. If the result is very nearly zero, then the root is positive; otherwise, it is negative.

The root-squaring process can be terminated when two successive approximations are very nearly the same.

Graeffe's method has the advantage of providing approximations to all the roots of a polynomial equation simultaneously. Once the approximate values to all the roots are known, iterative methods can be used to obtain accurate value of each zero.

Example 2.28 Using Graeffe's method, find the real roots of the equation

$$x^3 - 6x^2 + 11x - 6 = 0$$

Let

$$f(x) = x^3 - 6x^2 + 11x - 6 = 0 \tag{i}$$

then

$$f(-x) = -x^3 - 6x^2 - 11x - 6 = 0$$

therefore,

$$f(x)f(-x) = (-1)^3 (x^6 - 14x^4 + 49x^2 - 36)$$

let

$$\phi(x) = z^3 - 14z^2 + 49z - 36, \quad \text{where } z = x^2.$$

Hence, roots of $f(x) = 0$ are given by

$$\sqrt{\frac{36}{49}} = 0.857, \sqrt{\frac{49}{14}} = 1.871 \text{ and } \sqrt{14} = 3.742$$

Now,

$$\phi(-z) = -z^3 - 14z^2 - 49z - 36$$

Therefore,

$$\phi(z)\ \phi(-z) = (-1)^3(z^6 - 98z^4 + 1393z^2 - 1296)$$

Setting $\phi(u) = u^3 - 98u^2 + 1393u - 1296$, we obtain the next approximation to the roots of $f(x) = 0$ as

$$\left(\frac{1296}{1393}\right)^{1/4} = 0.9822, \left(\frac{1393}{98}\right)^{1/4} = 1.942 \text{ and } (98)^{1/4} = 3.147.$$

It is seen that the approximations are converging to the actual roots 1, 2 and 3, respectively.

Suppose we use $x_0 = 0.857$ and apply Newton's method. We obtain

$$x_1 = 0.857 + \frac{0.35027}{2.919} = 0.977.$$

Similarly, better approximations can be obtained for the other two roots.

2.10 LIN–BAIRSTOW'S METHOD

This method is useful in determining a quadratic factor of a polynomial. We shall explain the mathematical basis of the method by considering a cubic equation, viz.

$$f(x) = a_3x^3 + a_2x^2 + a_1x + a_0 \tag{2.56}$$

Let $x^2 + Rx + S$ be a quadratic factor of $f(x)$ and also let an approximate factor be $x^2 + rx + s$. If we divide $f(x)$ by $x^2 + rx + s$, then both the quotient and remainder would be linear factors. Hence we write

$$f(x) = (x^2 + rx + s)(b_3x + b_2) + b_1x + b_0 \tag{2.57}$$

It is clear that if $r = R$ and $s = S$, i.e., if the quadratic factor is exact, then the remainder term will be zero, which means that $b_1 = b_0 = 0$. Equating the coefficients of like powers of x in Eqs. (2.56) and (2.57), we obtain

$$\left.\begin{aligned} b_3 &= a_3, \\ b_2 &= a_2 - rb_3, \\ b_1 &= a_1 - rb_2 - sb_3, \\ b_0 &= a_0 - sb_2 \end{aligned}\right\} \tag{2.58}$$

From Eq. (2.58), we see that the b_i are functions of both r and s. If the factor $x^2 + rx + s$ is exact, i.e., if $x^2 + rx + s$ divides exactly the function $f(x)$, then we must have

$$b_0 = b_1 = 0$$

i.e.,

$$b_0(r, s) = 0 \quad \text{and} \quad b_1(r, s) = 0 \tag{2.59}$$

Now, by Taylor's theorem,

$$b_0(r, s) = b_0(r_0, s_0) + \frac{\partial b_0}{\partial r}\Delta r_0 + \frac{\partial b_0}{\partial s}\Delta s_0 = 0 \tag{2.60}$$

and

$$b_1(r, s) = b_1(r_0, s_0) + \frac{\partial b_1}{\partial r}\Delta r_0 + \frac{\partial b_1}{\partial s}\Delta s_0 = 0, \tag{2.61}$$

where the higher order partial derivatives are neglected and the first order partial derivatives are calculated at the point (r_0, s_0).

If the initial approximation (r_0, s_0) is assumed, then a correction $(\Delta r_0, \Delta s_0)$ can be computed from Eqs. (2.60) and (2.61), so that we have the next approximation as

These relations are

$$\left.\begin{aligned} r_1 &= r_0 + \Delta r_0 \\ \text{and} \quad s_1 &= s_0 + \Delta s_0 \end{aligned}\right\} \tag{2.62}$$

For the next approximation, we take $r_0 = r_1$ and $s_0 = s_1$, and proceed as above.

Example 2.29 Find a quadratic factor of the polynomial

$$f(x) = x^3 - x - 1.$$

We have

$$a_3 = 1, \quad a_2 = 0, \quad a_1 = -1, \quad a_0 = -1$$

Let

$$r_0 = s_0 = 1.0.$$

Then,

$$\begin{aligned} b_3 &= a_3, \\ b_2 &= a_2 - rb_3 = a_2 - ra_3, \\ b_1 &= a_1 - rb_2 - sb_3 = a_1 - r(a_2 - ra_3) - sa_3 \\ &= a_1 - ra_2 + r^2a_3 - sa_3 \\ b_0 &= a_0 - sb_2 = a_0 - s(a_2 - ra_3) \\ &= a_0 - sa_2 + sra_3 \end{aligned}$$

Therefore,

$$\frac{\partial b_0}{\partial r} = sa_3, \quad \frac{\partial b_0}{\partial s} = -a_2 + ra_3$$

$$\frac{\partial b_1}{\partial r} = -a_2 + 2ra_3, \quad \frac{\partial b_1}{\partial s} = -a_3.$$

Then, Eqs. (2.60) and (2.61) give

$$\Delta r_0 + \Delta s_0 = 0$$

and

$$2\Delta r_0 - \Delta s_0 = 1.$$

Hence,

$$\Delta r_0 = \frac{1}{3} = 0.3333$$

and

$$\Delta s_0 = -\frac{1}{3} = -0.3333.$$

It follows that

$$r_1 = r_0 + \Delta r_0 = 1 + \frac{1}{3} = 1.3333,$$

and

$$s_1 = s_0 + \Delta s_0 = 1 - \frac{1}{3} = 0.6667.$$

For the second approximation, we get

$$r_0 = 1.3333 \quad \text{and} \quad s_0 = 0.6667.$$

Then

$$b_0 = -0.1111 \quad \text{and} \quad b_1 = 0.1110,$$

$$\frac{\partial b_0}{\partial r} = 0.6667, \quad \frac{\partial b_0}{\partial s} = 1.3333, \quad \frac{\partial b_1}{\partial r} = 2.6666, \quad \frac{\partial b_1}{\partial s} = -1.$$

With these values, we obtain $\Delta r_0 = -0.00874$ and $\Delta s_0 = 0.0877$ which give $r_2 = 1.3246$ and $s_2 = 0.7544$, both of which are correct to three decimal places.

2.11 QUOTIENT–DIFFERENCE METHOD

This is a general method to determine the approximate roots of a polynomial equation and is described in Henrici [1964]. It is originally due to Rutishauser [1954]. Only an outline of the method is given here and for details the reader is referred to Henrici's book. We consider the cubic equation

$$f(x) = a_0x^3 + a_1x^2 + a_2x + a_3 = 0, \tag{2.63}$$

whose roots x_1, x_2 and x_3 are such that $0 < |x_1| < |x_2| < |x_3|$.
We write $f(x)$ in the form

$$f(x) = 1 + \frac{a_2}{a_3}x + \frac{a_1}{a_3}x^2 + \frac{a_0}{a_3}x^3$$

Now, let

$$\begin{aligned}\left[1 + \frac{a_2}{a_3}x + \frac{a_1}{a_3}x^2 + \frac{a_0}{a_3}x^3\right]^{-1} &= \sum_{r=1}^{3} \frac{k_r}{x - x_r} \\ &= \sum_{r=1}^{3} \frac{-k_r}{x_r}\left(1 - \frac{x}{x_r}\right)^{-1} \\ &= \sum_{i=0}^{\infty} \sum_{r=1}^{3} \frac{-k_r}{x_r^{i+1}} x^i \\ &= \sum_{i=0}^{\infty} b_i x^i, \end{aligned} \tag{2.64}$$

where

$$b_i = \sum_{r=1}^{3} \frac{-k_r}{x_r^{i+1}} \tag{2.65}$$

The method derives its name from the definitions of *quotients*

$$Q_1^{(i)} = \frac{b_i}{b_{i-1}} \tag{2.66}$$

and *differences*

$$D_1^{(i)} = Q_1^{(i+1)} - Q_1^{(i)} \tag{2.67}$$

In Ramanujan's method (Section 2.6), we have seen that the quantity

$$\lim_{i \to \infty} \frac{b_{i-1}}{b_i} \tag{2.68}$$

tends to the smallest root of the equation $f(x) = 0$, but the Q–D method leads to the approximations of all the roots. With the initial values of $Q_1^{(i)}$ and $D_1^{(i)}$ obtained from the definitions, approximations to the roots are computed by using the formulae:

$$D_r^{(i+1)}\, Q_r^{(i+1)} = D_r^{(i)}\, Q_{r+1}^{(i)} \tag{2.69}$$

and

$$D_r^{(i)} = D_{r-1}^{(i+1)} + Q_r^{(i+1)} - Q_r^{(i)} \tag{2.70}$$

$$\Delta_0^{(i)} = \Delta_3^{(i)} = 0 \text{ for all } i.$$

Proofs of these formulae are left as exercises to the reader.

Using Eqs. (2.69) and (2.70), a table of quotients and differences can be generated and a typical table is given below (Table 2.1).

Table 2.1 A Typical *Q-D* Table

D_0	Q_1	D_1	Q_2	D_2	Q_3	D_3
	$Q_1^{(1)}$		0		0	
0		$D_1^{(1)}$		$D_2^{(0)}$		0
	$Q_1^{(2)}$		$Q_2^{(1)}$		$Q_3^{(0)}$	
0		$D_1^{(2)}$		$D_2^{(1)}$		0
	$Q_1^{(3)}$		$Q_2^{(2)}$		$Q_3^{(1)}$	

When the first two rows are known, the other rows can be constructed by using Eqs. (2.69) and (2.70) alternately. The quantities $Q_1^{(i)}$, $Q_2^{(i)}$ and $Q_3^{(i)}$ tend to the reciprocals of the roots of $f(x) = 0$. Hence, instead of Eq. (2.63), if we consider the transformed equation

$$a_3x^3 + a_2x^2 + a_1x + a_0 = 0 \tag{2.71}$$

and proceed as above, we obtain directly the roots of Eq. (2.63). This procedure is illustrated in the following example.

Example 2.30 Using the *Q–D* method, find the roots of the cubic equation

$$f(x) = x^3 - 9x^2 + 26x - 24 = 0.$$

To compute the roots directly, we consider the transformed equation

$$-24x^3 + 26x^2 - 9x + 1 = 0.$$

We then have

$$(-24x^3 + 26x^2 - 9x + 1)(b_0 + b_1x + b_2x^2 + b_3x^3 + \cdots) = 1.$$

Comparing the coefficients of like powers of x on both sides of the above equation, we obtain

$$b_0 = 1; \qquad -9b_0 + b_1 = 0;$$
$$b_2 - 9b_1 + 26b_0 = 0;$$
$$b_3 - 9b_2 + 26b_1 - 24b_0 = 0$$

The above relations give

$$b_0 = 1, \quad b_1 = 9, \quad b_2 = 55 \quad \text{and} \quad b_3 = 285,$$

so that

$$\frac{b_1}{b_0} = 9 = Q_1^{(1)}$$

$$\frac{b_2}{b_1} = \frac{55}{9} = 6.1111 = Q_1^{(2)}$$

and

$$\frac{b_3}{b_2} = \frac{285}{55} = 5.1818 = Q_1^{(3)}.$$

We obtain the differences

$$D_1^{(1)} = Q_1^{(2)} - Q_1^{(1)} = -2.8889,$$

and

$$D_1^{(2)} = Q_1^{(3)} - Q_1^{(2)} = -0.9293.$$

To determine $Q_2^{(1)}$, we have

$$Q_2^{(1)} = \frac{D_1^{(2)}Q_1^{(2)}}{D_1^{(1)}} = 1.9658.$$

With $Q_2^{(0)} = 0,$ we can now compute $D_2^{(0)}$ from the formula

$$D_2^{(0)} = Q_2^{(1)} + D_1^{(1)} - Q_2^{(0)}$$
$$= -0.9231.$$

It follows immediately that $Q_3^{(0)} = 0.9231$.

We have thus computed all the elements in the first two rows of Table 2.1. This enables us to use Eqs. (2.69) and (2.70) alternately to generate the quotients and differences row by row.

To compute $D_1^{(2)}$, we use the elements

$$\begin{array}{ccc} & D_1^{(1)} & \\ Q_1^{(2)} & & Q_2^{(1)} \\ & D_1^{(2)} & \end{array}$$

where

$$D_1^{(2)} \cdot Q_1^{(2)} = D_1^{(1)} \cdot Q_2^{(1)}$$

Hence

$$D_1^{(2)} = \frac{-0.9293 \times 2.4616}{5.1818} = -0.4415.$$

Similarly,

$$D_2^{(1)} = -\frac{(0.9231)^2}{1.9658} = -0.4355.$$

Next row is a row of quotients. We have $Q_1^{(3)} = 5.1818.$
Also,

$$\begin{aligned} Q_2^{(2)} &= 1.9658 - 0.4335 + 0.9293 \\ &= 2.4616, \end{aligned}$$

and

$$\begin{aligned} Q_3^{(1)} &= 0.9231 - 0.4335 \\ &= 1.3566. \end{aligned}$$

Proceeding in this way, the following table of quotients and differences is formed (Table 2.2).

Table 2.2 Solution of Example 2.30

D_0	Q_1	D_1	Q_2	D_2	Q_3	D_3
	9		0		0	
0		–2.8889		–0.9231		
	6.1111		1.9658		0.9231	
0		–0.9293		–0.4335		0
	5.1818		2.4616		1.3566	
0		–0.4415		–0.2389		0
	4.7403		2.6642		1.5955	
0		–0.2481		–0.1431		0
	4.4922		2.7692		1.7386	
0		–0.1529		–0.0898		0
	4.3393		2.8323		1.8284	
		–0.0998		–0.0580		0
	4.2395		2.8741		1.8864	

It can be seen that Q_1, Q_2 and Q_3 are converging to the actual values 4, 3 and 2, respectively.

2.12 SOLUTION TO SYSTEMS OF NONLINEAR EQUATIONS

In this section, we consider two methods for the solution of simultaneous nonlinear equations: (i) the method of iteration and (ii) Newton–Raphson method. For simplicity, we consider a system of two equations:

$$\left.\begin{aligned} f(x, y) &= 0 \\ \text{and} \quad g(x, y) &= 0 \end{aligned}\right\} \tag{2.72}$$

2.12.1 Method of Iteration

As in the case of a single equation, we assume that Eq. (2.72) may be written in the form

$$x = F(x, y), \quad y = G(x, y), \tag{2.73}$$

where the functions F and G satisfy the following conditions in a closed neighbourhood R of the root (α, β):

(i) F and G and their first partial derivatives are continuous in R, and

(ii) $$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \quad \text{and} \quad \left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| < 1, \tag{2.74}$$

for all (x, y) in R.

If (x_0, y_0) is an initial approximation to the root (α, β), then Eq. (2.73) give the sequence

$$\left.\begin{aligned} x_1 &= F(x_0, y_0), & y_1 &= G(x_0, y_0) \\ x_2 &= F(x_1, y_1), & y_2 &= G(x_1, y_1) \\ \vdots \quad & \quad \vdots & \vdots \quad & \quad \vdots \\ x_{n+1} &= F(x_n, y_n), & y_{n+1} &= G(x_n, y_n) \end{aligned}\right\} \tag{2.75}$$

For faster convergence, recently computed values of x_i may be used in the evaluation of y_i in Eq. (2.75). Conditions in Eq. (2.74) are *sufficient* for convergence and in the limit, we obtain

$$\alpha = F(\alpha, \beta) \quad \text{and} \quad \beta = G(\alpha, \beta) \tag{2.76}$$

Hence, α and β are the roots of Eq. (2.73), and therefore, also of the Eq. (2.72).

The method can obviously be generalized to any number of equations.

Example 2.31 Find a real root of the equations

$$y^2 - 5y + 4 = 0 \quad \text{and}$$

$$3yx^2 - 10x + 7 = 0,$$

using the iteration method.

Clearly, a real root is $x = 1$ and $y = 1$.
To apply the iteration method, we rewrite the equations as

$$x = \frac{1}{10}(3yx^2 + 7) \tag{i}$$

and

$$y = \frac{1}{5}(y^2 + 4) \tag{ii}$$

Here

$$F(x, y) = \frac{1}{10}(3yx^2 + 7), \quad \frac{\partial F}{\partial x} = \frac{6xy}{10}, \quad \frac{\partial F}{\partial y} = \frac{3x^2}{10},$$

$$G(x, y) = \frac{1}{5}(y^2 + 4), \quad \frac{\partial G}{\partial x} = 0, \quad \frac{\partial G}{\partial y} = \frac{2y}{5}.$$

Let (0.5, 0.5) be an approximate root. Then

$$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| = \left|\frac{6xy}{10}\right|_{(0.5, 0.5)} + \left|\frac{3x^2}{10}\right|_{(0.5, 0.5)}$$
$$= 0.15 + 0.075 < 1$$

and

$$\left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| = \left|\frac{2y}{5}\right|_{0.5} = 0.2 < 1.$$

Hence the conditions for convergence are satisfied and the approximations are given by

$$x_{n+1} = \frac{1}{10}\left[3y_n x_n^2 + 7\right] \text{ and } y_{n+1} = \frac{1}{5}\left[y_n^2 + 4\right].$$

We obtain successively with $x_0 = 0.5$ and $y_0 = 0.5$

$$x_1 = \frac{1}{10}\left[\frac{3}{8} + 7\right] = 0.7375; \qquad y_1 = \frac{1}{5}\left[\frac{1}{4} + 4\right] = 0.85$$

$$x_2 = \frac{1}{10}\left[3(0.85)(0.7375)^2 + 7\right] = 0.8387; \qquad y_2 = \frac{1}{5}\left[(0.85)^2 + 4\right] = 0.9445$$

$$x_3 = \frac{1}{10}\left[3(0.8387)^2(0.9445) + 7\right] = 0.8993; \qquad y_3 = \frac{1}{5}\left[(0.9445)^2 + 4\right] = 0.9784.$$

Successive approximations are

$x_4 = 0.9374,$ $\qquad y_4 = 0.9914$

$$x_5 = 0.9613, \qquad y_5 = 0.9966$$
$$x_6 = 0.9763, \qquad y_6 = 0.9986$$
$$x_7 = 0.9855, \qquad y_7 = 0.9994.$$

Convergence to the root (1, 1) is obvious.

2.12.2 Newton–Raphson Method

Let (x_0, y_0) be an initial approximation to the root of Eq. (2.72). If $(x_0 + h, y_0 + k)$ is the root of the system, then we must have

$$f(x_0 + h, y_0 + k) = 0, \qquad g(x_0 + h, y_0 + k) = 0 \tag{2.77}$$

Assuming that f and g are sufficiently differentiable, we expand both the functions in Eq. (2.77) by Taylor's series to obtain

$$\left.\begin{aligned} f_0 + h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} + \cdots = 0 \\ g_0 + h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} + \cdots = 0, \end{aligned}\right\} \tag{2.78}$$

where

$$\frac{\partial f}{\partial x_0} = \left[\frac{\partial f}{\partial x}\right]_{x = x_0}, \; f_0 = f(x_0, y_0), \text{ etc.}$$

Neglecting the second and higher-order derivative terms, we obtain the following system of linear equations:

$$\left.\begin{aligned} h\frac{\partial f}{\partial x_0} + k\frac{\partial f}{\partial y_0} = -f_0 \\ \text{and} \quad h\frac{\partial g}{\partial x_0} + k\frac{\partial g}{\partial y_0} = -g_0 \end{aligned}\right\} \tag{2.79}$$

Equation (2.79) possesses a unique solution if

$$D = \begin{vmatrix} \dfrac{\partial f}{\partial x_0} & \dfrac{\partial f}{\partial y_0} \\ \dfrac{\partial g}{\partial x_0} & \dfrac{\partial g}{\partial y_0} \end{vmatrix} \neq 0. \tag{2.80}$$

By Cramer's rule, the solution of Eq. (2.79) is given by

$$h = \frac{1}{D}\begin{vmatrix} -f_0 & \dfrac{\partial f}{\partial y_0} \\ -g_0 & \dfrac{\partial g}{\partial y_0} \end{vmatrix} \text{ and } k = \frac{1}{D}\begin{vmatrix} \dfrac{\partial f}{\partial x_0} & -f_0 \\ \dfrac{\partial g}{\partial x_0} & -g_0 \end{vmatrix} \tag{2.81}$$

The new approximations are, therefore,

$$x_1 = x_0 + h \quad \text{and} \quad y_1 = y_0 + k \tag{2.82}$$

The process is to be repeated till we obtain the roots to the desired accuracy. As in the case of a single equation, the convergence is of second order.

Example 2.32 Solve the system given in Example 2.31 by Newton–Raphson method.

We have

$$f(x) = 3yx^2 - 10x + 7 = 0$$

$$g(x) = y^2 - 5y + 4 = 0$$

Then,

$$\frac{\partial f}{\partial x} = 6yx - 10, \qquad \frac{\partial f}{\partial y} = 3x^2,$$

$$\frac{\partial g}{\partial x} = 0, \qquad \frac{\partial g}{\partial y} = 2y - 5$$

Taking $x_0 = y_0 = 0.5$, we obtain

$$\frac{\partial f}{\partial x_0} = -8.5, \qquad \frac{\partial f}{\partial y_0} = 0.75, \qquad f_0 = 2.375,$$

$$\frac{\partial g}{\partial x_0} = 0, \qquad \frac{\partial g}{\partial y_0} = -4, \qquad g_0 = 1.75$$

Hence,

$$D = \begin{vmatrix} -8.5 & 0.75 \\ 0 & -4 \end{vmatrix} = 34.$$

Therefore,

$$h = \frac{1}{34}\begin{vmatrix} -2.375 & 0.75 \\ -1.75 & -4 \end{vmatrix} = 0.3180,$$

and

$$k = \frac{1}{34}\begin{vmatrix} -8.5 & -2.375 \\ 0 & -1.75 \end{vmatrix} = 0.4375.$$

It follows that

$$x_1 = 0.5 + 0.3180 = 0.8180$$

and

$$y_1 = 0.5 + 0.4375 = 0.9375$$

For the second approximation, we have

$$f_1 = 0.7019, \qquad g_1 = 0.1914,$$

$$\frac{\partial f}{\partial x_1} = -5.3988, \qquad \frac{\partial f}{\partial y_1} = 2.0074,$$

$$\frac{\partial g}{\partial x_1} = 0, \qquad \frac{\partial g}{\partial y_1} = -3.125.$$

Therefore,

$$D = \begin{vmatrix} -5.3988 & 2.0074 \\ 0 & -3.125 \end{vmatrix} = 16.8712.$$

Hence,

$$h = \frac{1}{16.8712}\begin{vmatrix} -0.7019 & 2.0074 \\ -0.1914 & -3.125 \end{vmatrix} = 0.1528,$$

and

$$k = \frac{1}{16.8712}\begin{vmatrix} -5.3918 & -0.7019 \\ 0 & -0.1914 \end{vmatrix} = 0.0612.$$

It follows that

$$x_2 = 0.8180 + 0.1528 = 0.9708$$

and

$$y_2 = 0.9375 + 0.0612 = 0.9987$$

Example 2.33 Solve the system $x^2 + y^2 = 1$ and $y = x^2$ by Newton–Raphson method.

Let

$$f = x^2 + y^2 - 1 \quad \text{and} \quad g = y - x^2.$$

From the graphs of the curves, we find that there are two points of intersection, one each in the first and second quadrants. We shall approximate to the solution in the first quadrant. We have

$$\frac{\partial f}{\partial x} = 2x, \qquad \frac{\partial f}{\partial y} = 2y,$$

$$\frac{\partial g}{\partial x} = -2x, \qquad \frac{\partial g}{\partial y} = 1.$$

We start with $x_0 = y_0 = 0.7071$ obtained from the approximation $y = x$. Then we compute

$$\frac{\partial f}{\partial x_0} = 1.4142, \qquad \frac{\partial f}{\partial y_0} = 1.4142,$$

$$\frac{\partial g}{\partial x_0} = -1.4142, \qquad \frac{\partial g}{\partial y_0} = 1.$$

Therefore,

$$D = \begin{vmatrix} 1.4142 & 1.4142 \\ -1.4142 & 1 \end{vmatrix} = 3.4142; \quad f_0 = 0,$$

Hence,

$$h = \frac{1}{3.4142}\begin{vmatrix} 0 & 1.4142 \\ -0.2071 & 1 \end{vmatrix} = 0.0858,$$

and

$$k = \frac{1}{3.4142}\begin{vmatrix} 1.4142 & 0 \\ -1.4142 & -0.2071 \end{vmatrix} = -0.0858,$$

It follows, therefore,

$$x_1 = 0.7071 + 0.0858 = 0.7858,$$

and

$$y_1 = 0.7071 - 0.0858 = 0.6213.$$

The process can be repeated to obtain a better approximation.

Eliminating x between the two given equations, we obtain the quadratic for y

$$y^2 + y - 1 = 0,$$

which gives $y = 0.6180$ for the first quadrant solution.

Then $x = 0.7861$. These values may be compared with the solution (x_1, y_1) obtained above.

Example 2.34 Solve the system

$$\sin x - y = -0.9793$$

$$\cos y - x = -0.6703$$

with $x_0 = 0.5$ and $y_0 = 1.5$ as the initial approximation.

We have

$$f(x, y) = \sin x - y + 0.9793$$

and

$$g(x, y) = \cos y - x + 0.6703.$$

For the first iteration, we have

$$f_0 = -0.0413,\ g_0 = 0.2410,$$

$$D = f_x g_y - g_x f_y = \cos(0.5)(-\sin 1.5) - (1) = -1.8754$$

$$h = \frac{fg_y - gf_y}{D} = -0.1505,\ k = \frac{gf_x - fg_x}{D} = -0.0908$$

Therefore,

$$x = 0.5 + 0.1505 = 0.6505$$

and

$$y = 1.5 + 0.0908 = 1.5908$$

For the second iteration, we have $x_0 = 0.6505$ and $y_0 = 1.5908$. Then we obtain

$$D = -1.7956$$

also

$$h = -0.003181 \quad \text{and} \quad k = 0.003384.$$

Hence the new approximation is

$$x = 0.6505 + 0.0032 = 0.6537,$$

$$y = 1.5908 - 0.0034 = 1.5874.$$

Substituting these values in the given equations, we find that these are correct to four decimal places.

EXERCISES

2.1 Explain the bisection method for finding a real root of the equation $f(x) = 0$ and write an algorithm for its implementation with a test for relative accuracy of the approximation.

Obtain a root, correct to three decimal places, of each of the following equations using the bisection method (Problems 2.2–2.5):

2.2 $x^3 - 4x - 9 = 0$

2.3 $x^3 + x^2 - 1 = 0$

2.4 $5x \log_{10} x - 6 = 0$

2.5 $x^2 + x - \cos x = 0$

2.6 Give the sequence of steps in the regula–falsi method for determining a real root of the equation $f(x) = 0$.

Use the method of false position to find a real root, correct to three decimal places, of the following equations (Problems 2.7–2.10):

2.7 $x^3 + x^2 + x + 7 = 0$

2.8 $x^3 - x - 4 = 0$

2.9 $x = 3e^{-x}$

2.10 $x \tan x + 1 = 0$

2.11 Find the real root, which lies between 2 and 3, of the equation

$$x \log_{10} x - 1.2 = 0$$

using the methods of bisection and false–position to a tolerance of 0.5%.

2.12 Explain briefly the method of iteration to compute a real root of the equation $f(x) = 0$, stating the condition of convergence of the sequence of approximations. Give a graphical representation of the method.

Use the method of iteration to find, correct to four significant figures, a real root of each of the following equations (Problems 2.13–2.16):

2.13 $e^x = 3x$

2.14 $x = \dfrac{1}{(x+1)^2}$

2.15 $1 + x^2 = x^3$

2.16 $x - \sin x = \dfrac{1}{2}$

2.17 Establish an iteration formula to find the reciprocal of a positive number N by Newton–Raphson method. Hence find the reciprocal of 154 to four significant figures.

2.18 Explain Newton–Raphson method to compute a real root of the equation $f(x) = 0$ and find the condition of convergence. Hence, find a non-zero root of the equation $x^2 + 4\sin x = 0$.

2.19 Using Newton–Raphson method, derive a formula for finding the kth root of a positive number N and hence compute the value of $(25)^{1/4}$.

Use the Newton–Raphson method to obtain a root, correct to three decimal places, of each of the following equations (Problems 2.20–2.25):

2.20 $x^{\sin 2} - 4 = 0$

2.21 $e^x = 4x$

2.22 $x^3 - 5x + 3 = 0$

2.23 $xe^x = \cos x$

2.24 $x = \dfrac{1 + \cos x}{3}$

2.25 $\cot x = -x$

2.26 Describe a computational procedure to implement Newton–Raphson method for computing the square root of a positive number to an accuracy ε. Write a flow-chart for the same.

2.27 Compute, to four decimal places, the root between 1 and 2 of the equation

$$x^3 - 2x^2 + 3x - 5 = 0$$

by (a) Method of False Position and (b) Newton–Raphson method.

Using Ramanujan's method, find the smallest root of each of the following equations (Problems 2.28–2.30):

2.28 $x^3 - 6x^2 + 11x - 6 = 0$

2.29 $x + x^3 - 1 = 0$

2.30 $\sin x + x - 1 = 0$

2.31 Use the secant method to determine the root, between 5 and 8, of the equation $x^{2.2} = 69$. Compare your result with that obtained in Example 2.7.

2.32 Determine the real root of the equation $x = e^{-x}$, using the secant method. Compare your result with that obtained in Example 2.26.

2.33 Point out the difference between the regula–falsi and the secant methods for finding a real root of $f(x) = 0$. Apply both the methods to find a real root of the equation $x^3 - 2x - 5 = 0$ to an accuracy of 4 decimal places.

2.34 Describe briefly Muller's method and use it to find (a) the root, between 2 and 3, of the equation $x^3 - 2x - 5 = 0$ and (b) the root, between 0 and 1, of the equation $x = e^{-x} \cos x$.

Find the real roots of the following equations using Graeffe's root-squaring method (Problems 2.35–2.36):

2.35 $x^3 - 4x^2 + 5x - 2 = 0$

2.36 $x^3 - 2x^2 - 5x + 6 = 0$

2.37 Find a quadratic factor of the polynomial

$$f(x) = x^3 - 2x^2 + x - 2$$

by Bairstow's method with two approximations starting with $r_0 = -0.5$ and $s_0 = 1$.

2.38 Determine a quadratic factor, nearer to $x^2 - 1.5x + 1.5$, of the polynomial

$$f(x) = x^4 - 8x^3 + 39x^2 - 62x + 50$$

by Bairstow's method. Give the computational steps of this method.

2.39 Using the *Q-D* method, find the real roots of the equation

$$f(x) = x^3 - 6x^2 + 11x - 6 = 0$$

2.40 In the notation of Section 2.11, prove the following results:

(a) $\lim\limits_{i\to\infty} \dfrac{b_i}{b_{i-1}} = \dfrac{1}{x_1}$

(b) $\lim\limits_{i\to\infty} \dfrac{\dfrac{1}{x_1} - Q_1^{(i)}}{\left(\dfrac{x_1}{x_2}\right)^i} = \dfrac{1}{x_1}\dfrac{k_2}{k_1}\left(1 - \dfrac{x_1}{x_2}\right)$

(c) $\lim\limits_{i\to\infty} \dfrac{\dfrac{1}{x_1} - Q_1^{(i+1)}}{\left(\dfrac{x_1}{x_2}\right)^i} = \dfrac{1}{x_2}\dfrac{k_2}{k_1}\left(1 - \dfrac{x_1}{x_2}\right).$

2.41 Prove the formula

$$D_r^{(i)}\, Q_{r+1}^{(i)} = D_r^{(i+1)}\, Q_r^{(i+1)}$$

Solve the following systems by Newton–Raphson method. (Problems 2.42–2.44).

2.42 $x^2 + y^2 = 1, \quad y = x^3$

2.43 $x^2 - y^2 = 4, \quad x^2 + y^2 = 16$

2.44 $x^2 + y = 11, \quad y^2 + x = 7$

2.45 To find the smallest root of the equation $f(x) = x^3 - x - 1 = 0$ by the iteration method, $f(x) = 0$ should be rewritten as

(a) $x = x^3 - 1$ (b) $x = (x + 1)^{1/3}$

(c) $x = \dfrac{1}{x^2 - 1}$ (d) $x = \dfrac{x+1}{x^2}$

Find the correct choice in the above.

2.46 Newton–Raphson formula converges if

(a) $\left| \dfrac{f'(x) f''(x)}{[f(x)]^2} \right| < 1$ (b) $\left| \dfrac{f(x) f''(x)}{[f'(x)]^2} \right| < 1$

(c) $\left| \dfrac{f(x) f'(x)}{[f''(x)]^2} \right| < 1$ (d) None of these.

2.47 Which one of the following is not correct?

(a) Newton–Raphson method has quadratic convergence.
(b) The bisection method converges slowly.
(c) To solve $f(x) = 0$ by iteration method, the given equation is written in the form $x = \phi(x)$ where $|\phi'(x)| < 1$ in an interval containing the root.
(d) The method of regula–falsi converges faster than the secant method.

2.48 Which one of the following is correct?

(a) The bisection method has quadratic convergence.
(b) The iteration method is a self-correction method.
(c) The equation $x^3 - 2x - 5 = 0$ has two positive roots.
(d) The equation $x = \cos x$ can be solved by Graeffe's method.

Answers to Exercises

2.2 2.706 **2.3** 0.755
2.4 2.741 **2.5** 0.550
2.7 –2.105 **2.8** 1.796
2.9 1.0499 **2.10** 2.798

2.11 Successive values are (By bisection method)

2.5, 2.75, 2.625, 2.6875, 2.7187, 2.7344, 2.7422.

By false position method

2.7210, 2.7402, 2.7407, 2.7406, 2.7406

2.13 0.6190 **2.14** 0.4656
2.15 1.466 **2.16** 1.497
2.17 0.006494 **2.18** –1.9338
2.19 2.23607 **2.20** 4.5932

2.21 0.3574

2.22 With $x_0 = 0.5$, $x_1 = 0.6470$, $x_2 = 0.65656$, …

2.23 With $x_0 = 0.5$, $x_1 = 0.5180$, $x_2 = 0.5180$, …

2.24 $x_0 = 1.047$, $x_1 = 0.6224$, $x_2 = 0.6071$, $x_3 = 0.6071$.

2.25 2.798

2.27 1.8437; 1.8438

2.28 $\dfrac{b_1}{b_2} = 0.54546,$ $\dfrac{b_2}{b_3} = 0.77647,$

$\dfrac{b_3}{b_4} = 0.88696,$ $\dfrac{b_4}{b_5} = 0.94237, \ldots$

Smallest root is 1.

2.29 Convergents are:

1.0, 0.5, 0.66666, 0.75, 0.66666, 0.66666, 0.69231, 0.68421, …

2.30 Convergents are:

0.5, 0.5, 0.51064, 0.51087, 0.51097, …

Required root is 0.5110.

2.31 Root = 6.85236.

2.32 Fifth iteration value = 0.567143

2.33 Regula–falsi: 2.0945 in 7 iterations.
Secant method: 2.0946.

2.34 (a) 2.09462409 (b) 0.51752

2.35 Third approximation to the roots: 0.9168, 1.0897, 2.0019

2.36 Third approximation: 3.014443, 1.991425, 0.999494.

2.37 $x^2 - 0.0165x + 0.9394$ (Exact factor is $x^2 + 1$).

2.38 $x^2 - 1.9485x + 1.942982$ (Exact factor is $x^2 - 2x + 2$).

2.39 $Q_1 \approx 3.13145$, $Q_2 \approx 1.898$, $Q_3 \approx 0.9706$.

2.42 $x_1 = 0.8261$, $y_1 = 0.5636$.

2.43 $x_0 = 3$, $y_0 = 2$, $x_1 = 3.1667$, $y_1 = 2.5$

2.44 $x = 3.584$, $y = -1.848$

2.45 (b)

2.46 (b)

2.47 (d)

2.48 (a)

Chapter 3

Interpolation

3.1 INTRODUCTION

The statement

$$y = f(x), \qquad x_0 \le x \le x_n$$

means: corresponding to every value of x in the range $x_0 \le x \le x_n$, there exists one or more values of y. Assuming that $f(x)$ is single-valued and continuous and that it is known explicitly, then the values of $f(x)$ corresponding to certain given values of x, say $x_0, x_1, \ldots, x_n$ can easily be computed and tabulated. The central problem of numerical analysis is the converse one: Given the set of tabular values $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ satisfying the relation $y = f(x)$ where the explicit nature of $f(x)$ is not known, it is required to find a simpler function, say $\phi(x)$, such that $f(x)$ and $\phi(x)$ *agree at the set of tabulated points*. Such a process is called *interpolation*. If $\phi(x)$ is a polynomial, then the process is called *polynomial interpolation* and $\phi(x)$ is called the *interpolating polynomial*. Similarly, different types of interpolation arise depending on whether $\phi(x)$ is a finite trigonometric series, series of Bessel functions, etc. In this chapter, we shall be concerned with polynomial interpolation only. As a justification for the approximation of an unknown function by means of a polynomial, we state here, without proof, a famous theorem due to Weierstrass (1885): if $f(x)$ is continuous in $x_0 \le x \le x_n$, then given any $\varepsilon > 0$, there exists a polynomial $P(x)$ such that

$$|f(x) - P(x)| < \varepsilon, \quad \text{for all } x \text{ in } (x_0, x_n).$$

This means that it is possible to find a polynomial $P(x)$ whose graph remains within the region bounded by $y = f(x) - \varepsilon$ and $y = f(x) + \varepsilon$ for all x between x_0 and x_n, however small ε may be.

When approximating a given function $f(x)$ by means of polynomial $\phi(x)$, one may be tempted to ask: (i) How should the closeness of the approximation be measured? and (ii) What is the criterion to decide the best polynomial approximation to the function? Answers to these questions, important though they are for the practical problem of interpolation, are outside the scope of this book and will not be attempted here. We will, however, derive in the next section a formula for finding the error associated with the approximation of a tabulated function by means of a polynomial.

3.2 ERRORS IN POLYNOMIAL INTERPOLATION

Let the function $y(x)$, defined by the $(n + 1)$ points (x_i, y_i), $i = 0, 1, 2, \ldots, n$, be continuous and differentiable $(n + 1)$ times, and let $y(x)$ be approximated by a polynomial $\phi_n(x)$ of degree not exceeding n such that

$$\phi_n(x_i) = y_i, \quad i = 0, 1, 2, \ldots, n \tag{3.1}$$

If we now use $\phi_n(x)$ to obtain approximate values of $y(x)$ at some points other than those defined by Eq. (3.1), what would be the accuracy of this approximation? Since the expression $y(x) - \phi_n(x)$ vanishes for $x = x_0, x_1, \ldots, x_n$, we put

$$y(x) - \phi_n(x) = L\Pi_{n+1}(x), \tag{3.2}$$

where

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \tag{3.3}$$

and L is to be determined such that Eq. (3.2) holds for any intermediate value of x, say $x = x'$, $x_0 < x' < x_n$. Clearly,

$$L = \frac{y(x') - \phi_n(x')}{\Pi_{n+1}(x')}. \tag{3.4}$$

We construct a function $F(x)$ such that

$$F(x) = y(x) - \phi_n(x) - L\Pi_{n+1}(x), \tag{3.5}$$

where L is given by Eq. (3.4) above,

It is clear that

$$F(x_0) = F(x_1) = \cdots = F(x_n) = F(x') = 0,$$

that is, $F(x)$ vanishes $(n + 2)$ times in the interval $x_0 \le x \le x_n$; consequently, by the repeated application of Rolle's theorem (see Theorem 1.3, Section 1.2), $F'(x)$ must vanish $(n + 1)$ times, $F''(x)$ must vanish n times, etc., in the interval $x_0 \le x \le x_n$. In particular, $F^{(n+1)}(x)$ must vanish once in the interval.

Let this point be given by $x = \xi$, $x_0 < \xi < x_n$. On differentiating Eq. (3.5) $(n + 1)$ times with respect to x and putting $x = \xi$, we obtain

$$0 = y^{(n+1)}(\xi) - L(n + 1)!$$

so that

$$L = \frac{y^{(n+1)}(\xi)}{(n+1)!}. \tag{3.6}$$

Comparison of Eqs. (3.4) and (3.6) yields the results

$$y(x') - \phi_n(x') = \frac{y^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x').$$

Dropping the prime on x', we obtain

$$y(x) - \phi_n(x) = \frac{\Pi_{n+1}(x)}{(n+1)!}y^{(n+1)}(\xi), \quad x_0 < \xi < x_n, \tag{3.7}$$

which is the required expression for the error. Since $y(x)$ is, generally, unknown and hence we do not have any information concerning $y^{(n+1)}(x)$, formula (3.7) is almost useless in practical computations. On the other hand, it is extremely useful in theoretical work in different branches of numerical analysis. In particular, we will use it to determine errors in Newton's interpolating formulae which will be discussed in Section 3.6.

3.3 FINITE DIFFERENCES

Assume that we have a table of values (x_i, y_i), $i = 0, 1, 2, \ldots, n$ of any function $y = f(x)$, the values of x being equally spaced, i.e., $x_i = x_0 + ih$, $i = 0, 1, 2, \ldots, n$. Suppose that we are required to recover the values of $f(x)$ for some intermediate values of x, or to obtain the derivative of $f(x)$ for some x in the range $x_0 \le x \le x_n$. The methods for the solution to these problems are based on the concept of the 'differences' of a function which we now proceed to define.

3.3.1 Forward Differences

If $y_0, y_1, y_2, \ldots, y_n$ denote a set of values of y, then $y_1 - y_0$, $y_2 - y_1$, …, $y_n - y_{n-1}$ are called the *differences* of y. Denoting these differences by Δy_0, Δy_1, …, Δy_{n-1} respectively, we have

$$\Delta y_0 = y_1 - y_0, \qquad \Delta y_1 = y_2 - y_1, \ldots, \qquad \Delta y_{n-1} = y_n - y_{n-1},$$

where Δ is called the *forward difference operator* and Δy_0, Δy_1, …, are called *first forward differences*. The differences of the first forward differences are called *second forward differences* and are denoted by $\Delta^2 y_0$, $\Delta^2 y_1$, … Similarly, one can define *third forward differences*, *fourth forward differences*, etc.

Thus,

$$\begin{aligned}
\Delta^2 y_0 = \Delta y_1 - \Delta y_0 &= y_2 - y_1 - (y_1 - y_0) \\
&= y_2 - 2y_1 + y_0, \\
\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 &= y_3 - 2y_2 + y_1 - (y_2 - 2y_1 + y_0) \\
&= y_3 - 3y_2 + 3y_1 - y_0 \\
\Delta^4 y_0 = \Delta^3 y_1 - \Delta^3 y_0 &= y_4 - 3y_3 + 3y_2 - y_1 - (y_3 - 3y_2 + 3y_1 - y_0) \\
&= y_4 - 4y_3 + 6y_2 - 4y_1 + y_0.
\end{aligned}$$

It is, therefore, clear that any higher-order difference can easily be expressed in terms of the ordinates, since the coefficients occurring on the right side are the binomial coefficients.

Table 3.1 shows how the forward differences of all orders can be formed:

Table 3.1 Forward Difference Table

x	y_0	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
x_0	y_0						
		Δy_0					
x_1	y_1		$\Delta^2 y_0$				
		Δy_1		$\Delta^3 y_0$			
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$		
		Δy_2		$\Delta^3 y_1$		$\Delta^5 y_0$	
x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$		$\Delta^6 y_0$
		Δy_3		$\Delta^3 y_2$		$\Delta^5 y_1$	
x_4	y_4		$\Delta^2 y_3$		$\Delta^4 y_2$		
		Δy_4		$\Delta^3 y_3$			
x_5	y_5		$\Delta^2 y_4$				
		Δy_5					
x_6	y_6						

In practical computations, the forward difference table can be formed in the following way. For the data points (x_i, y_i), $i = 0, 1, 2, \ldots, n$ and $x_i = x_0 + ih$, we have

$$\Delta y_j = y_{j+1} - y_i, j = 0, 1, \ldots, n - 1.$$

Denoting y_j as DEL(0, j), the above equation can be written as

$$\Delta y_j = \text{DEL}(0, j + 1) - \text{DEL}(0, j) = \text{DEL}(1, j)$$

It follows that

$$\Delta^i y_j = \text{DEL}(i - 1, j + 1) - \text{DEL}(i - 1, j),$$

which is the ith forward difference of y_j.

For the data points (x_i, y_i), $i = 0, 1, 2, \ldots, 6$, we have difference Table 3.2.

Table 3.2 Forward Difference Table

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
x_0	DEL(0, 0)						
		DEL(1, 0)					
x_1	DEL(0, 1)		DEL(2, 0)				
		DEL(1, 1)		DEL(3, 0)			
x_2	DEL(0, 2)		DEL(2, 1)		DEL(4, 0)		
		DEL(1, 2)		DEL(3, 1)		DEL(5, 0)	
x_3	DEL(0, 3)		DEL(2, 2)		DEL(4, 1)		DEL(6, 0)
		DEL(1, 3)		DEL(3, 2)		DEL(5, 1)	
x_4	DEL(0, 4)		DEL(2, 3)		DEL(4, 2)		
		DEL(1, 4)		DEL(3, 3)			
x_5	DEL(0, 5)		DEL(2, 4)				
		DEL(1, 5)					
x_6	DEL(0, 6)						

In Table 3.2

$$\begin{aligned}
\text{DEL}(4,\ 0) &= \text{DEL}(3,\ 1) - \text{DEL}(3,\ 0)\\
&= \text{DEL}(2,\ 2) - \text{DEL}(2,\ 1) - [\text{DEL}(2,\ 1) - \text{DEL}(2,\ 0)]\\
&= \text{DEL}(1,\ 3) - \text{DEL}(1,\ 2) - 2[\text{DEL}(1,\ 2) - \text{DEL}(1,\ 1)]\\
&\quad + \text{DEL}(1,\ 1) - \text{DEL}(1,\ 0)\\
&= \text{DEL}(0,\ 4) - \text{DEL}(0,\ 3) - 3[\text{DEL}(0,\ 3) - \text{DEL}(0,\ 2)]\\
&\quad +3[\text{DEL}(0,\ 2) - \text{DEL}(0,\ 1)] - [\text{DEL}(0,\ 1) - \text{DEL}(0,\ 0)]\\
&= \text{DEL}(0,\ 4) - 4\text{DEL}(0,\ 3) + 6\text{DEL}(0,\ 2) - 4\text{DEL}(0,\ 1) + \text{DEL}(0,\ 0)\\
&= y_4 - 4y_3 + 6y_2 - 4y_1 + y_0
\end{aligned}$$

The forward difference table can now be formed by the simple statements:

```
Do i = 1(1)n
Do j = 0(1)n - i
DEL(i, j) = DEL(i - 1, j + 1) - DEL(i - 1, j)
Next j
Next i
End
```

3.3.2 Backward Differences

The differences $y_1 - y_0$, $y_2 - y_1$, …, $y_n - y_{n-1}$ are called first *backward differences* if they are denoted by ∇y_1, ∇y_2, …, ∇y_n respectively, so that

$$\begin{aligned}
\nabla y_1 &= y_1 - y_0, \quad \nabla y_2 = y_2 - y_1,\\
&\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots\\
\nabla y_n &= y_n - y_{n-1},
\end{aligned}$$

where ∇ is called the *backward difference operator.* In a similar way, one can define backward differences of higher orders.

Thus, we obtain

$$\begin{aligned}\nabla^2 y_2 &= \nabla y_2 - \nabla y_1 \\ &= y_2 - y_1 - (y_1 - y_0) = y_2 - 2y_1 + y_0, \\ \nabla^3 y_3 &= \nabla^2 y_3 - \nabla^2 y_2 \\ &= y_3 - 3y_2 + 3y_1 - y_0, \text{ etc.}\end{aligned}$$

With the same values of x and y as in Table 3.1, a backward difference Table 3.3 can be formed:

Table 3.3 Backward Difference Table

x	y	∇	∇^2	∇^3	∇^4	∇^5	∇^6
x_0	y_0						
x_1	y_1	∇y_1					
x_2	y_2	∇y_2	$\nabla^2 y_2$				
x_3	y_3	∇y_3	$\nabla^2 y_3$	$\nabla^3 y_3$			
x_4	y_4	∇y_4	$\nabla^2 y_4$	$\nabla^3 y_4$	$\nabla^4 y_4$		
x_5	y_5	∇y_5	$\nabla^2 y_5$	$\nabla^3 y_5$	$\nabla^4 y_5$	$\nabla^5 y_5$	
x_6	y_6	∇y_6	$\nabla^2 y_6$	$\nabla^3 y_6$	$\nabla^4 y_6$	$\nabla^5 y_6$	$\nabla^6 y_6$

3.3.3 Central Differences

The *central difference operator* δ is defined by the relations

$$y_1 - y_0 = \delta y_{1/2}, \quad y_2 - y_1 = \delta y_{3/2}, \ldots, \quad y_n - y_{n-1} = \delta y_{n-1/2}.$$

Similarly, higher-order central differences can be defined. With the values of x and y as in the preceding two tables, a central difference Table 3.4 can be formed:

Table 3.4 Central Difference Table

x	y	δ	δ^2	δ^3	δ^4	δ^5	δ^6
x_0	y_0						
		$\delta y_{1/2}$					
x_1	y_1		$\delta^2 y_1$				
		$\delta y_{3/2}$		$\delta^3 y_{3/2}$			
x_2	y_2		$\delta^2 y_2$		$\delta^4 y_2$		
		$\delta y_{5/2}$		$\delta^3 y_{5/2}$		$\delta^5 y_{5/2}$	
x_3	y_3		$\delta^2 y_3$		$\delta^4 y_3$		$\delta^6 y_3$
		$\delta y_{7/2}$		$\delta^3 y_{7/2}$		$\delta^5 y_{7/2}$	
x_4	y_4		$\delta^2 y_4$		$\delta^4 y_4$		
		$\delta y_{9/2}$		$\delta^3 y_{9/2}$			
x_5	y_5		$\delta^2 y_5$				
		$\delta y_{11/2}$					
x_6	y_6						

It is clear from all the four tables that in a definite numerical case, the same numbers occur in the same positions whether we use forward, backward or central differences. Thus, we obtain

$$\Delta y_0 = \nabla y_1 = \delta\, y_{1/2}, \qquad \Delta^3 y_2 = \nabla^3 y_5 = \delta^3 y_{7/2}, \ldots$$

3.3.4 Symbolic Relations and Separation of Symbols

Difference formulae can easily be established by symbolic methods, using the *shift* operator E and the *averaging* or the *mean* operator μ, in addition to the operators, Δ, ∇ and δ already defined.

The averaging operator μ is defined by the equation:

$$\mu y_r = \frac{1}{2}(y_{r+1/2} + y_{r-1/2}).$$

The shift operator E is defined by the equation:

$$Ey_r = y_{r+1},$$

which shows that the effect of E is to shift the functional value y_r to the next higher value y_{r+1}. A second equation with E gives

$$E^2 y_r = E(Ey_r) = Ey_{r+1} = y_{r+2},$$

and in general,

$$E^n y_r = y_{r+n}.$$

It is now easy to derive a relationship between Δ and E, for we have

$$\Delta y_0 = y_1 - y_0 = Ey_0 - y_0 = (E-1)\, y_0$$

and hence

$$\Delta \equiv E - 1 \quad \text{or} \quad E \equiv 1 + \Delta. \tag{3.8a)*}$$

We can now express any higher-order forward difference in terms of the given function values. For example,

$$\Delta^3 y_0 = (E-1)^3 y_0 = (E^3 - 3E^2 + 3E - 1)\, y_0 = y_3 - 3y_2 + 3y_1 - y_0.$$

From the definitions, the following relations can easily be established:

$$\left.\begin{aligned} &\nabla = 1 - E^{-1}, \\ &\delta = E^{1/2} - E^{-1/2}, \\ &\mu = (1/2)\,(E^{1/2} + E^{-1/2}),\ \mu^2 = 1 + (1/4)\,\delta^2 \\ &\Delta = \nabla E = \delta E^{1/2}. \end{aligned}\right\} \tag{3.8b}$$

*The student should note that Eq. (3.8a) does not mean that the operators E and Δ have any existence as separate entities; it merely implies that the effect of the operator E on y_0 is the same as that of the operator $(1 + \Delta)$ on y_0.

As an example, we prove the relation

$$\mu^2 \equiv 1 + (1/4)\,\delta^2.$$

We have, by definition,

$$\begin{aligned}\mu y_r &= \frac{1}{2}(y_{r+1/2} + y_{r-1/2})\\ &= \frac{1}{2}(E^{1/2} y_r + E^{-1/2} y_r)\\ &= \frac{1}{2}(E^{1/2} + E^{-1/2}) y_r.\end{aligned}$$

Hence

$$\mu = \frac{1}{2}(E^{1/2} + E^{-1/2})$$

and

$$\begin{aligned}\mu^2 &= \frac{1}{4}(E^{1/2} + E^{-1/2})^2\\ &= \frac{1}{4}(E + E^{-1} + 2)\\ &= \frac{1}{4}[(E^{1/2} - E^{-1/2})^2 + 4]\\ &= \frac{1}{4}(\delta^2 + 4).\end{aligned}$$

We therefore have

$$\mu \equiv \sqrt{1 + \frac{1}{4}\delta^2}.$$

Finally, we define the operator D such that

$$Dy(x) = \frac{d}{dx}\, y(x).$$

To relate D to E, we start with the Taylor's series

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!} y''(x) + \frac{h^3}{3!} y'''(x) + \cdots$$

This can be written in the symbolic form

$$Ey(x) = \left(1 + hD + \frac{h^2 D^2}{2!} + \frac{h^3 D^3}{3!} + \cdots\right) y(x).$$

Since the series in the brackets is the expansion of e^{hD}, we obtain the interesting result

$$E \equiv e^{hD}. \tag{3.8c}$$

Using the relation (3.8a), a number of useful identities can be derived. This relation is used to separate the effect of E into that of the powers of Δ and this method of separation is called the *method of separation of symbols*. The following examples demonstrate the use of this method.

Example 3.1 Using the method of separation of symbols, show that

$$\Delta^n u_{x-n} = u_x - nu_{x-1} + \frac{n(n-1)}{2} u_{x-2} + \cdots + (-1)^n u_{x-n}.$$

To prove this result, we start with the right-hand side. Thus,

$$\begin{aligned}
&u_x - nu_{x-1} + \frac{n(n-1)}{2} u_{x-2} + \cdots + (-1)^n u_{x-n} \\
&= u_x - nE^{-1}u_x + \frac{n(n-1)}{2} E^{-2}u_x + \cdots + (-1)^n E^{-n}u_x \\
&= \left[1 - nE^{-1} + \frac{n(n-1)}{2} E^{-2} + \cdots + (-1)^n E^{-n}\right] u_x \\
&= (1 - E^{-1})^n u_x \\
&= \left(1 - \frac{1}{E}\right)^n u_x \\
&= \left(\frac{E-1}{E}\right)^n u_x \\
&= \frac{\Delta^n}{E^n} u_x \\
&= \Delta^n E^{-n} u_x \\
&= \Delta^n u_{x-n},
\end{aligned}$$

which is the left-hand side.

Example 3.2 Show that

$$e^x\left(u_0 + x\Delta u_0 + \frac{x^2}{2!}\Delta^2 u_0 + \cdots\right) = u_0 + u_1 x + u_2 \frac{x^2}{2!} + \cdots$$

Now,

$$\begin{aligned}
e^x\left(u_0 + x\,\Delta u_0 + \frac{x^2}{2!}\Delta^2 u_0 + \cdots\right) &= e^x\left(1 + x\Delta + \frac{x^2\Delta^2}{2!} + \cdots\right)u_0 \\
&= e^x e^{x\Delta} u_0 = e^{x(1+\Delta)} u_0 \\
&= e^{xE} u_0
\end{aligned}$$

$$= \left(1 + xE + \frac{x^2E^2}{2!} + \cdots\right)u_0$$

$$= u_0 + xu_1 + \frac{x^2}{2!}u_2 + \cdots,$$

which is the required result.

3.4 DETECTION OF ERRORS BY USE OF DIFFERENCE TABLES

Difference tables can be used to check errors in tabular values. Suppose that there is an error of +1 unit in a certain tabular value. As higher differences are formed, the error spreads out fanwise, and is at the same time, considerably magnified, as shown in Table 3.5.

Table 3.5 Detection of Errors using Difference Table

y	Δ	Δ^2	Δ^3	Δ^4	Δ^5
0					
	0				
0		0			
	0		0		
0		0		0	
	0		0		1
0		0		1	
	0		1		–5
0		1		–4	
	1		–3		10
1		–2		6	
	–1		3		–10
0		1		–4	
	0		–1		5
0		0		–1	
	0		0		–1
0		0		0	
	0		0		
0		0			
	0				
0					

This table shows the following characteristics:

(i) The effect of the error increases with the order of the differences.

(ii) The errors in any one column are the binomial coefficients with alternating signs.

(iii) The algebraic sum of the errors in any difference column is zero, and

(iv) The maximum error occurs opposite the function value containing the error. These facts can be used to detect errors by difference tables. We illustrate this by means of an example.

Example 3.3 Consider the following difference table:

x	y	Δ	Δ^2	Δ^3	Δ^4
1	3010				
		414			
2	3424		–36		
		378		–39	
3	3802		–75		+178
		303		+139	
4	4105		+ 64		–271
		367		–132	
5	4472		– 68		+181
		299		+ 49	
6	4771		–19		– 46
		280		+3	
7	5051		–16		
		264			
8	5315				

The term –271 in the fourth difference column has fluctuations of 449 and 452 on either side of it. Comparison with Table 3.5 suggests that there is an error of –45 in the entry for $x = 4$. The correct value of y is therefore 4105 + 45 = 4150, which shows that the last-two digits have been transposed, a very common form of error. The reader is advised to form a new difference table with this correction, and to check that the third differences are now practically constant.

If an error is present in a given data, the differences of some order will become alternating in sign. Hence, higher-order differences should be formed till the error is revealed as in the above example. If there are errors in several tabular values, then it is not easy to detect the errors by differencing.

3.5 DIFFERENCES OF A POLYNOMIAL

Let $y(x)$ be a polynomial of the nth degree so that

$$y(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n.$$

Then we obtain

$$y(x+h) - y(x) = a_0[(x+h)^n - x^n] + a_1[(x+h)^{n-1} - x^{n-1}] + \cdots$$

$$= a_0 (nh)\, x^{n-1} + a_1' x^{n-2} + \cdots + a_n',$$

where $a_1', a_2', \ldots, a_n'$ are the new coefficients.

The above equation can be written as

$$\Delta y(x) = a_0 (nh)\, x^{n-1} + a_1' x^{n-2} + \cdots + a_n',$$

which shows that the first difference of a polynomial of the nth degree is a polynomial of degree $(n-1)$. Similarly, the second difference will be a polynomial of degree $(n-2)$, and the coefficient of x^{n-2} will be $a_0 n(n-1)h^2$.

Thus the nth difference is $a_0 n! h^n$, which is a constant. Hence, the $(n + 1)$th, and higher differences of a polynomial of nth degree will be zero. Conversely, if the nth differences of a tabulated function are constant and the $(n + 1)$th, $(n + 2)$th, …, differences all vanish, then the tabulated function represents a polynomial of degree n. It should be noted that these results hold good only if the values of x are equally spaced. The converse is important in numerical analysis since it enables us to approximate a function by a polynomial if its differences of some order become nearly constant.

3.6 NEWTON'S FORMULAE FOR INTERPOLATION

Given the set of $(n + 1)$ values, viz., (x_0, y_0), (x_1, y_1), (x_2, y_2), …, (x_n, y_n), of x and y, it is required to find $y_n(x)$, a polynomial of the nth degree such that y and $y_n(x)$ agree at the tabulated points. Let the values of x be equidistant, i.e. let

$$x_i = x_0 + ih, \quad i = 0, 1, 2, \ldots, n.$$

Since $y_n(x)$ is a polynomial of the nth degree, it may be written as

$$\left.\begin{aligned} y_n(x) = a_0 &+ a_1(x - x_0) + a_2(x - x_0)(x - x_1) \\ &+ a_3(x - x_0)(x - x_1)(x - x_2) + \cdots \\ &+ a_n(x - x_0)(x - x_1)(x - x_2)\ldots(x - x_{n-1}). \end{aligned}\right\} \tag{3.9}$$

Imposing now the condition that y and $y_n(x)$ should agree at the set of tabulated points, we obtain

$$a_0 = y_0;\ a_1 = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y_0}{h};\ a_2 = \frac{\Delta^2 y_0}{h^2 2!};\ a_3 = \frac{\Delta^3 y_0}{h^3 3!}; \cdots;\ a_n = \frac{\Delta^n y_0}{h^n n!};$$

Setting $x = x_0 + ph$ and substituting for $a_0, a_1, \ldots, a_n$, Eq. (3.9) gives

$$y_n(x) = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 + \cdots$$

$$+ \frac{p(p-1)(p-2)\ldots(p-n+1)}{n!}\Delta^n y_0, \tag{3.10}$$

which is *Newton's forward difference interpolation formula* and is useful for interpolation *near the beginning* of a set of tabular values.

To find the error committed in replacing the function $y(x)$ by means of the polynomial $y_n(x)$, we use Eq. (3.7) to obtain

$$y(x) - y_n(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_n)}{(n+1)!} y^{(n+1)}(\xi), \qquad x_0 < \xi < x_n \tag{3.11}$$

As remarked earlier we do not have any information concerning $y^{(n+1)}(x)$, and therefore, formula given in Eq. (3.11) is useless in practice. Nevertheless,

if $y^{(n+1)}(x)$ does not vary too rapidly in the interval, a useful estimate of the derivative can be obtained in the following way. Expanding $y(x+h)$ by Taylor's series (see Theorem 1.4), we obtain

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!}y''(x) + \cdots$$

Neglecting the terms containing h^2 and higher powers of h, this gives

$$y'(x) \approx \frac{1}{h}[y(x+h) - y(x)] = \frac{1}{h}\Delta y(x).$$

Writing $y'(x)$ as $Dy(x)$ where $D \equiv d/dx$, the differentiation operator, the above equation gives the operator relation

$$D \equiv \frac{1}{h}\Delta \quad \text{and so} \quad D^{n+1} \equiv \frac{1}{h^{n+1}}\Delta^{n+1}.$$

We thus obtain

$$y^{(n+1)}(x) \approx \frac{1}{h^{n+1}}\Delta^{n+1}y(x). \tag{3.12}$$

Equation (3.11) can, therefore, be written as

$$y(x) - y_n(x) = \frac{p(p-1)(p-2)\ldots(p-n)}{(n+1)!}\Delta^{n+1}y(\xi) \tag{3.13}$$

in which form it is suitable for computation.

Instead of assuming $y_n(x)$ as in Eq. (3.9), if we choose it in the form

$$\begin{aligned} y_n(x) = a_0 &+ a_1(x-x_n) + a_2(x-x_n)(x-x_{n-1}) \\ &+ a_3(x-x_n)(x-x_{n-1})(x-x_{n-2}) + \cdots \\ &+ a_n(x-x_n)(x-x_{n-1})\ldots(x-x_1). \end{aligned}$$

and then impose the condition that y and $y_n(x)$ should agree at the tabulated points x_n, x_{n-1}, …, x_2, x_1, x_0, we obtain (after some simplification)

$$y_n(x) = y_n + p\nabla y_n + \frac{p(p+1)}{2!}\nabla^2 y_n + \cdots + \frac{p(p+1)\ldots(p+n-1)}{n!}\nabla^n y_n, \tag{3.14}$$

where $p = (x - x_n)/h$.

This is *Newton's backward difference interpolation formula* and it uses tabular values to the left of y_n. This formula is therefore useful for interpolation *near the end of* the tabular values.

It can be shown that the error in this formula may be written as

$$y(x) - y_n(x) = \frac{p(p+1)(p+2)\ldots(p+n)}{(n+1)!}\nabla^{n+1}y(\xi), \tag{3.15}$$

where $x_0 < x < x_n$ and $x = x_n + ph$.

The following examples illustrate the use of these formulae.

Example 3.4 Find the cubic polynomial which takes the following values: $y(1) = 24$, $y(3) = 120$, $y(5) = 336$, and $y(7) = 720$. Hence, or otherwise, obtain the value of $y(8)$.

We form the difference table:

x	y	Δ	Δ^2	Δ^3
1	24			
		96		
3	120		120	
		216		48
5	336		168	
		384		
7	720			

Here $h = 2$. With $x_0 = 1$, we have $x = 1 + 2p$ or $p = (x - 1)/2$. Substituting this value of p in Eq. (3.10), we obtain

$$y(x) = 24 + \frac{x-1}{2}(96) + \frac{\left(\frac{x-1}{2}\right)\left(\frac{x-1}{2}-1\right)}{2}(120) + \frac{\left(\frac{x-1}{2}\right)\left(\frac{x-1}{2}-1\right)\left(\frac{x-1}{2}-2\right)}{6}(48)$$

$$= x^3 + 6x^2 + 11x + 6.$$

To determine $y(8)$, we observe that $p = 7/2$. Hence, Eq. (3.10) gives:

$$y(8) = 24 + \frac{7}{2}(96) + \frac{(7/2)(7/2-1)}{2}(120) + \frac{(7/2)(7/2-1)(7/2-2)}{6}(48) = 990.$$

Direct substitution in $y(x)$ also yields the same value.

Note: This process of finding the value of y for some value of x *outside* the given range is called extrapolation and this example demonstrates the fact that if a tabulated function is a polynomial, then both interpolation and extrapolation would give exact values.

Example 3.5 Using Newton's forward difference formula, find the sum

$$S_n = 1^3 + 2^3 + 3^3 + \cdots + n^3.$$

We have

$$S_{n+1} = 1^3 + 2^3 + 3^3 + \cdots + n^3 + (n+1)^3$$

Hence

$$S_{n+1} - S_n = (n+1)^3,$$

or

$$\Delta S_n = (n+1)^3. \tag{i}$$

It follows that

$$\Delta^2 S_n = \Delta S_{n+1} - \Delta S_n = (n+2)^3 - (n+1)^3 = 3n^2 + 9n + 7,$$

$$\Delta^3 S_n = 3(n+1)^2 + 9n + 7 - (3n^2 + 9n + 7) = 6n + 12$$

$$\Delta^4 S_n = 6(n+1) + 12 - (6n + 12) = 6.$$

Since $\Delta^5 S_n = \Delta^6 S_n = \cdots = 0$, S_n is a fourth-degree polynomial in n.

Further,

$$S_1 = 1, \qquad \Delta S_1 = 8, \qquad \Delta^2 S_1 = 19, \qquad \Delta^3 S_1 = 18, \qquad \Delta^4 S_1 = 6.$$

Equation (3.10) gives

$$\begin{aligned} S_n &= 1 + (n-1)(8) + \frac{(n-1)(n-2)}{2}(19) + \frac{(n-1)(n-2)(n-3)}{6}(18) \\ &\quad + \frac{(n-1)(n-2)(n-3)(n-4)}{24}(6) \\ &= \frac{1}{4}n^4 + \frac{1}{2}n^3 + \frac{1}{4}n^2 \\ &= \left[\frac{n(n+1)}{2}\right]^2. \end{aligned}$$

Example 3.6 Values of x (in degrees) and sin x are given in the following table:

x (in degrees)	sin x
15	0.2588190
20	0.3420201
25	0.4226183
30	0.5
35	0.5735764
40	0.6427876

Determine the value of sin 38°.

The difference table is

x	sin x	Δ	Δ^2	Δ^3	Δ^4	Δ^5
15	0.2588190					
		0.0832011				
20	0.3420201		–0.0026029			
		0.0805982		–0.0006136		
25	0.4226183		–0.0032165		0.0000248	
		0.0773817		–0.0005888		0.0000041
30	0.5		–0.0038053		0.0000289	
		0.0735764		–0.0005599		
35	0.5735764		–0.0043652			
		0.0692112				
40	0.6427876					

To find sin 38°, we use Newton's backward difference formula with $x_n = 40$ and $x = 38$. This gives

$$p = \frac{x - x_n}{h} = \frac{38 - 40}{5} = -\frac{2}{5} = -0.4.$$

Hence, using Eq. (3.14), we obtain

$$\begin{aligned} y(38) &= 0.6427876 - 0.4\,(0.0692112) + \frac{-0.4(-0.4+1)}{2}(-0.0043652) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)}{6}(-0.0005599) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)(-0.4+3)}{24}(0.0000289) \\ &\quad + \frac{(-0.4)(-0.4+1)(-0.4+2)(-0.4+3)(-0.4+4)}{120}(0.0000041) \\ &= 0.6427876 - 0.02768448 + 0.00052382 + 0.00003583 - 0.00000120 \\ &= 0.6156614. \end{aligned}$$

Example 3.7 Find the missing term in the following table:

x	y
0	1
1	3
2	9
3	–
4	81

Explain why the result differs from $3^3 = 27$.

Since four points are given, the given data can be approximated by a third degree polynomial in x. Hence $\Delta^4 y_0 = 0$. Substituting $\Delta = E - 1$ and simplifying, we get

$$E^4 y_0 - 4E^3 y_0 + 6E^2 y_0 - 4Ey_0 + y_0 = 0.$$

Since $E^r y_0 = y_r$, the above equation becomes

$$y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0.$$

Substituting for y_0, y_1, y_2 and y_4 in the above, we obtain

$$y_3 = 31.$$

The tabulated function is 3^x and the exact value of $y(3)$ is 27. The error is due to the fact that the exponential function 3^x is approximated by means of a polynomial in x of degree 3.

Example 3.8 The table below gives the values of tan x for $0.10 \le x \le 0.30$:

x	$y = \tan x$
0.10	0.1003
0.15	0.1511
0.20	0.2027
0.25	0.2553
0.30	0.3093

Find : (a) tan 0.12 (b) tan 0.26, (c) tan 0.40 and (d) tan 0.50.

The table of difference is

x	y	Δ	Δ^2	Δ^3	Δ^4
0.10	0.1003				
		0.0508			
0.15	0.1511		0.0008		
		0.0516		0.0002	
0.20	0.2027		0.0010		0.0002
		0.0526		0.0004	
0.25	0.2553		0.0014		
		0.0540			
0.30	0.3093				

(a) To find tan (0.12), we have $0.12 = 0.10 + p(0.05)$, which gives $p = 0.4$. Hence, Eq. (3.10) gives

$$\begin{aligned}\tan(0.12) &= 0.1003 + 0.4(0.0508) + \frac{0.4(0.4-1)}{2}(0.0008) \\ &\quad + \frac{0.4(0.4-1)(0.4-2)}{6}(0.0002) \\ &\quad + \frac{0.4(0.4-1)(0.4-2)(0.4-3)}{24}(0.0002) \\ &= 0.1205.\end{aligned}$$

(b) To find tan (0.26), we have 0.26 = 0.30 + p(0.05), which gives p = –0.8. Hence, Eq. (3.14) gives

$$\tan(0.26) = 0.3093 - 0.8(0.0540) + \frac{-0.8(-0.8+1)}{2}(0.0014)$$
$$+ \frac{-0.8(-0.8+1)(-0.8+2)}{6}(0.0004)$$
$$+ \frac{-0.8(-0.8+1)(-0.8+2)(-0.8+3)}{24}(0.0002)$$
$$= 0.2662.$$

Proceeding as in the case (i) above, we obtain

(c) tan (0.40) = 0.4241, and

(d) tan (0.50) = 0.5543.

The actual values, correct to four decimal places, of tan (0.12), tan (0.26), tan (0.40) and tan (0.50) are respectively 0.1206, 0.2660, 0.4228 and 0.5463. Comparison of the computed and actual values shows that in the first-two cases (i.e. of interpolation) the results obtained are fairly accurate whereas in the last-two cases (i.e. of extrapolation) the errors are quite considerable. The example therefore demonstrates the important result that if a tabulated function is other than a polynomial, then extrapolation very far from the table limits would be dangerous—although interpolation can be carried out very accurately.

3.7 CENTRAL DIFFERENCE INTERPOLATION FORMULAE

In the preceding section, we derived and discussed Newton's forward and backward interpolation formulae, which are applicable for interpolation near the beginning and end respectively, of tabulated values. We shall, in the present section, discuss the central difference formulae which are most suited for interpolation near the middle of a tabulated set. The central difference operator δ was already introduced in Section 3.3.3.

The most important central difference formulae are those due to Stirling, Bessel and Everett. These will be discussed in Sections 3.7.2, 3.7.3 and 3.7.4, respectively. Gauss's formulae, introduced in Section 3.7.1 below, are of interest from a theoretical stand-point only.

3.7.1 Gauss' Central Difference Formulae

In this section, we will discuss Gauss' forward and backward formulae.

Gauss' forward formula

We consider the following difference table in which the central ordinate is taken for convenience as y_0 corresponding to $x = x_0$.

The differences used in this formula lie on the line shown in Table 3.6. The formula is, therefore, of the form

$$y_p = y_0 + G_1\Delta y_0 + G_2\Delta^2 y_{-1} + G_3\Delta^3 y_{-1} + G_4\Delta^4 y_{-2} + \cdots \qquad (3.16)$$

where G_1, G_2, … have to be determined. The y_p on the left side can be expressed in terms of y_0, Δy_0 and higher-order differences of y_0, as follows:

Table 3.6 Gauss' Forward Formula

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
x_{-3}	y_{-3}						
		Δy_{-3}					
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$				
		Δy_{-2}		$\Delta^3 y_{-3}$			
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$		
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		
		Δy_1		$\Delta^3 y_0$			
x_2	y_2		$\Delta^2 y_1$				
		Δy_2					
x_3	y_3						

Clearly,

$$y_p = E^p y_0$$

$$= (1+\Delta)^p y_0, \text{ using relation (3.8a)}$$

$$= y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 + \cdots$$

Similarly, the right side of Eq. (3.16) can also be expressed in terms of y_0, Δy_0 and higher-order differences. We have

$$\Delta^2 y_{-1} = \Delta^2 E^{-1} y_0$$

$$= \Delta^2 (1+\Delta)^{-1} y_0$$

$$= \Delta^2 (1 - \Delta + \Delta^2 - \Delta^3 + \cdots) y_0$$

$$= \Delta^2 (y_0 - \Delta y_0 + \Delta^2 y_0 - \Delta^3 y_0 + \cdots)$$

$$= \Delta^2 y_0 - \Delta^3 y_0 + \Delta^4 y_0 - \Delta^5 y_0 + \cdots$$

$$\Delta^3 y_{-1} = \Delta^3 y_0 - \Delta^4 y_0 + \Delta^5 y_0 - \Delta^6 y_0 + \cdots$$

$$\begin{aligned}\Delta^4 y_{-2} &= \Delta^4 E^{-2} y_0 \\ &= \Delta^4 (1+\Delta)^{-2} y_0 \\ &= \Delta^4 (y_0 - 2\Delta y_0 + 3\Delta^2 y_0 - 4\Delta^3 y_0 + \cdots) \\ &= \Delta^4 y_0 - 2\Delta^5 y_0 + 3\Delta^6 y_0 - 4\Delta^7 y_0 + \cdots\end{aligned}$$

Hence Eq. (3.16) gives the identity

$$\begin{aligned}&y_0 + p\Delta y_0 + \frac{p(p-1)}{2!}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{3!}\Delta^3 y_0 \\ &\quad + \frac{p(p-1)(p-2)(p-3)}{4!}\Delta^4 y_0 + \cdots \\ &= y_0 + G_1 \Delta y_0 + G_2(\Delta^2 y_0 - \Delta^3 y_0 + \Delta^4 y_0 - \Delta^5 y_0 + \cdots) \\ &\quad + G_3(\Delta^3 y_0 - \Delta^4 y_0 + \Delta^5 y_0 - \Delta^6 y_0 + \cdots) \\ &\quad + G_4(\Delta^4 y_0 - 2\Delta^5 y_0 + 3\Delta^6 y_0 - 4\Delta^7 y_0 + \cdots) + \cdots \qquad (3.17)\end{aligned}$$

Equating the coefficients of Δy_0, $\Delta^2 y_0$, $\Delta^3 y_0$, etc., on both sides of Eq. (3.17), we obtain

$$\left.\begin{aligned} G_1 &= p, \\ G_2 &= \frac{p(p-1)}{2!}, \\ G_3 &= \frac{(p+1)\,p(p-1)}{3!}, \\ G_4 &= \frac{(p+1)\,p(p-1)(p-2)}{4!}. \end{aligned}\right\} \qquad (3.18)$$

Gauss' backward formula

This formula uses the differences which lie on the line shown in Table 3.7.

Table 3.7 Gauss' Backward Formula

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
$\vdots$	$\vdots$						
x_{-1}	y_{-1}						
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$	
x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		Δy_0		$\Delta^3 y_{-1}$		$\Delta^5 y_{-2}$	
x_1	y_1						
$\vdots$	$\vdots$						

Gauss' backward formula can therefore be assumed to be of the form

$$y_p = y_0 + G_1'\Delta y_{-1} + G_2'\Delta^2 y_{-1} + G_3'\Delta^3 y_{-2} + G_4'\Delta^4 y_{-2} + \cdots \tag{3.19}$$

where $G_1', G_2', \ldots$ have to be determined. Following the same procedure as in Gauss' forward formula, we obtain

$$\left.\begin{aligned} G_1' &= p, \\ G_2' &= \frac{p(p+1)}{2!}, \\ G_3' &= \frac{(p+1)\,p(p-1)}{3!} \\ G_4' &= \frac{(p+2)\,(p+1)\,p(p-1)}{4!} \\ &\vdots \end{aligned}\right\} \tag{3.20}$$

Example 3.9 From the following table, find the value of $e^{1.17}$ using Gauss' forward formula:

x	e^x
1.00	2.7183
1.05	2.8577
1.10	3.0042
1.15	3.1582
1.20	3.3201
1.25	3.4903
1.30	3.6693

We have

$$1.17 = 1.15 + p(0.05),$$

which gives

$$p = \frac{0.02}{0.05} = \frac{1}{4}.$$

The difference table is given below.

x	e^x	Δ	Δ^2	Δ^3	Δ^4
1.00	2.7183				
		0.1394			
1.05	2.8577		0.0071		
		0.1465		0.0004	
1.10	3.0042		0.0075		0
		0.1540		0.0004	
1.15	3.1582		0.0079		0
		0.1619		0.0004	
1.20	3.3201		0.0083		0.0001
		0.1702		0.0005	
1.25	3.4903		0.0088		
		0.1790			
1.30	3.6693				

Using formulae (3.16) and (3.18), we obtain

$$\begin{aligned} e^{1.17} &= 3.1582 + \frac{2}{5}(0.1619) + \frac{(2/5)(2/5-1)}{2}(0.0079) \\ &\quad + \frac{(2/5+1)\,(2/5)\,(2/5-1)}{6}(0.0004) \\ &= 3.1582 + 0.0648 - 0.0009 \\ &= 3.2221. \end{aligned}$$

3.7.2 Stirling's Formula

Taking the mean of Gauss' forward and backward formulae, we obtain

$$y_p = y_0 + p\frac{\Delta y_{-1} + \Delta y_0}{2} + \frac{p^2}{2}\Delta^2 y_{-1} + \frac{p(p^2-1)}{3!}\,\frac{\Delta^3 y_{-1} + \Delta^3 y_{-2}}{2}$$

$$+ \frac{p^2(p^2-1)}{4!}\Delta^4 y_{-2} + \cdots \tag{3.21}$$

Formula given in Eq. (3.21) is called *Stirling's formula*.

3.7.3 Bessel's Formula

This is a very useful formula for practical interpolation, and it uses the differences as shown in the following table, where the brackets mean that the average of the values has to be taken.

$\vdots$	$\vdots$						
x_{-1}	y_{-1}						
x_0	$\begin{pmatrix} y_0 \\ y_1 \end{pmatrix}$	Δy_0	$\begin{pmatrix} \Delta^2 y_{-1} \\ \Delta^2 y_0 \end{pmatrix}$	$\Delta^3 y_{-1}$	$\begin{pmatrix} \Delta^4 y_{-2} \\ \Delta^4 y_{-1} \end{pmatrix}$	$\Delta^5 y_{-2}$	$\begin{pmatrix} \Delta^6 y_{-3} \\ \Delta^6 y_{-2} \end{pmatrix}$
x_1							
$\vdots$	$\vdots$						

Hence, Bessel's formula can be assumed in the form

$$y_p = \frac{y_0 + y_1}{2} + B_1 \Delta y_0 + B_2 \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + B_3 \Delta^3 y_{-1}$$

$$+ B_4 \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots$$

$$= y_0 + \left(B_1 + \frac{1}{2} \right) \Delta y_0 + B_2 \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + B_3 \Delta^3 y_{-1}$$

$$+ B_4 \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots \tag{3.22}$$

Using the method outlined in Section 3.7.1, i.e., Gauss' forward formula, we obtain

$$\left.\begin{aligned} B_1 + \frac{1}{2} &= p, \\ B_2 &= \frac{p(p-1)}{2!}, \\ B_3 &= \frac{p(p-1)(p-1/2)}{3!}, \\ B_4 &= \frac{(p+1)\, p(p-1)(p-1)}{4!}, \\ &\vdots \end{aligned}\right\} \tag{3.23}$$

Hence, Bessel's interpolation formula may be written as

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!} \Delta^3 y_{-1}$$

$$+ \frac{(p+1)\, p(p-1)(p-2)}{4!} \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots \tag{3.24}$$

3.7.4 Everett's Formula

This is an extensively used interpolation formula and uses only even order differences, as shown in the following table:

x_0	y_0		$\Delta^2 y_{-1}$		$\Delta^4 y_{-2}$		$\Delta^6 y_{-3}$
		–		–		–	
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		$\Delta^6 y_{-2}$

Hence the formula has the form

$$y_p = E_0 y_0 + E_2 \Delta^2 y_{-1} + E_4 \Delta^4 y_{-2} + \cdots + F_0 y_1 + F_2 \Delta^2 y_0 + F_4 \Delta^4 y_{-1} + \cdots \quad (3.25)$$

The coefficients $E_0, F_0, E_2, F_2, E_4, F_4, \ldots$ can be determined by the same method as in the preceding cases, and we obtain

$$\left.\begin{aligned} E_0 &= 1 - p = q, & F_0 &= p, \\ E_2 &= \frac{q(q^2 - 1^2)}{3!}, & F_2 &= \frac{p(p^2 - 1^2)}{3!}, \\ E_4 &= \frac{q(q^2 - 1^2)(q^2 - 2^2)}{5!}, & F_4 &= \frac{p(p^2 - 1^2)(p^2 - 2^2)}{5!}, \\ &\vdots & &\vdots \end{aligned}\right\} \quad (3.26)$$

Hence Everett's formula is given by

$$\left.\begin{aligned} y_p = q y_0 &+ \frac{q(q^2 - 1^2)}{3!} \Delta^2 y_{-1} + \frac{q(q^2 - 1^2)(q^2 - 2^2)}{5!} \Delta^4 y_{-2} + \cdots \\ &+ p y_1 + \frac{p(p^2 - 1^2)}{3!} \Delta^2 y_0 + \frac{p(p^2 - 1^2)(p^2 - 2^2)}{5!} \Delta^4 y_{-1} + \cdots \end{aligned}\right\} \quad (3.27)$$

where $q = 1 - p$.

3.7.5 Relation between Bessel's and Everett's Formulae

These formulae are very closely related, and it is possible to deduce one from the other by a suitable rearrangement. To see this we start with Bessel's formula

$$y_p = y_0 + p\Delta y_0 + \frac{p(p-1)}{2!} \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!} \Delta^3 y_{-1}$$

$$+ \frac{(p+1)\, p(p-1)(p-2)}{4!} \frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots$$

$$= y_0 + p(y_1 - y_0) + \frac{p(p-1)}{2!}\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{p(p-1)(p-1/2)}{3!}(\Delta^2 y_0 - \Delta^2 y_{-1})$$

$$+ \frac{(p+1)\,p(p-1)(p-2)}{4!}\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \cdots$$

expressing the odd order differences in terms of low even order differences. This gives on simplification

$$y_p = (1-p)y_0 + \left[\frac{p(p-1)}{4} - \frac{(p-1)\,p(p-1/2)}{6}\right]\Delta^2 y_{-1} + \cdots$$

$$+ py_1 + \left[\frac{p(p-1)}{4} + \frac{p(p-1)(p-1/2)}{6}\right]\Delta^2 y_0 + \cdots$$

$$= qy_0 + \frac{q(q^2-1^2)}{3!}\Delta^2 y_{-1} + \cdots + py_1 + \frac{p(p^2-1^2)}{3!}\Delta^2 y_0 + \cdots$$

which is *Everett's formula* truncated after second differences. Hence we have a result of practical importance that Everett's formula truncated after second differences is equivalent to Bessel's formula truncated after third differences. In a similar way, Bessel's formula may be deduced from Everett's.

3.8 PRACTICAL INTERPOLATION

In the preceding sections, we have derived some interpolation formulae of great practical importance. A natural question is: Which one of these formulae gives the most accurate result?

(i) For interpolation at the beginning or end of a table of values, Newton's forward and backward interpolation formulae have to be used respectively.

(ii) For interpolation near the middle of a set of values, the following are the choices:

$$\text{Stirling's formula if } -\frac{1}{4} \le p \le \frac{1}{4},$$

and

$$\text{Bessel's formula for } \frac{1}{4} \le p \le \frac{3}{4}.$$

It can be shown that if the third differences are negligible, then Bessel's formula is about seven times more accurate than Stirling's formula. If the third differences are more than 60 in magnitude, then Everett's formula should be preferred.

Example 3.10 The following table gives the values of e^x for certain equidistant values of x. Find the value of e^x when $x = 0.644$.

x	$y = e^x$
0.61	1.840431
0.62	1.858928
0.63	1.877610
0.64	1.896481
0.65	1.915541
0.66	1.934792
0.67	1.954237

The table of differences is

x	$y = e^x$	Δ	Δ^2	Δ^3	Δ^4
0.61	1.840431				
		0.018497			
0.62	1.858928		0.000185		
		0.018682		0.000004	
0.63	1.877610		0.000189		–0.000004
		0.018871		0	
0.64	1.896481		0.000189		0.000002
		0.019060		0.000002	
0.65	1.915541		0.000191		0.000001
		0.019251		0.000003	
0.66	1.934792		0.000194		
		0.019445			
0.67	1.954237				

Clearly,

$$p = \frac{0.644 - 0.64}{0.01} = 0.4.$$

The third difference contribution to both Stirling's and Bessel's formulae is negligible, and using Stirling's formula, we obtain

$$y(0.644) = 1.896481 + 0.4\frac{0.018871 + 0.019060}{2} + \frac{0.16}{2}(0.000189)$$

$$= 1.896481 + 0.0075862 + 0.00001512$$

$$= 1.904082,$$

while Bessel's formula gives

$$y(0.644) = 1.896481 + 0.4(0.019060) + \frac{0.4(0.4-1)}{2} \cdot \frac{0.000189 + 0.000191}{2}$$

$$= 1.896481 + 0.0076240 - 0.0000228$$

$$= 1.904082.$$

Using Everett's formula, we find that

$$y(0.644) = 0.6(1.896481) + \frac{0.6(0.36-1)}{2}(0.000189)$$
$$+ 0.4(1.915541) + \frac{0.4(0.16-1)}{2}(0.000191)$$
$$= 1.1378886 - 0.000012096 + 0.7662164 - 0.000010696$$
$$= 1.904082.$$

In all the above cases, the value obtained is correct to six decimal places.

It is known from algebra that the nth degree polynomial which passes through $(n+1)$ points is *unique*. Hence the various interpolation formulae derived here are actually only different forms of the same polynomial. It, therefore, follows that all the interpolation formulae should give the same functional value. This is illustrated in the above example where we found that the interpolated value of 0.644 is 1.904082 regardless of which formula is used.

Example 3.11 From the table of Example 3.10, find the value of e^x when $x = 0.638$, using Stirling's and Bessel's formulae.

It was mentioned in Section 3.8 that Stirling's formula gives the most accurate result for $-1/4 \le p \le 1/4$, and Bessel's formula is most efficient for $1/4 \le p \le 3/4$. In order to use these formulae, we therefore, have to choose x_0 so that p satisfies the appropriate inequality.

To use Stirling's formula, we choose $x_0 = 0.64$ and $x_n = 0.638$ so that $p = -0.2$. Hence,

$$y(0.638) = 1.896481 - 0.2 \cdot \frac{0.018871 + 0.019060}{2} + \frac{0.04}{2}(0.000189)$$
$$= 1.896481 - 0.0037931 + 0.0000038$$
$$= 1.892692,$$

which is correct to the last decimal place.

For Bessel's formula, we choose $x_0 = 0.63$, $x_n = 0.638$ so that $p = 0.8$. Hence, we obtain

$$y(0.638) = 1.877610 + 0.8(0.018871) + \frac{0.8(0.8-1)}{2}(0.000189)$$
$$= 1.877610 + 0.0150968 - 0.0000151$$
$$= 1.892692, \text{ as before.}$$

Example 3.12 The values of x and e^{-x} are given in the following table. Find the value of e^{-x} when $x = 1.7475$.

x	$y = e^{-x}$	Δ	Δ^2	Δ^3	Δ^4
1.72	0.1790661479				
		–17817379			
1.73	0.1772844100		177285		
		–17640094		–1762	
1.74	0.1755204006		175523		13
		–17464571		–1749	
1.75	0.1737739435		173774		22
		–17290797		–1727	
1.76	0.1720448638		172047		15
		–17118750		–1712	
1.77	0.1703329888		170335		
		–16948415			
1.78	0.1686381473				

It should be noted that in writing the differences in the above table, the zeros between the decimal point and the first significant digit to its right are omitted. Thus, in the column of second differences, the number 173774 should be taken as 0.0000173774 in the computations.

To compute $y(1.7475)$, we choose $x_0 = 1.74$ and $x_p = 1.7475$ so that $p = 3/4$. We shall obtain the solution by using both Bessel's and Everett's formulae.

(i) If we use Bessel's formula, the third differences need to be taken into account since they exceed 60 units in magnitude. Hence Bessel's formula gives

$$y(1.7475) = 0.1755204006 - \frac{3}{4}(0.0017464571) + \frac{(3/4)(3/4-1)}{2}\,\frac{0.0000175523 + 0.0000173774}{2}$$

$$= 0.1755204006 - 0.00130984284 - 0.00000163734 + 0.00000000137$$

$$= 0.1742089218, \text{ correct to ten decimal places.}$$

(ii) On the other hand, if we use Everett's formula up to second differences only, we obtain

$$y(1.7475) = \frac{1}{4}(0.1755204006) + \frac{(1/4)(1/16-1)}{6}(0.0000175523) + \frac{3}{4}(0.1737739435) + \frac{(3/4)(9/16-1)}{6}(0.0000173774)$$

$$= 0.04388010015 - 0.00000068564 + 0.13033045764 - 0.00000095033$$

$$= 0.1742089218, \text{ as before.}$$

This example verifies the result of Section 3.7.5 that Everett's formula turncated after second differences is equivalent to Bessel's formula truncated after third differences. When the fourth difference contribution becomes significant (i.e. when they exceed 20 units in magnitude), Everett's formula will be easier to apply since it uses only the even order differences.

3.9 INTERPOLATION WITH UNEVENLY SPACED POINTS

In the preceding sections, we have derived interpolation formulae of utmost importance and discussed their practical use in some detail. But, as is well known, they possess the disadvantage of requiring the values of the independent variable to be equally spaced. It is therefore desirable to have interpolation formulae with unequally spaced values of the argument. We discuss, in the present section and the next, four such formulae: (i) Lagrange's interpolation formula which uses only the function values, (ii) Hermite's interpolation formula which is similar to Lagrange's formula, (iii) Newton's general interpolation formula which uses what are called divided differences and (iv) Aitken's method of interpolation by iteration.

3.9.1 Lagrange's Interpolation Formula

Let $y(x)$ be continuous and differentiable $(n + 1)$ times in the interval (a, b). Given the $(n + 1)$ points (x_0, y_0), (x_1, y_1), …, (x_n, y_n) where the values of x need not necessarily be equally spaced, we wish to find a polynomial of degree n, say $L_n(x)$, such that

$$L_n(x_i) = y(x_i) = y_i, \quad i = 0, 1, \ldots, n \tag{3.28}$$

Before deriving the general formula, we first consider a simpler case, viz., the equation of a straight line (a linear polynomial) passing through two points (x_0, y_0) and (x_1, y_1). Such a polynomial, say $L_1(x)$, is easily seen to be

$$\begin{aligned} L_1(x) &= \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 \\ &= l_0(x) y_0 + l_1(x) y_1 \\ &= \sum_{i=0}^{1} l_i(x) y_i, \end{aligned} \tag{3.29}$$

where

$$l_0(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{and} \quad l_1(x) = \frac{x - x_0}{x_1 - x_0}. \tag{3.30}$$

From Eq. (3.30), it is seen that

$$l_0(x_0)=1, \qquad l_0(x_1)=0, \qquad l_1(x_0)=0, \qquad l_1(x_1)=1.$$

These relations can be expressed in a more convenient form as

$$l_i(x_j)=\begin{cases} 1, \text{ if } i=j \\ 0, \text{ if } i\neq j. \end{cases} \tag{3.31}$$

The $l_i(x)$ in Eq. (3.29) also have the property

$$\sum_{i=0}^{1} l_i(x)=l_0(x)+l_1(x)=\frac{x-x_1}{x_0-x_1}+\frac{x-x_0}{x_1-x_0}=1. \tag{3.32}$$

Equation (3.29) is the *Lagrange polynomial of degree one passing through* two points (x_0, y_0) and (x_1, y_1). In a similar way, the *Lagrange polynomial of degree two* passing through three points (x_0, y_0), (x_1, y_1) and (x_2, y_2) is written as

$$\begin{aligned} L_2(x) &= \sum_{i=0}^{2} l_i(x)\, y_i \\ &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}y_0+\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1+\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}y_2, \end{aligned} \tag{3.33}$$

where the $l_i(x)$ satisfy the conditions given in Eqs. (3.31) and (3.32).

To derive the general formula, let

$$L_n(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n \tag{3.34}$$

be the desired polynomial of the nth degree such that conditions given in Eq. (3.28) (called the *interpolatory conditions*) are satisfied. Substituting these conditions in Eq. (3.34), we obtain the system of equations

$$\left.\begin{aligned} y_0 &= a_0+a_1x_0+a_2x_0^2+\cdots+a_nx_0^n \\ y_1 &= a_0+a_1x_1+a_2x_1^2+\cdots+a_nx_1^n \\ y_2 &= a_0+a_1x_2+a_2x_2^2+\cdots+a_nx_2^n \\ &\vdots \\ y_n &= a_0+a_1x_n+a_2x_n^2+\cdots+a_nx_n^n. \end{aligned}\right\} \tag{3.35}$$

The set of Eqs. (3.35) will have a solution if

$$\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} \neq 0. \tag{3.36}$$

The value of this determinant, called *Vandermonde's determinant*, is

$$(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)(x_1 - x_2)\ldots(x_1 - x_n)\ldots(x_{n-1} - x_n).$$

Eliminating a_0, a_1, …, a_n from Eqs. (3.34) and (3.35), we obtain

$$\begin{vmatrix} L_n(x) & 1 & x & x^2 & \cdots & x^n \\ y_0 & 1 & x_0 & x_0^2 & \cdots & x_0^n \\ y_1 & 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ y_n & 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix} = 0, \tag{3.37}$$

which shows that $L_n(x)$ is a linear combination of y_0, y_1, y_2, …, y_n. Hence we write

$$L_n(x) = \sum_{i=0}^{n} l_i(x) y_i, \tag{3.38}$$

where $l_i(x)$ are polynomials in x of degree n. Since $L_n(x_j) = y_j$ for $j = 0, 1, 2, \ldots, n$, Eq. (3.32) gives

$$\left.\begin{aligned} l_i(x_j) &= 0 \qquad \text{if} \quad i \neq j \\ l_j(x_j) &= 1 \qquad \text{for all} \quad j \end{aligned}\right\},$$

which are the same as Eq. (3.31). Hence $l_i(x)$ may be written as

$$l_i(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}, \tag{3.39}$$

which obviously satisfies the conditions (3.31).

If we now set

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_{i-1})(x - x_i)(x - x_{i+1})\ldots(x - x_n), \tag{3.40}$$

then

$$\begin{aligned} \Pi'_{n+1}(x_i) &= \frac{d}{dx}[\Pi_{n+1}(x)]_{x = x_i} \\ &= (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n) \end{aligned} \tag{3.41}$$

so that Eq. (3.39) becomes

$$l_i(x) = \frac{\Pi_{n+1}(x)}{(x - x_i)\,\Pi'_{n+1}(x_i)}. \tag{3.42}$$

Hence Eq. (3.38) gives

$$L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x - x_i)\,\Pi'_{n+1}(x_i)} y_i, \tag{3.43}$$

which is called *Lagrange's interpolation formula*. The coefficients $l_i(x)$, defined in Eq. (3.39), are called *Lagrange interpolation coefficients*. Interchanging x and y in Eq. (3.43), we obtain the formula

$$L_n(y) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(y)}{(y - y_i)\,\Pi'_{n+1}(y_i)} x_i, \tag{3.44}$$

which is useful for *inverse interpolation*.

It is trivial to show that the Lagrange interpolating polynomial is *unique*. To prove this, we assume the contrary. Let $\bar{L}_n(x)$ be a polynomial, distinct from $L_n(x)$, of degree not exceeding n and such that

$$\bar{L}_n(x_i) = y_i, \qquad i = 0, 1, 2, \ldots, n.$$

Then the polynomial defined by $M(x)$, where

$$M(x) = L_n(x) - \bar{L}_n(x)$$

vanishes at the $(n + 1)$ points x_i, $i = 0, 1, \ldots, n$. Hence we have

$$M_n(x) \equiv 0,$$

which shows that $L_n(x)$ and $\bar{L}_n(x)$ are identical.

A major advantage of this formula is that the coefficients in Eq. (3.44) are easily determined. Further, it is more general in that it is applicable to either equal or unequal intervals and the abscissae $x_0, x_1, \ldots, x_n$ need not be in order. Using this formula it is, however, inconvenient to pass from one interpolation polynomial to another of degree one greater.

The following examples illustrate the use of Lagrange's formula.

Example 3.13 Certain corresponding values of x and $\log_{10} x$ are (300, 2.4771), (304, 2.4829), (305, 2.4843) and (307, 2.4871). Find $\log_{10}301$.

From formula given in Eq. (3.43), we obtain

$$\begin{aligned}
\log_{10} 301 &= \frac{(-3)\,(-4)\,(-6)}{(-4)\,(-5)\,(-7)}(2.4771) + \frac{(1)\,(-4)\,(-6)}{(4)\,(-1)\,(-3)}(2.4829) \\
&\quad + \frac{(1)\,(-3)\,(-6)}{(5)\,(1)\,(-2)}(2.4843) + \frac{(1)\,(-3)\,(-4)}{(7)\,(3)\,(2)}(2.4871) \\
&= 1.2739 + 4.9658 - 4.4717 + 0.7106 \\
&= 2.4786.
\end{aligned}$$

Example 3.14 If $y_1 = 4$, $y_3 = 12$, $y_4 = 19$ and $y_x = 7$, find x.

Using Eq. (3.44), we have

$$x = \frac{(-5)(-12)}{(-8)(-15)}(1) + \frac{(3)(-12)}{(8)(-7)}(3) + \frac{(3)(-5)}{(15)(7)}(4)$$

$$= \frac{1}{2} + \frac{27}{14} - \frac{4}{7}$$

$$= 1.86.$$

The actual value is 2.0 since the above values were obtained from the polynomial $y(x) = x^2 + 3$.

Example 3.15 Find the Lagrange interpolating polynomial of degree 2 approximating the function $y = \ln x$ defined by the following table of values. Hence determine the value of ln 2.7.

x	$y = \ln x$
2	0.69315
2.5	0.91629
3.0	1.09861

We have

$$l_0(x) = \frac{(x-2.5)(x-3.0)}{(-0.5)(-1.0)} = 2x^2 - 11x + 15.$$

Similarly, we find

$$l_1(x) = -(4x^2 - 20x + 24) \quad \text{and} \quad l_2(x) = 2x^2 - 9x + 10.$$

Hence

$$L_2(x) = (2x^2 - 11x + 15)(0.69315) - (4x^2 - 20x + 24)(0.91629)$$
$$+ (2x^2 - 9x + 10)(1.09861)$$
$$= -0.08164x^2 + 0.81366x - 0.60761,$$

which is the required quadratic polynomial.

Putting $x = 2.7$, in the above polynomial, we obtain

$$\ln 2.7 \approx L_2(2.7) = -0.08164\,(2.7)^2 + 0.81366\,(2.7) - 0.60761 = 0.9941164.$$

Actual value of ln 2.7 = 0.9932518, so that

$$|\text{Error}| = 0.0008646.$$

Example 3.16 The function $y = \sin x$ is tabulated below

x	$y = \sin x$
0	0
$\pi/4$	0.70711
$\pi/2$	1.0

Using Lagrange's interpolation formula, find the value of $\sin(\pi/6)$.

We have

$$\sin\frac{\pi}{6} \approx \frac{(\pi/6-0)(\pi/6-\pi/2)}{(\pi/4-0)(\pi/4-\pi/2)}(0.70711) + \frac{(\pi/6-0)(\pi/6-\pi/4)}{(\pi/2-0)(\pi/2-\pi/4)}(1)$$

$$= \frac{8}{9}(0.70711) - \frac{1}{9}$$

$$= \frac{4.65688}{9}$$

$$= 0.51743.$$

Example 3.17 Using Lagrange's interpolation formula, find the form of the function $y(x)$ from the following table

x	y
0	-12
1	0
3	12
4	24

Since $y = 0$ when $x = 1$, it follows that $x - 1$ is a factor. Let $y(x) = (x-1)\,R(x)$. Then $R(x) = y/(x-1)$. We now tabulate the values of x and $R(x)$.

x	$R(x)$
0	12
3	6
4	8

Applying Lagrange's formula to the above table, we find

$$R(x) = \frac{(x-3)(x-4)}{(-3)(-4)}(12) + \frac{(x-0)(x-4)}{(3-0)(3-4)}(6) + \frac{(x-0)(x-3)}{(4-0)(4-3)}(8)$$

$$= (x-3)(x-4) - 2x(x-4) + 2x(x-3)$$

$$= x^2 - 5x + 12.$$

Hence the required polynomial approximation to $y(x)$ is given by

$$y(x) = (x-1)(x^2 - 5x + 12).$$

3.9.2 Error in Lagrange's Interpolation Formula

Equation (3.7) can be used to estimate the error of the Lagrange interpolation formula for the class of functions which have continuous derivatives of order up to $(n+1)$ on $[a, b]$. We, therefore, have

$$y(x) - L_n(x) = R_n(x) = \frac{\Pi_{n+1}(x)}{(n+1)!} y^{(n+1)}(\xi), \quad a < \xi < b \tag{3.45}$$

and the quantity E_L, where

$$E_L = \max_{[a,b]} |R_n(x)| \tag{3.46}$$

may be taken as an estimate of error. Further, if we assume that

$$|y^{(n+1)}(\xi)| \le M_{n+1}, \quad a \le \xi \le b \tag{3.47}$$

then

$$E_L \le \frac{M_{n+1}}{(n+1)!} \max_{[a,b]} |\Pi_{n+1}(x)| \tag{3.48}$$

The following examples illustrate the computation of the error.

Example 3.18 Estimate the error in the value of y obtained in Example 3.15.

Since $y = \ln x$, we obtain $y' = 1/x$, $y'' = -1/x^2$ and $y''' = 2/x^3$. It follows that $y'''(\xi) = 2/\xi^3$. Thus the continuity conditions on $y(x)$ and its derivatives are satisfied in [2, 3]. Hence

$$R_n(x) = \frac{(x-2)(x-2.5)(x-3)}{6} \frac{2}{\xi^3}, \qquad 2 < \xi < 3$$

But

$$\left| \frac{1}{\xi^3} \right| < \frac{1}{2^3} = \frac{1}{8}.$$

When $x = 2.7$, we therefore obtain

$$|R_n(x)| \le \left| \frac{(2.7-2)(2.7-2.5)(2.7-3)}{6} \frac{2}{8} \right| = \frac{0.7 \times 0.2 \times 0.3}{3 \times 8} = 0.00175,$$

which agrees with the actual error given in Example 3.15.

Example 3.19 Estimate the error in the solution computed in Example 3.16.

Since $y(x) = \sin x$, we have

$$y'(x) = \cos x, \qquad y''(x) = -\sin x, \qquad y'''(x) = -\cos x.$$

Hence $|y'''(\xi)| < 1$.

When $x = \pi/6$.

$$|R_n(x)| \le \left| \frac{(\pi/6 - 0)(\pi/6 - \pi/4)(\pi/6 - \pi/2)}{6} \right| = \frac{1}{6}\frac{\pi}{6}\frac{\pi}{12}\frac{\pi}{3} = 0.02392,$$

which agrees with the actual error in the solution obtained in Example 3.16.

3.9.3 Hermite's Interpolation Formula

The interpolation formulae so far considered make use of only a certain number of function values. We now derive an interpolation formula in which both the function and its first derivative values are to be assigned at each point of interpolation. This is referred to as *Hermite's interpolation formula*. The interpolation problem is then defined as follows: Given the set of data points (x_i, y_i, y_i'), $i = 0, 1, \ldots, n$, it is required to determine a polynomial of the least degree, say $H_{2n+1}(x)$, such that

$$H_{2n+1}(x_i) = y_i \quad \text{and} \quad H'_{2n+1}(x_i) = y_i'; \qquad i = 0, 1, \ldots, n, \tag{3.49}$$

where the primes denote differentiation with respect to x. The polynomial $H_{2n+1}(x)$ is called *Hermite's interpolation polynomial*. We have here $(2n + 2)$ conditions and therefore the number of coefficients to be determined is $(2n + 2)$ and the degree of the polynomial is $(2n + 1)$. In analogy with the Lagrange interpolation formula (3.43), we seek a representation of the form

$$H_{2n+1}(x) = \sum_{i=0}^{n} u_i(x) y_i + \sum_{i=0}^{n} v_i(x) y_i', \tag{3.50}$$

where $u_i(x)$ and $v_i(x)$ are polynomials in x of degree $(2n + 1)$. Using conditions (3.49), we obtain

$$\left.\begin{aligned} u_i(x_j) &= \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \ne j \end{cases}; \qquad v_i(x) = 0, \text{ for all } i \\ u_i'(x) &= 0, \text{ for all } i; \qquad v_i'(x_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \ne j. \end{cases} \end{aligned}\right\} \tag{3.51}$$

Since $u_i(x)$ and $v_i(x)$ are polynomials in x of degree $(2n + 1)$, we write

$$u_i(x) = A_i(x)\,[l_i(x)]^2 \quad \text{and} \quad v_i(x) = B_i(x)\,[l_i(x)]^2, \tag{3.52}$$

where $l_i(x)$ are given by Eq. (3.42). It is easy to see that $A_i(x)$ and $B_i(x)$ are both linear functions in x. We therefore write

$$u_i(x) = (a_i x + b_i)\,[l_i(x)]^2 \quad \text{and} \quad v_i(x) = (c_i x + d_i)\,[l_i(x)]^2 \tag{3.53}$$

Using conditions Eq. (3.51) in Eq. (3.53), we obtain

$$\left.\begin{aligned} a_i x_i + b_i &= 1 \\ c_i x_i + d_i &= 0 \end{aligned}\right\} \tag{3.54a}$$

and

$$\left.\begin{aligned} a_i + 2l_i'(x_i) &= 0 \\ c_i &= 1. \end{aligned}\right\} \tag{3.54b}$$

From Eq. (3.54), we deduce

$$\left.\begin{aligned} a_i &= -2l_i'(x_i), \quad & b_i &= 1 + 2x_i\ l_i'(x_i) \\ c_i &= 1, & d_i &= -x_i. \end{aligned}\right\} \tag{3.55}$$

Hence Eq. (3.53) become

$$u_i(x) = [-2x\, l_i'(x_i) + 1 + 2x_i l_i'(x_i)]\,[l_i(x)]^2$$

$$= [1 - 2\,(x - x_i)\, l_i'(x_i)]\,[l_i(x)]^2 \tag{3.56a}$$

and

$$v_i(x) = (x - x_i)\,[l_i(x)]^2. \tag{3.56b}$$

Using the above expressions for $u_i(x)$ and $v_i(x)$ in Eq. (3.50), we obtain finally

$$H_{2n+1}(x) = \sum_{i=0}^{n} [1 - 2(x - x_i)\, l_i'(x_i)]\,[l_i(x)]^2\, y_i + \sum_{i=0}^{n} (x - x_i)\,[l_i(x)]^2\, y_i', \tag{3.57}$$

which is the required *Hermite interpolation formula*.

The following examples demonstrate the application of Hermite's formula.

Example 3.20 Find the third-order Hermite polynomial passing through the points $(x_i,\ y_i,\ y'_i)$, $i = 0,\ 1$.

Putting $n = 1$ in Hermite's formula (3.57), we obtain

$$H_3(x) = [1 - 2(x - x_0)\, l_0'(x_0)]\,[l_0(x)]^2\, y_0 + [1 - 2\,(x - x_1)\, l_1'(x_1)]\,[l_1(x)]^2\, y_1$$
$$+ (x - x_0)\,[l_0(x)]^2\, y_0' + (x - x_1)\,[l_1(x)]^2\, y_1'. \tag{i}$$

Since

$$l_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x_1 - x}{h_1} \quad \text{and} \quad l_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - x_0}{h_1},$$

where $h_1 = x_1 - x_0$. Hence

$$l_0'(x) = -\frac{1}{h_1} \quad \text{and} \quad l_1'(x) = \frac{1}{h_1}.$$

Then, Eq. (i) simplifies to

$$H_3(x) = \left[1 + \frac{2(x - x_0)}{h_1}\right] \frac{(x_1 - x)^2}{h_1^2} y_0 + \left[1 + \frac{2(x_1 - x)}{h_1}\right] \frac{(x - x_0)^2}{h_1^2} y_1$$

$$+ (x - x_0) \frac{(x_1 - x)^2}{h_1^2} y_0' + (x - x_1) \frac{(x - x_0)^2}{h_1^2} y_1', \qquad \text{(ii)}$$

which is the required Hermite formula.

Example 3.21 Determine the Hermite polynomial of degree 5, which fits the following data and hence find an approximate value of ln 2.7.

x	$y = \ln x$	$y' = 1/x$
2.0	0.69315	0.5
2.5	0.91629	0.4000
3.0	1.09861	0.33333

The polynomials $l_i(x)$ have already been computed in Example 3.15. These are

$$l_0(x) = 2x^2 - 11x + 15, \qquad l_1(x) = -(4x^2 - 20x + 24), \qquad l_2(x) = 2x^2 - 9x + 10.$$

We therefore obtain

$$l_0'(x) = 4x - 11, \qquad l_1'(x) = -8x + 20, \qquad l_2'(x) = 4x - 9.$$

Hence

$$l_0'(x_0) = -3, \qquad l_1'(x_1) = 0, \qquad l_2'(x_2) = 3$$

Equations (3.56) give

$$u_0(x) = (6x - 11)(2x^2 - 11x + 15)^2, \qquad v_0(x) = (x - 2)(2x^2 - 11x + 15)^2$$

$$u_1(x) = (4x^2 - 20x + 24)^2, \qquad v_1(x) = (x - 2.5)(4x^2 - 20x + 24)^2,$$

$$u_2(x) = (19 - 6x)(2x^2 - 9x + 10)^2, \qquad v_2(x) = (x - 3)(2x^2 - 9x + 10)^2,$$

Substituting these expressions in Eq. (3.57), we obtain the required Hermite polynomial

$$\begin{aligned} H_5(x) = {} & (6x - 11)(2x^2 - 11x + 15)^2 (0.69315) \\ & + (4x^2 - 20x + 24)^2 (0.91629) \\ & + (19 - 6x)(2x^2 - 9x + 10)^2 (1.09861) \\ & + (x - 2)(2x^2 - 11x + 15)^2 (0.5) \\ & + (x - 2.5)(4x^2 - 20x + 24)^2 (0.4) \\ & + (x - 3)(2x^2 - 9x + 10)^2 (0.33333). \end{aligned}$$

Putting $x = 2.7$ and simplifying, we obtain

$$\ln(2.7) \approx H_5(2.7) = 0.993252,$$

which is correct to six decimal places. This is therefore a more accurate result than that obtained by using the Lagrange interpolation formula.

3.10 DIVIDED DIFFERENCES AND THEIR PROPERTIES

The Lagrange interpolation formula, derived in Section 3.9.1, has the disadvantage that if another interpolation point were added, then the interpolation coefficients $l_i(x)$ will have to be recomputed. We therefore seek an interpolation polynomial which has the property that a polynomial of higher degree may be derived from it by simply adding new terms. Newton's general interpolation formula is one such formula and it employs what are called *divided differences*. It is our principal purpose in this section to define such differences and discuss certain of their properties to obtain the basic formula due to Newton.

Let $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ be the given $(n+1)$ points. Then the divided differences of order $1, 2, \ldots, n$ are defined by the relations:

$$\left.\begin{aligned}
[x_0, x_1] &= \frac{y_1 - y_0}{x_1 - x_0}, \\
[x_0, x_1, x_2] &= \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0}, \\
&\vdots \\
[x_0, x_1, \ldots, x_n] &= \frac{[x_1, x_2, \ldots, x_n] - [x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.
\end{aligned}\right\} \tag{3.58}$$

Even if the arguments are equal, the divided differences may still have a meaning. We then set $x_1 = x_0 + \varepsilon$ so that

$$\begin{aligned}
[x_0, x_1] &= \lim_{\varepsilon \to 0} [x_0, x_0 + \varepsilon] \\
&= \lim_{\varepsilon \to 0} \frac{y(x_0 + \varepsilon) - y(x_0)}{\varepsilon} \\
&= y'(x_0), \quad \text{if } y(x) \text{ is differentiable.}
\end{aligned}$$

Similarly,

$$\underbrace{[x_0, x_0, \ldots, x_0]}_{(r+1)\text{ arguments}} = \frac{y^r(x_0)}{r!}. \tag{3.59}$$

From Eq. (3.58), it is easy to see that

$$[x_0, x_1] = \frac{y_0}{x_0 - x_1} + \frac{y_1}{x_1 - x_0} = [x_1, x_0].$$

Again,

$$\begin{aligned}
[x_0, x_1, x_2] &= \frac{1}{x_2 - x_0}\left(\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}\right) \\
&= \frac{1}{x_2 - x_0}\left[\frac{y_2}{x_2 - x_1} - y_1\left(\frac{1}{x_2 - x_1} + \frac{1}{x_1 - x_0}\right) + \frac{y_0}{x_1 - x_0}\right] \\
&= \frac{y_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{y_1}{(x_1 - x_0)(x_1 - x_2)} \\
&\quad + \frac{y_2}{(x_2 - x_0)(x_2 - x_1)}. \qquad (3.60)
\end{aligned}$$

Similarly it can be shown that

$$\begin{aligned}
[x_0, x_1, \ldots, x_n] &= \frac{y_0}{(x_0 - x_1)\ldots(x_0 - x_n)} + \frac{y_1}{(x_1 - x_0)\ldots(x_1 - x_n)} + \cdots \\
&\quad + \frac{y_n}{(x_n - x_0)\ldots(x_n - x_{n-1})}. \qquad (3.61)
\end{aligned}$$

Hence the divided differences are symmetrical in their arguments.

Now let the arguments be equally spaced so that $x_1 - x_0 = x_2 - x_1 = \cdots = x_n - x_{n-1} = h$. Then we obtain

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{1}{h}\Delta y_0 \qquad (3.62)$$

$$[x_0, x_1, x_2] = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0} = \frac{1}{2h}\left(\frac{\Delta y_1}{h} - \frac{\Delta y_0}{h}\right) = \frac{1}{2h^2}\Delta^2 y_0 = \frac{1}{h^2 2!}\Delta^2 y_0 \qquad (3.63)$$

and in general,

$$[x_0, x_1, \ldots, x_n] = \frac{1}{h^n n!}\Delta^n y_0. \qquad (3.64)$$

If the tabulated function is a polynomial of nth degree, then $\Delta^n y_0$ would be a constant and hence the nth divided difference would also be a constant.

For the set of values (x_i, y_i), $i = 0, 1, 2, \ldots, n$, divided differences can be generated by the following statements.

```
Define y(x_j) = y_j = DD(0, j), j = 0, 1, 2, …, n
Do i = 1(1)n
Do j = 0(1)(n - i)
DD(i, j) = (DD(i-1, j+1) - DD(i-1, j)) / (X(i+j) - X(j))
Next j
Next i
```

3.10.1 Newton's General Interpolation Formula

By definition, we have

$$[x, x_0] = \frac{y - y_0}{x - x_0},$$

so that

$$y = y_0 + (x - x_0)[x, x_0] \tag{3.65}$$

Again

$$[x, x_0, x_1] = \frac{[x, x_0] - [x_0, x_1]}{x - x_1}$$

which gives

$$[x, x_0] = [x_0, x_1] + (x - x_1)\ [x, x_0, x_1]$$

Substituting this value of $[x, x_0]$ in Eq. (3.65), we obtain

$$y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x, x_0, x_1] \tag{3.66}$$

But

$$[x, x_0, x_1, x_2] = \frac{[x, x_0, x_1] - [x_0, x_1, x_2]}{x - x_2},$$

and so

$$[x, x_0, x_1] = [x_0, x_1, x_2] + (x - x_2)[x, x_0, x_1, x_2] \tag{3.67}$$

Equation (3.66) now gives

$$\begin{aligned} y = y_0 &+ (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &+ (x - x_0)\,(x - x_1)(x - x_2)[x, x_0, x_1, x_2] \end{aligned} \tag{3.68}$$

Proceeding in this way, we obtain

$$\begin{aligned} y = y_0 &+ (x - x_0)\ [x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ &+ (x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] + \cdots \\ &+ (x - x_0)(x - x_1)(x - x_2)\ \cdots\ (x - x_n)[x, x_0, x_1, \ldots, x_n] \end{aligned} \tag{3.69}$$

This formula is called *Newton's general interpolation formula with divided differences,* the last term being the remainder term after $(n + 1)$ terms.

After generating the divided differences, interpolation can be carried out by the following statements.

Let y_k be required corresponding to the value $x = x_k$. Then

```
yk = y0
factor = 1.0
Do i = 0(1)(n - 1)
factor = factor * (xk - xi)
yk = yk + factor * DD (i + 1, 0)
Next i
End
```

Example 3.22 As our first example to illustrate the use of Newton's divided difference formula, we consider the data of Example 3.13.

The divided difference table is

x	$\log_{10} x$		
300	2.4771		
		0.00145	
304	2.4829		0.00001
		0.00140	
305	2.4843		0
		0.00140	
307	2.4871		

Hence Eq. (3.69) gives

$$\log_{10} 301 = 2.4771 + 0.00145 + (-3)\,(-0.00001) = 2.4786, \text{ as before.}$$

It is clear that the arithmetic in this method is much simpler when compared to that in Lagrange's method.

Example 3.23 Using the following table find $f(x)$ as a polynomial in x.

x	$f(x)$
–1	3
0	–6
3	39
6	822
7	1611

The divided difference table is

x	$f(x)$				
–1	3				
		–9			
0	–6		6		
		15		5	
3	39		41		1
		261		13	
6	822		132		
		789			
7	1611				

Hence Eq. (3.69) gives

$$f(x) = 3 + (x+1)(-9) + x(x+1)(6) + x(x+1)(x-3)(5) + x(x+1)(x-3)(x-6)$$
$$= x^4 - 3x^3 + 5x^2 - 6.$$

3.10.2 Interpolation by Iteration

Newton's general interpolation formula may be considered as one of a class of methods which generate successively higher-order interpolation formulae. We now describe another method of this class, due to A.C. Aitken, which has the advantage of being very easily programmed for a digital computer.

Given the $(n + 1)$ points (x_0, y_0), (x_1, y_1), …, (x_n, y_n), where the values of x need not necessarily be equally spaced, then to find the value of y corresponding to any given value of x we proceed iteratively as follows: obtain a first approximation to y by considering the first-two points only; then obtain its second approximation by considering the first-three points, and so on. We denote the different interpolation polynomials by $\Delta(x)$, with suitable subscripts, so that at the first stage of approximation, we have

$$\Delta_{01}(x) = y_0 + (x - x_0)[x_0, x_1] = \frac{1}{x_1 - x_0}\begin{vmatrix} y_0 & x_0 - x \\ y_1 & x_1 - x \end{vmatrix}. \tag{3.70}$$

Similarly, we can form $\Delta_{02}(x)$, $\Delta_{03}(x), \ldots$

Next, we form Δ_{012} by considering the first-three points:

$$\Delta_{012}(x) = \frac{1}{x_2 - x_1}\begin{vmatrix} \Delta_{01}(x) & x_1 - x \\ \Delta_{02}(x) & x_2 - x \end{vmatrix}. \tag{3.71}$$

Similarly, we obtain $\Delta_{013}(x)$, $\Delta_{014}(x)$, etc. At the nth stage of approximation, we obtain

$$\Delta_{012\ldots n}(x) = \frac{1}{x_n - x_{n-1}}\begin{vmatrix} \Delta_{012\ldots\overline{n-1}}(x) & x_{n-1} - x \\ \Delta_{012\ldots\overline{n-2}n}(x) & x_n - x \end{vmatrix}. \tag{3.72}$$

The computations may conveniently be arranged as in Table 3.8 below:

Table 3.8 Aitken's Scheme

x	y				
x_0	y_0				
		$\Delta_{01}(x)$			
x_1	y_1		$\Delta_{012}(x)$		
		$\Delta_{02}(x)$		$\Delta_{0123}(x)$	
x_2	y_2		$\Delta_{013}(x)$		$\Delta_{01234}(x)$
		$\Delta_{03}(x)$		$\Delta_{0124}(x)$	
x_3	y_3		$\Delta_{014}(x)$		
		$\Delta_{04}(x)$			
x_4	y_4				

A modification of this scheme, due to Neville, is given in Table 3.9. Neville's scheme is particularly suited for iterated inverse interpolation.

Table 3.9 Neville's Scheme

x	y				
x_0	y_0				
		$\Delta_{01}(x)$			
x_1	y_1		$\Delta_{012}(x)$		
		$\Delta_{12}(x)$		$\Delta_{0123}(x)$	
x_2	y_2		$\Delta_{123}(x)$		$\Delta_{01234}(x)$
		$\Delta_{23}(x)$		$\Delta_{1234}(x)$	
x_3	y_3		$\Delta_{234}(x)$		
		$\Delta_{34}(x)$			
x_4	y_4				

As an illustration of Aitken's method, we consider, again, Example 3.22.

Example 3.24 Aitken's scheme is

x	$\log_{10} x$			
300	2.4771			
		2.47855		
304	2.4829		2.47858	
		2.47854		2.47860
305	2.4843		2.47857	
		2.47853		
307	2.4871			

Hence $\log_{10} 301 = 2.4786$, as before.

An obvious advantage of Aitken's method is that *it gives a good idea of the accuracy of the result at any stage*.

3.11 INVERSE INTERPOLATION

Given a set of values of x and y, the process of finding the value of x for a certain value of y is called *inverse interpolation*. When the values of x are at unequal intervals, the most obvious way of performing this process is by interchanging x and y in Lagrange's or Aitken's methods. Use of Lagrange's formula was already illustrated in Example 3.14. We will now solve the same example by means of Aitken's and Neville's schemes.

Aitken's scheme (see Table 3.8) is

y	x		
4	1		
		1.750	
12	3		1.857
		1.600	
19	4		

whereas Neville's scheme (see Table 3.9) gives

y	x		
4	1		
		1.750	
12	3		1.857
		2.286	
19	4		

Hence both the schemes lead to the same result ultimately. In practice, however, Neville's scheme should be preferred for the simple reason that in this scheme those points which are nearest to x_r are used for interpolation at $x = x_r$. It is, of course, important to remember that inverse interpolation is, in general, meaningful only *if the function is single-valued in the interval*.

When the values of x are equally spaced, the method of successive approximations, described below, should be used.

Method of successive approximations

We start with Newton's forward difference formula [see Eq. (3.10), Section 3.6] written as

$$y_u = y_0 + u\Delta y_0 + \frac{u(u-1)}{2}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{6}\Delta^3 y_0 + \cdots \tag{3.73}$$

From this, we obtain

$$u = \frac{1}{\Delta y_0}\left[y_u - y_0 - \frac{u(u-1)}{2}\Delta^2 y_0 - \frac{u(u-1)(u-2)}{6}\Delta^3 y_0 - \cdots\right]. \tag{3.74}$$

Neglecting the second and higher differences, we obtain the first approximation to u and this, we write, as follows

$$u_1 = \frac{1}{\Delta y_0}(y_u - y_0). \tag{3.75}$$

Next, we obtain the second approximation to u by including the term containing the second differences. Thus,

$$u_2 = \frac{1}{\Delta y_0}\left[y_u - y_0 - \frac{u_1(u_1-1)}{2}\Delta^2 y_0\right], \tag{3.76}$$

where we have used the value of u_1 for u in the coefficient of $\Delta^2 y_0$. Similarly, we obtain

$$u_3 = \frac{1}{\Delta y_0}\left[y_u - y_0 - \frac{u_2(u_2-1)}{2}\Delta^2 y_0 - \frac{u_2(u_2-1)(u_2-2)}{6}\Delta^3 y_0\right] \tag{3.77}$$

and so on. This process should be continued till two successive approximations to u agree with each other to the required accuracy. The method is illustrated by means of the following example.

Example 3.25 Tabulate $y = x^3$ for x = 2, 3, 4 and 5, and calculate the cube root of 10 correct to *three* decimal places.

x	$y = x^3$	Δ	Δ^2	Δ^3
2	8			
		19		
3	27		18	
		37		6
4	64		24	
		61		
5	125			

Here $y_u = 10,\ y_0 = 8,\ \Delta y_0 = 19,\ \Delta^2 y_0 = 18$ and $\Delta^3 y_0 = 6$. The successive approximations to u are, therefore,

$$u_1 = \frac{1}{19}(2) - 0.1$$

$$u_2 = \frac{1}{19}\left[2 - \frac{0.1\,(0.1-1)}{2}(18)\right] = 0.15$$

$$u_3 = \frac{1}{19}\left[2 - \frac{0.15\,(0.15-1)}{2}(18) - \frac{0.15\,(0.15-1)\,(0.15-2)}{6}(6)\right] = 0.1532$$

$$u_4 = \frac{1}{19}\left[2 - \frac{0.1532\,(0.1532-1)}{2}(18) - \frac{0.1532\,(0.1532-1)\,(0.1532-2)}{6}(6)\right] = 0.1541$$

$$u_5 = \frac{1}{19}\left[2 - \frac{0.1541\,(0.1541-1)}{2}(18) - \frac{0.1541\,(0.1541-1)\,(0.1541-2)}{6}(6)\right] = 0.1542.$$

We, therefore, take u = 0.154 correct to three decimal places. Hence the value of x (which corresponds to y = 10), i.e. the cube root of 10 is given by $x_0 + uh$ = 2.154.

This example demonstrates the relationship between the inverse interpolation and the solution of algebraic equations.

3.12 DOUBLE INTERPOLATION

In the preceding sections we have derived interpolation formulae to approximate a function of a single variable. For a function of two or more variables, the formulae become complicated but a simpler procedure is to interpolate with respect to the first variable keeping the others constant, then interpolate with respect to the second variable, and so on. The method is illustrated below for a function of two variables. For a more efficient procedure for multivariate interpolation, see Section 5.3.

Example 3.26 The following table gives the values of z for different values of x and y. Find z when $x = 2.5$ and $y = 1.5$.

	x				
y	0	1	2	3	4
0	0	1	4	9	16
1	2	3	6	11	18
2	6	7	10	15	22
3	12	13	16	21	28
4	18	19	22	27	34

We first interpolate with respect to x keeping y constant. For $x = 2.5$, we obtain the following table using *linear interpolation*.

y	z
0	6.5
1	8.5
2	12.5
3	18.5
4	24.5

Now, we interpolate with respect to y using linear interpolation once again. For $y = 1/5$, we obtain

$$z = \frac{8.5 + 12.5}{2} = 10.5$$

so that $z(2.5, 1.5) = 10.5$. Actually, the tabulated function is $z = x^2 + y^2 + y$ and hence $z(2.5, 1.5) = 10.0$, so that the computed value has an error of 5%.

EXERCISES

3.1 Form a table of differences for the function

$$f(x) = x^3 + 5x - 7$$

for $x = -1, 0, 1, 2, 3, 4, 5$. Continue the table to obtain $f(6)$ and $f(7)$.

3.2 Evaluate

(a) $\Delta^2 x^3$ (b) $\Delta^2(\cos x)$ (c) $\Delta[(x + 1)(x + 2)]$

(d) $\Delta(\tan^{-1} x)$ (e) $\Delta\left[\dfrac{f(x)}{g(x)}\right]$.

3.3 Locate and correct the error in the following table:

x	2.5	3.0	3.5	4.0	4.5	5.0	5.5
y	4.32	4.83	5.27	5.47	6.26	6.79	7.23

3.4 Locate and correct the error in the following table:

x	1.00	1.05	1.10	1.15	1.20	1.25	1.30
e^x	2.7183	2.8577	3.0042	3.1528	3.3201	3.4903	3.6693

3.5 Prove the following:

(a) $y_x = y_{x-1} + \Delta y_{x-2} + \Delta^2 y_{x-3} + \cdots + \Delta^{n-1} y_{x-n} + \Delta^n y_{x-(n+1)}$

(b) $\Delta^n y_x = y_{x+n} - {}^nC_1\, y_{x+n-1} + {}^nC_2\, y_{x+n-2} + \cdots + (-1)^n y_x$

(c) $y_1 + y_2 + \cdots + y_n = {}^nC_1 y_1 + {}^nC_2\, \Delta y_1 + \cdots + \Delta^{n-1} y_1$

3.6 From the following table, find the number of students who obtained marks between 60 and 70:

Marks obtained	0–40	40–60	60–80	80–100	100–120
No. of students	250	120	100	70	50

3.7 Find the polynomial which approximates the following values:

x	3	4	5	6	7	8	9
y	13	21	31	43	57	73	91

If the number 31 is the fifth term of the series, find the first and the tenth terms of the series.

3.8 Find $f(0.23)$ and $f(0.29)$ from the following table:

x	0.20	0.22	0.24	0.26	0.28	0.30
$f(x)$	1.6596	1.6698	1.6804	1.6912	1.7024	1.7139

3.9 Prove that

(a) $\Delta = \mu\delta + \dfrac{\delta^2}{2}$ (b) $\Delta^3 y_2 = \nabla^3 y_5$

3.10 From the table of cubes given below, find $(6.36)^3$ and $(6.61)^3$.

x	6.1	6.2	6.3	6.4	6.5	6.6	6.7
x^3	226.981	238.328	250.047	262.144	274.625	287.496	300.763

3.11 Define the operators Δ, ∇, δ, E and E^{-1} and show that

(a) $\Delta^r y_k = \nabla^r y_{k+r} = \delta^r y_{k+\frac{r}{2}}$ (b) $\Delta\nabla y_k = \nabla\Delta y_k = \delta^2 y_k$

(c) $\mu\delta = \dfrac{\Delta + \nabla}{2}$ (d) $1 + \mu^2\delta^2 = \left(1 + \dfrac{1}{2}\delta^2\right)^2$

(e) $\Delta^2 = (1 + \Delta)\delta^2$ (f) $\Delta\left(\dfrac{1}{y_k}\right) = -\dfrac{\Delta y_k}{y_k y_{k+1}}$.

3.12 Show that

$$\left(\Delta - \frac{1}{2}\delta^2\right) = \delta\left(1 + \frac{\delta^2}{4}\right)^{1/2}$$

3.13 Find the missing terms in the following:

x	0	5	10	15	20	25	30
y	1	3	?	73	225	?	1153

3.14 Derive expressions for the errors in Newton's formulae of forward and backward differences. Estimate the maximum error made in any value of sin x in Example 3.6 obtained by interpolation in the range $15° \le x \le 40°$.

3.15 Certain values of x and $f(x)$ are given below. Find $f(1.235)$.

x	1.00	1.05	1.10	1.15	1.20	1.25
$f(x)$	0.682689	0.706282	0.728668	0.749856	0.769861	0.788700

3.16 Prove the following relations:

(a) $\delta^2 E = \Delta^2$ (b) $E^{-1/2} = \mu - \dfrac{\delta}{2}$ (c) $\nabla = \delta E^{-1/2}$

(d) $\Delta - \nabla = \delta^2$ (e) $\mu = \cosh \dfrac{hD}{2}$.

3.17 Using Gauss's forward formula, find the value of $f(32)$ given that $f(25) = 0.2707$, $f(30) = 0.3027$, $f(35) = 0.3386$ and $f(40) = 0.3794$.

3.18 State Gauss's backward formula and use it to find the value of $\sqrt{12525}$, given that $\sqrt{12500} = 111.8034$, $\sqrt{12510} = 111.8481$, $\sqrt{12520} = 111.8928$, $\sqrt{12530} = 111.9375$ and $\sqrt{12540} = 111.9822$.

3.19 State Stirling's formula for interpolation at the middle of a table of values and find $e^{1.91}$ from the following table:

x	1.7	1.8	1.9	2.0	2.1	2.2
e^x	5.4739	6.0496	6.6859	7.3891	8.1662	9.0250

3.20 Using Stirling's formula, find cos(0.17), given that cos(0) = 1, cos(0.05) = 0.9988, cos(0.10) = 0.9950, cos(0.15) = 0.9888, cos(0.20) = 0.9801, cos(0.25) = 0.9689, and cos(0.30) = 0.9553.

3.21 State Bessel's formula for interpolation and mention its limitations. Use this formula to solve the problem in 3.20.

3.22 The complete elliptic integral of the second kind is defined as

$$k(m) = \int_0^{\pi/2} \frac{1}{\sqrt{1 - m\sin^2\theta}}\, d\theta$$

Find $k(0.25)$ given that
$k(0.20) = 1.6596$, $k(0.22) = 1.6698$, $k(0.24) = 1.6804$, $k(0.26) = 1.6912$, $k(0.28) = 1.7024$, $k(0.30) = 1.7139$.

3.23 Using Bessel's formula, find $y(5)$ given that
$y(0) = 14.27$, $y(4) = 15.81$, $y(8) = 17.72$, and $y(12) = 19.96$.

3.24 From Bessel's formula, derive the following formula for midway interpolation:

$$y_{1/2} = \frac{1}{2}(y_0 + y_1) - \frac{1}{16}\left(\Delta^2 y_{-1} + \Delta^2 y_0\right) + \frac{3}{256}\left(\Delta^4 y_{-2} + \Delta^4 y_{-1}\right) + \cdots$$

Evaluate sin(0.20) given that
sin(0.15) = 0.1494, sin(0.17) = 0.1692, sin(0.19) = 0.1889, sin(0.21) = 0.2085, sin(0.23) = 0.2280.

3.25 Deduce Everett's formula from Bessel's formula and show that Everett's formula truncated after second differences is equivalent to Bessel's formula truncated after third differences. Use Everett's formula to find cos(12.5°) given that
cos(0°) = 1, cos(5°) = 0.9962, cos(10°) = 0.9848, cos(15°) = 0.9659, cos(20°) = 0.9397.

3.26 Using Everett's formula, evaluate $f(25)$ from the set of values
$f(20) = 2854$, $f(24) = 3162$, $f(28) = 3544$, $f(32) = 3992$.

3.27 State Lagrange's interpolation formula and find a bound for the error in linear interpolation.

3.28 Write an algorithm for Lagrange's formula. Find the polynomial which fits the following data

(–1, 7), (1, 5) and (2, 15).

3.29 Find $y(2)$ from the following data using Lagrange's formula

x	0	1	3	4	5
y	0	1	81	256	625

3.30 Let the values of the function $y = \sin x$ be tabulated at the abscissae 0, $\pi/4$ and $\pi/2$. If the Lagrange polynomial $L_2(x)$ is fitted to this data, find a bound for the error in the interpolated value.

3.31 Find a cubic polynomial which fits the data

(–2, –12), (–1, –8), (2, 3) and (3, 5).

3.32 Show that

$$\sum_{i=0}^{n} \frac{\Pi_{n+1}(x)}{(x - x_i)\Pi'_{n+1}(x_i)} = 1,$$

where $\Pi_{n+1}(x) = (x - x_0)(x - x_1)(x - x_2) \cdots (x - x_n)$.

3.33 In complex analysis, the residue theorem is used in the evaluation of contour integrals. If a function $f(z)$ is analytic inside and on a closed contour C, and $x_0, x_1, x_2, \ldots, x_n$ are simple poles inside C, then

$$\int_C f(z)dz = 2\pi i \quad \text{(sum of the residues at } x_0, x_1, \ldots, x_n)$$

[The residue of $f(z)$ at $z = a$ is defined as $\lim_{z \to a} (z - a) f(z)$]

If $x_0, x_1, x_2, \ldots, x_n$ are simple poles of a function $y(x)$ which is analytic inside and an a closed contour C, then show that

$$|y(x) - L_n(x)| = \frac{1}{2\pi i} \int_C \frac{y(t)\Pi_{n+1}(x)\,dt}{(t-x)\Pi_{n+1}(t)}$$

3.34 Lagrange's formula can be used to express a rational function as a sum of partial fractions (see, Stanton [1967]). Express

$$f(x) = \frac{x^2 + x - 3}{x^3 - 2x^2 - x + 2}$$

as a sum of partial fractions.

3.35 Establish Newton's divided-difference interpolation formula and give an estimate of the remainder term. Deduce Newton's forward and backward difference interpolation formulae as particular cases.

3.36 Given

$$f(x) = \frac{1}{x^2},$$

find the divided differences $[a, b]$ and $[a, b, c]$.

3.37 Given the set of tabulated points (0, 2), (1, 3), (2, 12) and (15, 3587) satisfying the function $y = f(x)$, compute $f(4)$ using Newton's divided difference formula.

3.38 The nth divided difference $[x_0, x_1, x_2, \ldots, x_n]$ can be expressed as the quotient of two determinants. Show that

$$[x_0, x_1, x_2] = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{vmatrix}}$$

3.39 Show that

$$[x_0, x_1] = \int_0^1 y'(x_0 t_0 + x_1 t_1)\, dt_1,$$

where $t_0 \geq 0$ and $t_0 + t_1 = 1$.

3.40 Tabulate values of $\sin x$ for $x = 0(0.01)$, 1.0 and find the maximum errors in linear and quadratic interpolations using Lagrange's formula.

3.41 If $f(x) = \dfrac{1}{x}$, prove that

$$[x_0, x_1, \ldots, x_r] = \frac{(-1)^r}{x_0\, x_1 \cdots x_r}.$$

3.42 Using Hermite's interpolation formula, estimate the value of ln 4.2 from the data (values of x, $\ln x$ and $\dfrac{1}{x}$):

(4.0, 1.38629, 0.25000), (4.5, 1.50408, 0.22222), (5.0, 1.60944, 0.20000).

3.43 Find the Hermite polynomial of the third degree approximating the function $y(x)$ such that

$$y(0) = 1, \quad y'(0) = 0,$$
$$y(1) = 3, \quad y'(1) = 5.$$

3.44 Values of x and $\sqrt[3]{x}$ are given below

(51, 3.708), (55, 3.803), (57, 3.848). Find x when $\sqrt[3]{x} = 3.780$.

3.45 From the table of values of x and e^x, viz. (1.4, 4.0552), (1.5, 4.4817), (1.6, 4.9530), (1.7, 5.4739), find x when $e^x = 4.7115$, using the method of successive approximations.

3.46 The second degree polynomial which satisfies the set of values (0, 1), (1, 2) and (2, 1) is

(a) $1 + 2x - x^2$ (b) $1 - 2x + x^2$ (c) $1 - 2x - x^2$ (d) $1 + 2x + x^2$

Find the correct alternative in the above.

3.47 If $\Delta y = 1 + 2x + 3x^2$, which one of the following is not true?

(a) $\Delta^2 y = 6x + 5$ (b) $\Delta^3 y = 6$

(c) $\Delta^4 y = 0$ (d) $y = x^2 + x^3$

3.48 Which of the following statements are true?

(a) $\Delta(\tan^{-1} x) = \tan^{-1} \dfrac{h}{1 + x(x+h)}$ (b) $\Delta(\cos 2x) = 2 \sin x\, (x + h)$

(c) $\Delta^2(x^3) = 6x$ (d) $\delta = \nabla E^{1/2}$

(e) $\Delta^2 = (1 - \Delta)\delta^2$ (f) $\mu\delta = \dfrac{\Delta + \nabla}{2}$

(g) $\Delta x^n = nx^{n-1}$ (h) $\left(\frac{\Delta^2}{E}\right)x^3 = 6x$

(i) $\Delta^n e^x = e^x$

3.49 For the function $z = f(x, y)$, the following values are given

$$f(0, 0) = 0, \quad f(0, 1) = 2, \quad f(0, 2) = 6,$$
$$f(1, 0) = 1, \quad f(1, 1) = 3, \quad f(1, 2) = 7,$$
$$f(2, 0) = 4, \quad f(2, 1) = 6, \quad f(2, 2) = 10.$$

Estimate the value of $f(0.5, 0.5)$ by the method of linear interpolation and compare the result with the actual value obtained from $z = x^2 + y^2 + y$.

3.50 Estimate the value of $f(1.5, 1.5)$ in Problem 3.49 and compare the value with its actual value.

Answers to Exercises

3.1 239, 371

3.2 (a) $6h^2(h + x)$, (b) $\cos(x + 2h) - 2\cos(x + h) + \cos x$

(c) $2x + 4$ (d) $\tan^{-1}\frac{h}{x(x+h)}$

(e) $\frac{f(x+h)g(x) - g(x+h)f(x)}{g(x)\,g(x+h)}$

3.3 5.74

3.4 3.1582 **3.6** 54

3.7 $y(1) = 3$, $y(10) = 111$ **3.8** 1.6751, 1.7082

3.10 257.259, 288.805 **3.13** 17,551

3.14 0.00000001 **3.15** 0.783172

3.17 0.3165 **3.18** 111.9152

3.19 6.7531 **3.20** 0.9856

3.21 0.9856 **3.22** 1.6858

3.23 16.25 **3.24** 0.1987

3.25 0.9763 **3.26** 3251

3.28 $\frac{1}{3}(11x^2 - 3x + 7)$ **3.29** 16

3.30 0.0239 **3.31** $-\frac{1}{15}x^3 - \frac{3}{20}x^2 + \frac{241}{60}x - \frac{39}{10}$

3.34 $-\frac{1}{2(x+1)} + \frac{1}{2(x-1)} + \frac{1}{x-2}$ **3.36** $\frac{-(a+b)}{a^2b^2}, \frac{ab+bc+ca}{a^2b^2c^2}$

3.37 1454 **3.40** $|y(x) - L_2(x)| = 6.415 \times 10^{-8}$

3.42 1.435081 **3.43** $1 + x^2 + x^3$

3.44 54 **3.45** 1.55

3.46 (a) **3.47** (d)

3.48 (a), (d), (f), (h) **3.49** 1.5

3.50 Proceed as in Problem 3.49.

Chapter 4

Least Squares and Fourier Transforms

4.1 INTRODUCTION

In experimental work, we often encounter the problem of fitting a curve to data which are subject to errors. The strategy for such cases is to derive an approximating function that *broadly* fits the data without necessarily passing through the given points. The curve drawn is such that the discrepancy between the data points and the curve is least. In the method of least squares, the sum of the squares of the errors is minimized. For continuous functions, the method is discussed in Section 4.4.

The problem of approximating a function by means of Chebyshev polynomials is described in Section 4.5. This is important from the standpoint of digital computation.

In Chapter 3, we concentrated on polynomial interpolation, i.e., interpolation based on a linear combination of functions 1, x, x^2, …, x^n. On the other hand, trigonometric interpolation, i.e., interpolation based on trigonometric functions such as, cos x, sin x, cos $2x$, sin $2x$, … plays an important role in modelling vibrating systems. The Fourier series is a useful tool for dealing with periodic systems; but for aperiodic systems, the Fourier transform is the primary tool available. The computations of discrete Fourier transform and the Fast Fourier Transform (FFT) are discussed in detail in Section 4.6.

4.2 LEAST SQUARES CURVE FITTING PROCEDURES

Let the set of data points be (x_i, y_i), $i = 1, 2, \ldots, m$, and let the curve given by $Y = f(x)$ be fitted to this data. At $x = x_i$, the given ordinate is y_i and the

corresponding value on the fitting curve is $f(x_i)$. If e_i is the error of approximation at $x = x_i$, then we have

$$e_i = y_i - f(x_i) \tag{4.1}$$

If we write

$$\begin{aligned} S &= [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \cdots + [y_m - f(x_m)]^2 \\ &= e_1^2 + e_2^2 + \cdots + e_m^2, \end{aligned} \tag{4.2}$$

then the method of least squares consists in minimizing S, i.e., the sum of the squares of the errors. In the following sections, we shall study the linear and nonlinear least squares fitting to given data (x_i, y_i), $i = 1, 2, \ldots, m$.

4.2.1 Fitting a Straight Line

Let $Y = a_0 + a_1x$ be the straight line to be fitted to the given data, viz. (x_i, y_i), $i = 1, 2, \ldots, m$. Then, corresponding to Eq. (4.2), we have

$$\begin{aligned} S = {} & [y_1 - (a_0 + a_1x_1)]^2 + [y_2 - (a_0 + a_1x_2)]^2 \\ & + \cdots + [y_m - (a_0 + a_1x_m)]^2 \end{aligned} \tag{4.3}$$

For S to be minimum, we have

$$\begin{aligned} \frac{\partial S}{\partial a_0} = 0 = {} & -2[y_1 - (a_0 + a_1x_1)] - 2[y_2 - (a_0 + a_1x_2)] \\ & - \cdots -2[y_m - (a_0 + a_1x_m)] \end{aligned} \tag{4.4a}$$

and

$$\begin{aligned} \frac{\partial S}{\partial a_1} = 0 = {} & -2x_1[y_1 - (a_0 + a_1x_1)] - 2x_2\,[y_2 - (a_0 + a_1x_2)] \\ & - \cdots -2x_m\,[y_m - (a_0 + a_1x_m)] \end{aligned} \tag{4.4b}$$

The above equations simplify to

$$ma_0 + a_1(x_1 + x_2 + \cdots + x_m) = y_1 + y_2 + \cdots + y_m \tag{4.5a}$$

and

$$a_0(x_1 + x_2 + \cdots + x_m) + a_1\left(x_1^2 + x_2^2 + \cdots + x_m^2\right) = x_1y_1 + x_2y_2 + \cdots + x_my_m \tag{4.5b}$$

or more compactly to

$$ma_0 + a_1\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \tag{4.6a}$$

and

$$a_0\sum_{i=1}^{m} x_i + a_1\sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_iy_i \tag{4.6b}$$

Equations (4.6) are called the *normal equations*, and can be solved for a_0 and a_1, since x_i and y_i are known quantities.

We can obtain easily

$$a_1 = \frac{m\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \cdot \sum_{i=1}^{m} y_i}{m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2} \tag{4.7}$$

and then

$$a_0 = \bar{y} - a_1\bar{x}. \tag{4.8}$$

Since $\frac{\partial^2 S}{\partial a_0^2}$ and $\frac{\partial^2 S}{\partial a_1^2}$ are both positive at the points a_0 and a_1, it follows that these values provide a *minimum* of S. In Eq. (4.8), $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively. From Eq. (4.8), we have

$$\bar{y} = a_0 + a_1\bar{x},$$

which shows that the fitted straight line passes through the centroid of the data points.

Sometimes, a goodness of fit is adopted. The correlation coefficient (cc) is defined as

$$\text{cc} = \sqrt{\frac{S_y - S}{S_y}}, \tag{4.9}$$

where

$$S_y = \sum_{i=1}^{m} (y_i - \bar{y})^2 \tag{4.10}$$

and S defined by Eq. (4.3).

If cc is close to 1, then the fit is considered to be good, although this is not always true.

Example 4.1 Find the best values of a_0 and a_1 if the straight line $Y = a_0 + a_1x$ is fitted to the data (x_i, y_i):

(1, 0.6), (2, 2.4), (3, 3.5), (4, 4.8), (5, 5.7)

Find also the correlation coefficient.

From the table of values given below, we find $\bar{x} = 3$, $\bar{y} = 3.4$, and

$$a_1 = \frac{5(63.6) - 15(17)}{5(55) - 225} = 1.26$$

Therefore,

$$a_0 = \bar{y} - a_1\bar{x} = -0.38.$$

x_i	y_i	x_i^2	x_iy_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1x_i)^2$
1	0.6	1	0.6	7.84	0.0784
2	2.4	4	4.8	1.00	0.0676
3	3.5	9	10.5	0.01	0.0100
4	4.8	16	19.2	1.96	0.0196
5	5.7	25	28.5	5.29	0.0484
15	17.0	55	63.6	16.10	0.2240

The correlation coefficient = $\sqrt{\dfrac{16.10 - 0.2240}{16.10}} = 0.9930$.

Example 4.2 Certain experimental values of x and y are given below:

$$(0, -1),\ (2, 5),\ (5, 12),\ (7, 20)$$

If the straight line $Y = a_0 + a_1x$ is fitted to the above data, find the approximate values of a_0 and a_1.

The table of values is given below.

x	y	x^2	xy
0	–1	0	0
2	5	4	10
5	12	25	60
7	20	49	140
14	36	78	210

The normal equations are

$$4a_0 + 14a_1 = 36$$

and

$$14a_0 + 78a_1 = 210$$

Solving the two equations, we obtain

$$a_0 = -1.1381 \quad \text{and} \quad a_1 = 2.8966$$

Hence the best straight line fit is given by

$$Y = -1.1381 + x(2.8966).$$

4.2.2 Multiple Linear Least Squares

Suppose that z is a linear function of two variables x and y. If the function $z = a_0 + a_1x + a_2y$ is fitted to the data (z_1, x_1, y_1), (z_2, x_2, y_2), $\ldots(z_m, x_m, y_m)$, then the sum

$$S = \sum_{i=1}^{m} (z_i - a_0 - a_1x_i - a_2y_i)^2$$

should be minimum. For this, we have

$$\frac{\partial S}{\partial a_0} = -2\sum(z_i - a_0 - a_1x_i - a_2y_i) = 0\,,$$

$$\frac{\partial S}{\partial a_1} = -2x_i\sum(z_i - a_0 - a_1x_i - a_2y_i) = 0\,,$$

and

$$\frac{\partial S}{\partial a_2} = -2y_i\sum(z_i - a_0 - a_1x_i - a_2y_i) = 0\,.$$

These equations simplify to

$$\left.\begin{aligned} ma_0 + a_1\Sigma x_i + a_2\Sigma y_i &= \Sigma z_i \\ a_0\Sigma x_i + a_1\Sigma x_i^2 + a_2\Sigma x_iy_i &= \Sigma z_ix_i \\ a_0\Sigma y_i + a_1\Sigma y_ix_i + a_2\Sigma y_i^2 &= \Sigma z_iy_i \end{aligned}\right\} \qquad (4.11)$$

from which a_0, a_1 and a_2 can be determined.

Example 4.3 Find the values of a_0, a_1 and a_2, so that the function $z = a_0 + a_1x + a_2y$ is fitted to the data (x, y, z) given below.

(0, 0, 2), (1, 1, 4), (2, 3, 3), (4, 2, 16) and (6, 8, 8).

We form the following table of values

x	y	z	x^2	xy	zx	y^2	yz
0	0	2	0	0	0	0	0
1	1	4	1	1	4	1	4
2	3	3	4	6	6	9	9
4	2	16	16	8	64	4	32
6	8	8	36	48	48	64	64
13	14	33	57	63	122	78	109

The normal equations are

$$5a_0 + 13a_1 + 14a_2 = 33$$

$$13a_0 + 57a_1 + 63a_2 = 122$$

$$14a_0 + 63a_1 + 78a_2 = 109$$

The solution of the above system is

$$a_0 = 2,\ a_1 = 5 \text{ and } a_2 = -3.$$

4.2.3 Linearization of Nonlinear Laws

The given data may not always follow a linear relationship. This can be ascertained from a plot of the given data. If a nonlinear model is to be fitted, it can be conveniently transformed to a linear relationship. Some nonlinear laws and their transformations are given as follows.

(a) $y = ax + \dfrac{b}{x}$

This can be written as

$$xy = ax^2 + b$$

Put $xy = Y$, $x^2 = X$. With these transformations, it becomes a linear model.

(b) $xy^a = b$

Taking logarithms of both sides, we get

$$\log_{10}x + a\log_{10}y = \log_{10}b.$$

In this case, we put

$$\log_{10}y = Y,\ \log_{10}x = X,$$

$$\frac{1}{a}\log_{10}b = A_0 \text{ and } -\frac{1}{a} = A_1,$$

so that

$$Y = A_0 + A_1X.$$

(c) $y = ab^x$

Taking logarithms of both sides, we obtain

$$\log_{10}y = \log_{10}a + x\log_{10}b$$
$$\Rightarrow Y = A_0 + A_1X,$$

where

$$Y = \log_{10}y,\ A_0 = \log_{10}a,$$
$$X = x, \text{ and } A_1 = \log_{10}b$$

(d) $y = ax^b$

We have

$$\log_{10}y = \log_{10}a + b\log_{10}x$$
$$\Rightarrow Y = A_0 + A_1X,$$

where

$$Y = \log_{10}y,\ A_0 = \log_{10}a,\ A_1 = b$$

and

$$X = \log_{10}x.$$

(e) $y = ae^{bx}$

In this case, we write

$$\ln y = \ln a + bx$$
$$\Rightarrow Y = A_0 + A_1X,$$

where

$$Y = \ln y,\ A_0 = \ln a,\ A_1 = b$$

and

$$X = x.$$

Example 4.4 Using the method of least squares, find constants a and b such that the function $y = ae^{bx}$ fits the following data:

(1.0, 2.473), (3.0, 6.722), (5.0, 18.274), (7.0, 49.673), (9.0, 135.026).

We have

$$y = ae^{bx}$$

Therefore,

$$\ln y = \ln a + bx$$

$$\Rightarrow Y = A_0 + A_1X,$$

where

$$Y = \ln y,\ A_0 = \ln a,\ A_1 = b \text{ and } X = x.$$

The table of values is given below

X	$Y = \ln y$	X^2	XY
1	0.905	1	0.905
3	1.905	9	5.715
5	2.905	25	14.525
7	3.905	49	27.335
9	4.905	81	44.145
25	14.525	165	92.625

We obtain

$$\bar{X} = 5,\ \bar{Y} = 2.905$$

$$A_1 = \frac{5(92.625) - 25(14.525)}{5(165) - 625} = 0.5 = b.$$

Then

$$A_0 = \bar{Y} - A_1\bar{X} = 2.905 - 0.5(5) = 0.405.$$

Hence,

$$a = e^{A_0} = e^{0.405} = 1.499.$$

It follows that the required curve is of the form

$$y = 1.499e^{0.5x}$$

Example 4.5 Using the method of least squares, fit a curve of the form $y = \dfrac{x}{a+bx}$ to the following data

(3, 7.148), (5, 10.231), (8, 13.509), (12, 16.434).

We have

$$y = \frac{x}{a+bx}$$

$$\Rightarrow \frac{1}{y} = \frac{a+bx}{x} = b + \frac{a}{x}$$

$$\Rightarrow Y = A_0 + A_1X,$$

where

$$A_0 = b,\ A_1 = a,\ X = \frac{1}{x} \text{ and } Y = \frac{1}{y}.$$

The table of values is

X	Y	X^2	XY
0.333	0.140	0.111	0.047
0.200	0.098	0.040	0.020
0.125	0.074	0.016	0.009
0.083	0.061	0.007	0.005
0.741	0.373	0.174	0.081

We obtain

$$A_1 = a = \frac{4(0.081) - 0.741(0.373)}{4(0.174) - (0.741)^2} = 0.324,\ \bar{X} = 0.185,\ \bar{Y} = 0.093$$

and $A_0 = b = \bar{Y} - a\bar{X} = 0.0331$.

Hence the required fit is $Y = 0.0331 + 0.324(X)$, which simplifies to

$$y = \frac{x}{0.324 + 0.0331(x)}.$$

$$\left[\text{Note: The given data is obtained from the relation } y = \frac{x}{0.3162 + 0.0345x}\right]$$

4.2.4 Curve Fitting by Polynomials

Let the polynomial of the nth degree,

$$Y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \tag{4.12}$$

be fitted to the data points (x_i, y_i), $i = 1, 2, \ldots, m$. We then have

$$\begin{aligned} S = &\left[y_1 - (a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_nx_1^n)\right]^2 \\ &+ \left[y_2 - (a_0 + a_1x_2 + a_2x_2^2 + \cdots + a_nx_2^n)\right]^2 \\ &+ \cdots + \left[y_m - (a_0 + a_1x_m + a_2x_m^2 + \cdots + a_nx_m^n)\right]^2 \end{aligned} \tag{4.13}$$

Equating to zero the first partial derivatives and simplifying, we obtain the normal equations:

$$\left.\begin{aligned} ma_0 + a_1\Sigma x_i + a_2\Sigma x_i^2 + \cdots + a_n\Sigma x_i^n &= \Sigma y_i, \\ a_0\Sigma x_i + a_1\Sigma x_i^2 + \cdots + a_n\Sigma x_i^{n+1} &= \Sigma x_i y_i, \\ &\vdots \\ a_0\Sigma x_i^n + a_1\Sigma x_i^{n+1} + \cdots + a_n\Sigma x_i^{2n} &= \Sigma x_i^n y_i \end{aligned}\right\} \tag{4.14}$$

where the summations are performed from $i = 1$ to $i = m$.

The system (4.14) constitutes $(n + 1)$ equations in $(n + 1)$ unknowns, and hence can be solved for $a_0, a_1, \ldots, a_n$. Equation (4.12) then gives the required polynomial of the nth degree.

For larger values of n, system (4.14) becomes unstable with the result that round off errors in the data may cause large changes in the solution. Such systems occur quite often in practical problems and are called *ill-conditioned* systems. Orthogonal polynomials are most suited to solve such systems and one particular form of these polynomials, the Chebyshev polynomials, will be discussed later in this chapter.

Example 4.6 Fit a polynomial of the second degree to the data points (x, y) given by

$$(0, 1), (1, 6) \text{ and } (2, 17).$$

For $n = 2$, Eq. (4.14) requires $\Sigma x_i, \Sigma x_i^2, \Sigma x_i^3, \Sigma x_i^4, \Sigma y_i, \Sigma x_i y_i$ and $\Sigma x_i^2 y_i$. The table of values is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
3	24	5	9	17	40	74

The normal equations are

$$3a_0 + 3a_1 + 5a_2 = 24$$
$$3a_0 + 5a_1 + 9a_2 = 40$$
$$5a_0 + 9a_1 + 17a_2 = 74$$

Solving the above system, we obtain

$$a_0 = 1, \; a_1 = 2 \text{ and } a_2 = 3.$$

The required polynomial is given by $Y = 1 + 2x + 3x^2$, and it can be seen that this fitting is *exact.*

Example 4.7 Fit a second degree parabola $y = a_0 + a_1x + a_2x^2$ to the data (x_i, y_i):

$$(1, 0.63), (3, 2.05), (4, 4.08), (6, 10.78).$$

The table of values is

x	y	x^2	x^3	x^4	xy	x^2y
1	0.63	1	1	1	0.63	0.63
3	2.05	9	27	81	6.15	18.45
4	4.08	16	64	256	16.32	65.28
6	10.78	36	216	1296	64.68	388.08
14	17.54	62	308	1634	87.78	472.44

The normal equations are

$$4a_0 + 14a_1 + 62a_2 = 17.54$$

$$14a_0 + 62a_1 + 308a_2 = 87.78$$

$$62a_0 + 308a_1 + 1634a_2 = 472.44,$$

from which we obtain

$$a_0 = 1.24,\ a_1 = -1.05 \quad \text{and} \quad a_2 = 0.44$$

4.2.5 Curve Fitting by a Sum of Exponentials

A frequently encountered problem in engineering and physics is that of fitting a sum of exponentials of the form

$$y = f(x) = A_1 e^{\lambda_1 x} + A_2 e^{\lambda_2 x} + \cdots + A_n e^{\lambda_n x} \tag{4.15}$$

to a set of data points (x_i, y_i), $i = 1, 2, \ldots, m$, where m is *much greater than* $2n$.

We describe here a computational technique due to Moore [1974]. For easy of presentation, we assume $n = 2$.

Then the function

$$y = A_1 e^{\lambda_1 x} + A_2 e^{\lambda_2 x} \tag{4.16}$$

is to be fitted to the data (x_i, y_i), $i = 1, 2, \ldots, m$, and $m \gg 4$. It is known that $y(x)$ satisfies a differential equation of the form

$$\frac{d^2 y}{dx^2} = a_1 \frac{dy}{dx} + a_2 y \tag{4.17}$$

where the constants a_1 and a_2 have to be determined. Integrating Eq. (4.17), we obtain

$$y'(x) - y'(x_0) = a_1[y(x) - y(0)] + a_2 \int_{x_0}^{x} y(x)\,dx, \tag{4.18}$$

where x_0 is the initial value of x and $y'(x) = \dfrac{dy}{dx}$. Integrating Eq. (4.18), we get

$$y(x) - y(0) - (x - x_0)\,y'(x_0) = a_1 \int_{x_0}^{x} y(x)\,dx - a_1(x - x_0)\,y(x_0)$$

$$+ a_2 \int_{x_0}^{x} \int_{x_0}^{x} y(x)\,dx\,dx \tag{4.19}$$

Now,

$$\int_{x_0}^{x}\int_{x_0}^{x} y(x)\,dx\,dx = \int_{x_0}^{x} (x-t)y(t)\,dt$$

Hence, Eq. (4.19) becomes

$$y(x) - y(0) - (x - x_0)\,y'(x_0) = a_1 \int_{x_0}^{x} y(x)\,dx - a_1(x - x_0)\,y(x_0)$$
$$+ a_2 \int_{x_0}^{x} (x - t)\,y(t)\,dt \tag{4.20}$$

In Eq. (4.20), $y'(x_0)$ is eliminated in the following way. Let x_1 and x_2 be two data points such that

$$x_0 - x_1 = x_2 - x_0 \tag{4.21}$$

Then Eq. (4.20) gives

$$y(x_1) - y(x_0) - (x_1 - x_0)\,y'(x_0) = a_1 \int_{x_0}^{x_1} y(x)\,dx - a_1(x_1 - x_0)\,y(x_0)$$
$$+ a_2 \int_{x_0}^{x_1} (x_1 - t)\,y(t)\,dt \tag{4.22}$$

and

$$y(x_2) - y(x_0) - (x_2 - x_0)\,y'(x_0) = a_1 \int_{x_0}^{x_2} y(x)\,dx - a_1(x_2 - x_0)\,y(x_0)$$
$$+ a_2 \int_{x_0}^{x_2} (x_2 - t)\,y(t)\,dt \tag{4.23}$$

Adding Eqs. (4.22) and (4.23) and using Eq. (4.21), we obtain

$$y(x_1) + y(x_2) - 2y(x_0) = a_1 \left[\int_{x_0}^{x_1} y(x)dx + \int_{x_0}^{x_2} y(x)\,dx \right]$$
$$+ a_2 \left[\int_{x_0}^{x_1} (x_1 - t)y(t)dt + \int_{x_0}^{x_2} (x_2 - t)\,y(t)\,dt \right] \tag{4.24}$$

Equation (4.24) can now be used to set up a linear system of equations for a_1 and a_2, and then we obtain λ_1 and λ_2 from the characteristic equation

$$\lambda^2 = a_1\lambda + a_2 \tag{4.25}$$

Finally, A_1 and A_2 can be obtained by the method of least squares or by the method of averages.

Example 4.8 Fit a function of the form

$$y = A_1 e^{\lambda_1 x} + A_2 e^{\lambda_2 x} \tag{i}$$

to the data defined by (x, y)

(1, 1.54), (1.1, 1.67), (1.2, 1.81), (1.3, 1.97), (1.4, 2.15),
(1.5, 2.35), (1.6, 2.58), (1.7, 2.83), (1.8, 3.11).

Let $x_0 = 1.2$, $x_1 = 1.0$, $x_2 = 1.4$. Then, Eq. (4.24) gives

$$0.07 = a_1\left[-\int_{1.0}^{1.2} y(x)\,dx + \int_{1.2}^{1.4} y(x)\,dx\right]$$

$$+ a_2\left[-\int_{1.0}^{1.2} (1.0 - t)y(t)\,dt + \int_{1.2}^{1.4} (1.4 - t)y(t)\,dt\right]$$

Evaluating the integrals by Simpson's rule* and simplifying, the above equation becomes

$$1.81a_1 + 2.180a_2 = 2.10 \tag{ii}$$

Again, choosing $x_1 = 1.4$, $x_0 = 1.6$ and $x_2 = 1.8$, and evaluating the integrals as before, we obtain the equation

$$2.88a_1 + 3.104a_2 = 3.00 \tag{iii}$$

Solving Eqs. (ii) and (iii), we get

$$a_1 = 0.03204 \quad \text{and} \quad a_2 = 0.9364.$$

Equation (4.25) now gives

$$\lambda^2 - 0.03204\lambda - 0.9364 = 0,$$

from which we obtain

$$\lambda_1 = 0.988 = 0.99,$$

and

$$\lambda_2 = -0.96.$$

Using the method of least squares, we finally obtain

$$A_1 = 0.499 \quad \text{and} \quad A_2 = 0.491.$$

The above data was actually constructed from the function $y = \cosh x$ so that $A_1 = A_2 = 0.5$, $\lambda_1 = 1.0$ and $\lambda_2 = -1.0$.

*See Section 6.4.2.

4.3 WEIGHTED LEAST SQUARES APPROXIMATION

In the previous section, we have minimized the sum of squares of the errors. A more general approach is to minimize the weighted sum of the squares of the errors taken over all data points. If this sum is denoted by S, then instead of Eq. (4.2), we have

$$S = W_1 [y_1 - f(x_1)]^2 + W_2 [y_2 - f(x_2)]^2 + \cdots + W_m [y_m - f(x_m)]^2$$
$$= W_1 e_1^2 + W_2 e_2^2 + \cdots + W_m e_m^2. \tag{4.26}$$

In Eq. (4.26), the W_i are prescribed positive numbers and are called *weights*. A weight is prescribed according to the relative accuracy of a data point. If all the data points are accurate, we set $W_i = 1$ for all i. We consider again the linear and nonlinear cases below.

4.3.1 Linear Weighted Least Squares Approximation

Let $Y = a_0 + a_1 x$ be the straight line to be fitted to the given data points, viz. $(x_1, y_1),\ldots,(x_m, y_m)$. Then

$$S(a_0, a_1) = \sum_{i=1}^{m} W_i [y_i - (a_0 + a_1 x_i)]^2. \tag{4.27}$$

For maxima or minima, we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = 0, \tag{4.28}$$

which give

$$\frac{\partial S}{\partial a_0} = -2 \sum_{i=1}^{m} W_i [y_i - (a_0 + a_1 x_i)] = 0 \tag{4.29}$$

and

$$\frac{\partial S}{\partial a_1} = -2 \sum_{i=1}^{m} W_i [y_i - (a_0 + a_1 x_i)] x_i = 0. \tag{4.30}$$

Simplification yields the system of equations for a_0 and a_1:

$$a_0 \sum_{i=1}^{m} W_i + a_1 \sum_{i=1}^{m} W_i x_i = \sum_{i=1}^{m} W_i y_i \tag{4.31}$$

and

$$a_0 \sum_{i=1}^{m} W_i x_i + a_1 \sum_{i=1}^{m} W_i x_i^2 = \sum_{i=1}^{m} W_i x_i y_i, \tag{4.32}$$

which are the *normal equations* in this case and are solved to obtain a_0 and a_1. We consider Example 4.2 again to illustrate the use of weights.

Example 4.9 Suppose that in the data of Example 4.2, the point (5, 12) is known to be more reliable than the others. Then we prescribe a weight (say, 10) corresponding to this point only and all other weights are taken as unity. The following table is then obtained.

x	y	W	Wx	Wx^2	Wy	Wxy
0	–1	1	0	0	–1	0
2	5	1	2	4	5	10
5	12	10	50	250	120	600
7	20	1	7	49	20	140
14	36	13	59	303	144	750

The normal Eqs. (4.31) and (4.32) then give

$$13a_0 + 59a_1 = 144 \tag{i}$$

$$59a_0 + 303a_1 = 750. \tag{ii}$$

Solution to Eqs. (i) and (ii) gives

$$a_0 = -1.349345 \quad \text{and} \quad a_1 = 2.73799.$$

The 'linear least squares approximation' is, therefore, given by

$$y = -1.349345 + 2.73799x.$$

Example 4.10 We consider Example 4.9 again with an increased weight, say 100, corresponding to $y(5.0)$. The following table is then obtained.

x	y	W	Wx	Wx^2	Wy	Wxy
0	–1	1	0	0	–1	0
2	5	1	2	4	5	10
5	12	100	500	2500	1200	6000
7	20	1	7	49	20	140
14	36	103	509	2553	1224	6150

The normal equations in this case are

$$103a_0 + 509a_1 = 1224 \tag{i}$$

and

$$509a_0 + 2553a_1 = 6150. \tag{ii}$$

Solving the preceding equations, we obtain

$$a_0 = -1.41258 \quad \text{and} \quad a_1 = 2.69056.$$

The required 'linear least squares approximation' is therefore given by

$$y = -1.41258 + 2.69056x,$$

and the value of $y(5) = 12.0402$.
It follows that the approximation becomes better when the weight is increased.

4.3.2 Nonlinear Weighted Least Squares Approximation

We now consider the least squares approximation of a set of m data points (x_i, y_i), $i = 1, 2, \ldots, m$, by a polynomial of degree $n < m$. Let

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{4.33}$$

be fitted to the given data points. We then have

$$S(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{m} W_i \, [y_i - (a_0 + a_1 x_i + \cdots + a_n x_i^n)]^2. \tag{4.34}$$

If a minimum occurs at $(a_0, a_1, \ldots, a_n)$, then we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \frac{\partial S}{\partial a_2} = \cdots = \frac{\partial S}{\partial a_n} = 0. \tag{4.35}$$

These conditions yield the normal equations

$$\left.\begin{aligned}
a_0 \sum_{i=1}^{m} W_i + a_1 \sum_{i=1}^{m} W_i x_i + \cdots + a_n \sum_{i=1}^{m} W_i x_i^n &= \sum_{i=1}^{m} W_i y_i \\
a_0 \sum_{i=1}^{m} W_i x_i + a_1 \sum_{i=1}^{m} W_i x_i^2 + \cdots + a_n \sum_{i=1}^{m} W_i x_i^{n+1} &= \sum_{i=1}^{m} W_i x_i y_i \\
&\vdots \\
a_0 \sum_{i=1}^{m} W_i x_i^n + a_1 \sum_{i=1}^{m} W_i x_i^{n+1} + \cdots + a_n \sum_{i=1}^{m} W_i x_i^{2n} &= \sum_{i=1}^{m} W_i x_i^n y_i.
\end{aligned}\right\} \tag{4.36}$$

Equations (4.36) are $(n + 1)$ equations in $(n + 1)$ unknowns $a_0, a_1, \ldots, a_n$. If the x_i are distinct with $n < m$, then the equations possess a 'unique' solution.

4.4 METHOD OF LEAST SQUARES FOR CONTINUOUS FUNCTIONS

In the previous sections, we considered the least squares approximations of discrete data. We shall, in the present section, discuss the least squares approximation of a continuous function on $[a, b]$. The summations in the normal equations are now replaced by definite integrals.

Let

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{4.37}$$

be chosen to minimize

$$S(a_0, a_1, \ldots, a_n) = \int_a^b W(x)\,[y(x) - (a_0 + a_1 x + \cdots + a_n x^n)]^2\, dx. \tag{4.38}$$

The necessary conditions for a minimum are given by

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \cdots = \frac{\partial S}{\partial a_n} = 0, \tag{4.39}$$

which yield

$$\left.\begin{aligned}
-2\int_a^b W(x)\,[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, dx &= 0 \\
-2\int_a^b W(x)\,[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x\, dx &= 0 \\
-2\int_a^b W(x)\,[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x^2\, dx &= 0 \\
&\vdots \\
-2\int_a^b W(x)\,[y(x) - (a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n)]\, x^n\, dx &= 0.
\end{aligned}\right\} \tag{4.40}$$

Rearrangement of terms in Eq. (4.40) gives the system

$$\left.\begin{aligned}
a_0 \int_a^b W(x)\,dx + a_1 \int_a^b xW(x)\,dx + \cdots + a_n \int_a^b x^n W(x)\,dx &= \int_a^b W(x)y(x)\,dx \\
a_0 \int_a^b xW(x)\,dx + a_1 \int_a^b x^2 W(x)\,dx + \cdots + a_n \int_a^b x^{n+1} W(x)\,dx &= \int_a^b x\,W(x)y(x)\,dx \\
&\vdots \\
a_0 \int_a^b x^n W(x)\,dx + a_1 \int_a^b x^{n+1} W(x)\,dx + \cdots + a_n \int_a^b x^{2n} W(x)\,dx &= \int_a^b x^n W(x)y(x)\,dx.
\end{aligned}\right\} \tag{4.41}$$

The system in Eq. (4.41) comprises $(n + 1)$ normal equations in $(n + 1)$ unknowns, viz. $a_0, a_1, a_2, \ldots, a_n$ and they always possess a 'unique' solution.

Example 4.11 Construct a least squares quadratic approximation to the function $y(x) = \sin x$ on $[0, \pi/2]$ with respect to the weight function $W(x) = 1$.

Let

$$y = a_0 + a_1 x + a_2 x^2 \tag{i}$$

be the required quadratic approximation. Then using Eq. (4.41), we obtain the system

$$\left.\begin{aligned}
a_0 \int_0^{\pi/2} dx + a_1 \int_0^{\pi/2} x\,dx + a_2 \int_0^{\pi/2} x^2 dx &= \int_0^{\pi/2} \sin x\,dx \\
a_0 \int_0^{\pi/2} x\,dx + a_1 \int_0^{\pi/2} x^2\,dx + a_2 \int_0^{\pi/2} x^3 dx &= \int_0^{\pi/2} x \sin x\,dx \\
a_0 \int_0^{\pi/2} x^2\,dx + a_1 \int_0^{\pi/2} x^3\,dx + a_2 \int_0^{\pi/2} x^4 dx &= \int_0^{\pi/2} x^2 \sin x\,dx.
\end{aligned}\right\} \tag{ii}$$

Simplifying Eq. (ii), we obtain

$$a_0 \frac{\pi}{2} + a_1 \frac{\pi^2}{8} + a_2 \frac{\pi^3}{24} = 1$$

$$a_0 \frac{\pi^2}{8} + a_1 \frac{\pi^3}{24} + a_2 \frac{\pi^4}{64} = 1$$

$$a_0 \frac{\pi^3}{24} + a_1 \frac{\pi^4}{64} + a_2 \frac{\pi^5}{160} = 2\left(\frac{\pi}{2} - 1\right),$$

whose solution is

$$\left.\begin{aligned}
a_0 &= \frac{18}{\pi} + \frac{96}{\pi^2} - \frac{480}{\pi^3} \\
a_1 &= -\frac{144}{\pi^2} - \frac{1344}{\pi^3} + \frac{5760}{\pi^4} \\
a_2 &= \frac{240}{\pi^3} + \frac{2880}{\pi^4} - \frac{11520}{\pi^5}.
\end{aligned}\right\} \tag{iii}$$

The required quadratic approximation to $y = \sin x$ on $[0, \pi/2]$ is then given by (i) and (iii),

As a check, we obtain, at $x = \pi/4$,

$$\sin x \approx -\frac{3}{\pi} - \frac{60}{\pi^2} + \frac{240}{\pi^3} = 0.706167587.$$

The true value of $\sin(\pi/4) = 0.707106781$, so that the error in the preceding solution is 0.000939194.

4.4.1 Orthogonal Polynomials

In the previous section, we have seen that the method of determining a least square approximation to a continuous function gives satisfactory results. However, this method possesses the disadvantage of solving a large linear system of equations. Besides, such a system may exhibit a peculiar tendency called *ill-conditioning*, which means that small change in any of its parameters introduces large errors in the solution—the degree of *ill-conditioning* increasing with the order of the system. Hence, alternative methods of solving the aforesaid least-squares problem have gained importance, and of these the method that employs 'orthogonal polynomials' is currently in use. This method possessess the great advantage that it does not require a linear system to be solved and is described below.

We choose the approximation in the form:

$$Y(x) = a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x), \tag{4.42}$$

where $f_j(x)$ is a polynomial in x of degree j.

Then, we write

$$S(a_0, a_1, \ldots, a_n) = \int_a^b W(x)\{y(x) - [a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)]\}^2 dx. \tag{4.43}$$

For S to be minimum, we must have

$$\left.\begin{aligned}
\frac{\partial S}{\partial a_0} &= 0 = -2\int_a^b W(x)\{y(x) - [a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)]\} f_0(x)\, dx \\
\frac{\partial S}{\partial a_1} &= 0 = -2\int_a^b W(x)\{y(x) - [a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)]\} f_1(x)\, dx \\
&\vdots \\
\frac{\partial S}{\partial a_n} &= 0 = -2\int_a^b W(x)\{y(x) - [a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)]\} f_n(x)\, dx
\end{aligned}\right\} \tag{4.44}$$

The normal equations are now given by

$$\left.\begin{aligned}
&a_0\int_a^b W(x)\,f_0^2(x)\,dx + a_1\int_a^b W(x)\,f_0(x)\,f_1(x)\,dx + \cdots + a_n\int_a^b W(x)\,f_0(x)\,f_n(x)\,dx \\
&\quad = \int_a^b W(x)\,y(x)\,f_0(x)\;dx \\
&a_0\int_a^b W(x)\,f_1(x)\;f_0(x)\,dx + a_1\int_a^b W(x)\,f_1^2(x)\,dx + \cdots + a_n\int_a^b W(x)\;f_1(x)\,f_n(x)\,dx \\
&\quad = \int_a^b W(x)\;y(x)\,f_1(x)\,dx \\
&\quad \vdots \\
&a_0\int_a^b W(x)\,f_n(x)\;f_0(x)\,dx + a_1\int_a^b W(x)\,f_n(x)\,f_1(x)\;dx + \cdots + a_n\int_a^b W(x)\;f_n^2(x)\,dx \\
&\quad = \int_a^b W(x)\;y(x)\,f_n(x)\,dx.
\end{aligned}\right\} \tag{4.45}$$

The above system can be written more simply as

$$\begin{aligned}
&a_0\int_a^b W(x)\;f_0(x)\;f_j(x)\,dx + a_1\int_a^b W(x)\;f_1(x)\;f_j(x)\,dx + \cdots \\
&\quad + a_n\int_a^b W(x)\,f_n(x)\,f_j(x)\,dx = \int_a^b W(x)\;y(x)\;f_j(x)\,dx, \qquad j = 0,\ 1,\ 2, \ldots, n.
\end{aligned} \tag{4.46}$$

In Eq. (4.45), we find products of the type $f_p(x)\,f_q(x)$ in the integrands, and if we assume that

$$\int_a^b W(x)\,f_p(x)\,f_q(x)\,dx = \begin{cases} 0, & p \neq q \\ \displaystyle\int_a^b W(x)\,f_p^2(x)\,dx, & p = q, \end{cases} \tag{4.47}$$

then the system (4.45) reduces to

$$\begin{aligned}
a_0\int_a^b W(x)\,f_0^2(x)\,dx &= \int_a^b W(x)\;y(x)\;f_0(x)\,dx \\
&\vdots \\
a_n\int_a^b W(x)\,f_n^2(x)\,dx &= \int_a^b W(x)\;y(x)\;f_n(x)\,dx.
\end{aligned}$$

From the preceding equations we obtain

$$a_j = \frac{\int_a^b W(x)\, y(x)\, f_j(x)\, dx}{\int_a^b W(x)\, f_j^2(x)\, dx}, \qquad j = 0, 1, 2, \ldots, n. \tag{4.48}$$

Substitution of a_0, a_1, …, a_n in Eq. (4.42) then yields the required least squares approximation, but the functions $f_0(x)$, $f_1(x)$, …, $f_n(x)$ are still not known. The $f_j(x)$, which are polynomials in x satisfying the conditions (4.47), are called *orthogonal polynomials* and are said to be orthogonal with respect to the weight function $W(x)$. They play an important role in numerical analysis and a few of them are listed below in Table 4.1.

Table 4.1 Orthogonal Polynomials*

Name	$f_j(x)$	*Interval*	$W(x)$
Jacobi	$P_n^{(\alpha,\beta)}(x)$	$[-1, 1]$	$(1-x)^{\alpha}(1+x)^{\beta}\ (\alpha, \beta > -1)$
Chebyshev (first kind)	$T_n(x)$	$[-1, 1]$	$(1-x^2)^{-1/2}$
Chebyshev (second kind)	$U_n(x)$	$[-1, 1]$	$(1-x^2)^{1/2}$
Legendre	$P_n(x)$	$(-1, 1)$	1
Laguerre	$L_n(x)$	$[0, \infty)$	e^{-x}
Hermite	$H_n(x)$	$(-\infty, \infty)$	e^{-x^2}

A brief discussion of some important properties of the Chebyshev polynomials $T_n(x)$ and their usefulness in the approximation of functions will be given in a later section of this chapter. We now return to our discussion of the problem of determining the least squares approximation. As we noted earlier, the functions $f_j(x)$ are yet to be determined. These are obtained by using the 'Gram–Schmidt orthogonalization process,' which has important applications in numerical analysis. This process is described in the next section.

4.4.2 Gram–Schmidt Orthogonalization Process

Suppose that the orthogonal polynomial $f_i(x)$, valid on the interval $[a, b]$, has the leading term x^i. Then, starting with

$$f_0(x) = 1 \tag{4.49}$$

*For more details concerning orthogonal polynomials, see Abramovitz and Stegun [1965].

we find that the linear polynomial $f_1(x)$, with leading term x, can be written as

$$f_1(x) = x + k_{1,0}\, f_0(x), \tag{4.50}$$

where $k_{1,0}$ is a constant to be determined. Since $f_1(x)$ and $f_0(x)$ are orthogonal, we have

$$\int_a^b W(x) f_0(x) f_1(x)\, dx = 0 = \int_a^b xW(x)\ f_0(x)\, dx + k_{1,0} \int_a^b W(x) f_0^2(x)\, dx$$

using Eqs. (4.47) and (4.49). From the above, we obtain

$$k_{1,0} = -\frac{\int_a^b x\, W(x) f_0(x)\, dx}{\int_a^b W(x) f_0^2(x)\, dx} \tag{4.51}$$

and Eq. (4.50) gives

$$f_1(x) = x - \frac{\int_a^b x\, W(x) f_0(x)\, dx}{\int_a^b W(x) f_0^2(x)\, dx}.$$

Now, the polynomial $f_2(x)$, of degree 2 in x and with leading term x^2, may be written as

$$f_2(x) = x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x), \tag{4.52}$$

where the constants $k_{2,0}$ and $k_{2,1}$ are to be determined by using the orthogonality conditions in Eq. (4.47). Since $f_2(x)$ is orthogonal to $f_0(x)$, we have

$$\int_a^b W(x) f_0(x)[x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x)]dx = 0.$$

Since $\int_a^b W(x)\ f_0(x)\ f_1(x)\, dx = 0$, the above equation gives

$$k_{2,0} = -\frac{\int_a^b x^2 W(x) f_0(x)\, dx}{\int_a^b W(x) f_0^2(x)\, dx} = -\frac{\int_a^b x^2 W(x)\, dx}{\int_a^b W(x)\, dx}. \tag{4.53}$$

Again, since $f_2(x)$ is orthogonal to $f_1(x)$, we have

$$\int_a^b W(x) f_1(x) [x^2 + k_{2,0} f_0(x) + k_{2,1} \ f_1(x)] \, dx = 0.$$

Using the condition that $\int_a^b W(x) \ f_0(x) \ f_1(x) \, dx = 0$, the above yields

$$k_{2,1} = -\frac{\int_a^b x^2 W(x) f_1(x) \, dx}{\int_a^b W(x) f_1^2(x) \, dx}. \tag{4.54}$$

Since $k_{2,0}$ and $k_{2,1}$ are known, Eq. (4.52) determines $f_2(x)$. Proceeding in this way, the method can be generalized and we write

$$f_j(x) = x^j + k_{j,0} f_0(x) + k_{j,1} f_1(x) + \cdots + k_{j,j-1} f_{j-1}(x), \tag{4.55}$$

where the constants $k_{j,i}$ are so chosen that $f_j(x)$ is orthogonal to $f_0(x)$, $f_1(x)$, …, $f_{j-1}(x)$. These conditions yield

$$k_{j,i} = -\frac{\int_a^b x^j W(x) f_i(x) \, dx}{\int_a^b W(x) f_i^2(x) \, dx}. \tag{4.56}$$

Since the a_i and $f_i(x)$ in Eq. (4.42) are known, the approximation $Y(x)$ can now be determined. The following example illustrates the method of procedure.

Example 4.12 Obtain the first-four orthogonal polynomials $f_n(x)$ on [–1, 1] with respect to the weight function $W(x) = 1$.

Let $f_0(x) = 1$. Then Eq. (4.51) gives

$$k_{1,0} = -\frac{\int_{-1}^{1} x \, dx}{\int_{-1}^{1} dx} = 0.$$

We then obtain from Eq. (4.50), $f_1(x) = x$. Equations (4.53) and (4.54) give respectively

$$k_{2,0} = -\frac{\int_{-1}^{1} x^2 dx}{\int_{-1}^{1} dx} = -\frac{1}{3}$$

and

$$k_{2,1} = -\frac{\int_{-1}^{1} x^2 x\, dx}{\int_{-1}^{1} x^2 dx} = 0.$$

Then Eq. (4.52) yields $f_2(x) = x^2 - 1/3$.
In a similar manner, we obtain

$$k_{3,0} = -\frac{\int_{-1}^{1} x^3 dx}{\int_{-1}^{1} dx} = 0,$$

$$k_{3,1} = -\frac{\int_{-1}^{1} x^3 x\, dx}{\int_{-1}^{1} x^2\, dx} = -\frac{3}{5},$$

$$k_{3,2} = -\frac{\int_{-1}^{1} x^3 (x^2 - 1/3)\, dx}{\int_{-1}^{1} (x^2 - 1/3)^2 dx} = 0.$$

It is easily verified that

$$f_3(x) = x^3 - \frac{3}{5}x.$$

Thus the required orthogonal polynomials are 1, x, $x^2 - 1/3$ and $x^3 - (3/5)x$. These polynomials are called *Legendre polynomials* and are usually denoted by $P_n(x)$. It is easy to verify that these polynomials satisfy the orthogonal property given in Eq. (4.47). An important application of Legendre polynomials occurs in numerical quadrature (see Chapter 6).

4.5 APPROXIMATION OF FUNCTIONS

The problem of approximating a function is a central problem in numerical analysis due to its importance in the development of software for digital computers. Function evaluation through interpolation techniques over stored

table of values has been found to be quite costlier when compared to the use of efficient function approximations.

Let $f_1, f_2, \ldots, f_n$ be the values of the given function and $\phi_1, \phi_2, \ldots, \phi_n$ be the corresponding values of the approximating function. Then the error vector is **e**, where the components of **e** are given by $e_i = f_i - \phi_i$. The approximation may be chosen in a number of ways. For example, we may find the approximation such that the quantity $\sqrt{(e_1^2 + e_2^2 + \cdots + e_n^2)}$ is *minimum.* This leads us to the least squares approximation which we have already studied. On the other hand, we may choose the approximation such that the maximum component of **e** is minimized. This leads us to the 'celebrated Chebyshev polynomials' which have found important application in the approximation of functions in digital computers.

In this section, we shall give a brief outline of Chebyshev polynomials and their applications in the economization of power series.*

4.5.1 Chebyshev Polynomials

The Chebyshev polynomial of degree n over the interval [–1, 1] is defined by the relation

$$T_n(x) = \cos(n \cos^{-1} x), \tag{4.57}$$

from which follows immediately the relation

$$T_n(x) = T_{-n}(x). \tag{4.58}$$

Let $\cos^{-1} x = \theta$ so that $x = \cos\theta$ and (4.57) gives

$$T_n(x) = \cos n\theta.$$

Hence

$$T_0(x) = 1 \quad \text{and} \quad T_1(x) = x.$$

Using the trigonometric identity

$$\cos(n-1)\theta + \cos(n+1)\theta = 2\cos n\theta \ \cos\theta,$$

we obtain easily

$$T_{n-1}(x) + T_{n+1}(x) = 2xT_n(x),$$

which is the same as

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{4.59}$$

This is the *recurrence relation* which can be used to successively compute all $T_n(x)$, since we know $T_0(x)$ and $T_1(x)$. The first seven Chebyshev polynomials are:

*Refer to Fox and Parker [1968] for further details and other applications of Chebyshev polynomials.

$$\left.\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1.
\end{aligned}\right\} \tag{4.60}$$

The graph of the first four Chebyshev polynomials are shown in Fig. 4.1

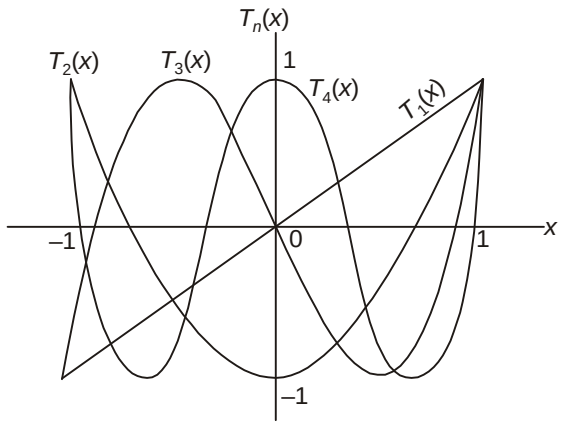


Figure 4.1 Chebyshev polynomials $T_n(x)$, $n = 1, 2, 3, 4$.

It is easy to see that the coefficient of x^n in $T_n(x)$ is always 2^{n-1}. Further, if we set $y = T_n(x) = \cos n\theta$, then we get

$$\frac{dy}{dx} = \frac{n \sin n\theta}{\sin\theta}$$

and

$$\frac{d^2y}{dx^2} = \frac{-n^2 \cos n\theta + n \sin n\theta \cot\theta}{\sin^2\theta} = \frac{-n^2 y + x(dy/dx)}{1 - x^2}$$

so that

$$(1 - x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + n^2 y = 0, \tag{4.61}$$

which is the *differential equation satisfied* by $T_n(x)$.

It is also possible to express powers of x in terms of Chebyshev polynomials. We find

$$\left.\begin{aligned}
1 &= T_0(x) \\
x &= T_1(x) \\
x^2 &= \frac{1}{2}[T_0(x)+T_2(x)] \\
x^3 &= \frac{1}{4}[3T_1(x)+T_3(x)] \\
x^4 &= \frac{1}{8}[3T_0(x)+4T_2(x)+T_4(x)] \\
x^5 &= \frac{1}{16}[10T_1(x)+5T_3(x)+T_5(x)] \\
x^6 &= \frac{1}{32}[10T_0(x)+15T_2(x)+6T_4(x)+T_6(x)].
\end{aligned}\right\} \tag{4.62}$$

and so on. These expressions will be useful in the economization of power series to be discussed later.

An important property of $T_n(x)$ is given by

$$\int_{-1}^{1} \frac{T_m(x)\,T_n(x)\,dx}{\sqrt{1-x^2}} = \begin{cases} 0, & m \neq n \\ \pi/2, & m = n \neq 0 \\ \pi, & m = n = 0 \end{cases} \tag{4.63}$$

that is, the polynomials $T_n(x)$ are *orthogonal* with the function $1/\sqrt{(1-x^2)}$. This property is easily proved since by putting $x = \cos\theta$, the above integral becomes

$$\int_0^{\pi} T_m(\cos\theta)\,T_n(\cos\theta)\,d\theta = \int_0^{\pi} \cos m\theta \cos n\theta\, d\theta$$

$$= \left[\frac{\sin(m+n)\,\theta}{2(m+n)} + \frac{\sin(m-n)\,\theta}{2(m-n)}\right]_0^{\pi},$$

from which follow the values given on the right side of Eq. (4.63).

We have seen above that $T_n(x)$ is a polynomial of degree n in x and that the coefficient of x^n in $T_n(x)$ is 2^{n-1}. In approximation theory, one uses *monic* polynomials, i.e. Chebyshev polynomials in which the coefficient of x^n is unity. If $P_n(x)$ is a monic polynomial, then we can write

$$P_n(x) = 2^{1-n}\,T_n(x), \qquad (n \geq 1). \tag{4.64}$$

A remarkable property of Chebyshev polynomials is that *of all monic polynomials*, $P_n(x)$, *of degree n whose leading coefficient equals unity, the*

polynomial $2^{1-n}T_n(x)$, *has the smallest least upper bound for its absolute value in the range* (–1, 1). Since $|T_n(x)| \le 1$, the upper bound referred to above is 2^{1-n}. Thus, in Chebyshev approximation, the maximum error is kept down to a minimum. This is often referred to as *minimax principle* and the polynomial in Eq. (4.64) is called the *minimax polynomial*. By this process we can obtain the best lower-order approximation, called the *minimax approximation*, to a given polynomial. This is illustrated in the following example.

Example 4.13 Find the best lower-order approximation to the cubic $2x^3+3x^2$. Using the relations given in Eq. (4.62), we write

$$\begin{aligned} 2x^3+3x^2 &= \frac{2}{4}[T_3(x)+3T_1(x)]+3x^2 \\ &= 3x^2+\frac{3}{2}T_1(x)+\frac{1}{2}T_3(x) \\ &= 3x^2+\frac{3}{2}x+\frac{1}{2}T_3(x), \qquad \text{since } T_1(x)=x. \end{aligned}$$

The polynomial $3x^2 + (3/2)\,x$ is the required lower-order approximation to the given cubic with a maximum error ±1/2 in the range (–1, 1).
A similar application of Chebyshev series in the *economization* of power series is discussed next.

4.5.2 Economization of Power Series

To describe this process, which is essentially due to Lanczos, we consider the power series expansion of $f(x)$ in the form

$$f(x) = A_0 + A_1x + A_2x^2 + \cdots + A_nx^n, \qquad (-1 \le x \le 1). \tag{4.65}$$

Using the relations given in Eq. (4.62), we convert the above series into an expansion in Chebyshev polynomials. We obtain

$$f(x) = B_0 + B_1T_1(x) + B_2T_2(x) + \cdots + B_nT_n(x). \tag{4.66}$$

For a large number of functions, an expansion as in Eq. (4.66) above, converges more rapidly than the power series given by Eq. (4.65). This is known as *economization of the power series* and is illustrated in Example 4.14.

Example 4.14 Economize the power series

$$\sin x \approx x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040}.$$

Since 1/5040 = 0.000198…, the truncated series, viz.,

$$\sin x \approx x - \frac{x^3}{6} + \frac{x^5}{120} \tag{i}$$

will produce a change in the fourth decimal place only. We now convert the powers of x in Eq. (i) into Chebyshev polynomials by using the relations given in Eq. (4.62). This gives

$$\sin x \approx T_1(x) - \frac{1}{24}[3T_1(x) + T_3(x)] + \frac{1}{120\times 16}[10T_1(x) + 5T_3(x) + T_5(x)].$$

Simplifying the above, we obtain

$$\sin x \approx \frac{169}{192}T_1(x) - \frac{5}{128}T_3(x) + \frac{1}{1920}T_5(x). \tag{ii}$$

Since 1/1920 = 0.00052…, the truncated series, viz.,

$$\sin x = \frac{169}{192}T_1(x) - \frac{5}{128}T_3(x) \tag{iii}$$

will produce a change in the fourth decimal place only. Using the relations given in Eq. (4.60), the economized series is, therefore, given by

$$\sin x \approx \frac{169}{192}x - \frac{5}{128}(4x^3 - 3x) = \frac{383}{384}x - \frac{5}{32}x^3.$$

4.6 FOURIER APPROXIMATION

The approximation of a function by means of Fourier series, i.e., by a series of sines and cosines, is found useful in applications involving oscillating or vibrating systems. Let the function $f(t)$ be a periodic function with period $T > 0$, i.e., let

$$f(t + T) = f(t), \tag{4.67}$$

where T is the smallest value satisfying Eq. (4.67). Then the Fourier series for $f(t)$ is written as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right), \tag{4.68}$$

where a_n and b_n are real numbers independent of t and $\omega_0 = 2\pi/T$ is called the *fundamental frequency.* The coefficients $2\pi k/T$, $k = 2, 3,\ldots$ are called *harmonics*.

Integrating both the sides of Eq. (4.68) from 0 to T, we obtain

$$\int_0^T f(t)\,dt = \frac{a_0}{2}\int_0^T dt + \int_0^T\left(a_n \cos\frac{2\pi nt}{T} + b_n \sin\frac{2\pi nt}{T}\right)dt = \frac{a_0}{2}T,$$

since

$$\int_0^T \cos\left(\frac{2\pi nt}{T}\right)dt = \int_0^T \sin\left(\frac{2\pi nt}{T}\right)dt = 0.$$

Hence

$$a_0 = \frac{2}{T}\int_0^T f(t)\,dt. \tag{4.69}$$

Again, multiplying both the sides of Eq. (4.68) by $\cos(2\pi nt/T)$ and then integrating from 0 to T, we get

$$a_n = \frac{2}{T}\int_0^T f(t)\cos\left(\frac{2\pi nt}{T}\right)dt, \tag{4.70}$$

since

$$\int_0^T \cos\left(\frac{2\pi nt}{T}\right)\sin\left(\frac{2\pi nt}{T}\right)dt = 0.$$

Finally, multiplying both the sides of Eq. (4.68) by $\sin(2\pi nt/T)$ and then integrating from 0 to T, we obtain

$$b_n = \frac{2}{T}\int_0^T f(t)\sin\left(\frac{2\pi nt}{T}\right)dt. \tag{4.71}$$

Thus the coefficients a_0, a_n and b_n in the representation (4.68) are evaluated. If $T = 2\pi$, i.e. if $f(t)$ is of period 2π, Eqs. (4.69)–(4.71) become:

$$\left.\begin{aligned} a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)dt, \\ a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos nt\,dt, \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt\,dt. \end{aligned}\right\} \tag{4.72}$$

The Fourier series becomes further simplified if $f(t)$ is an even or odd function. If $f(t)$ is even, then we have

$$\left.\begin{aligned} f(t) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt, \\ \text{where} \quad & \\ a_n &= \frac{2}{\pi}\int_0^{\pi} f(t)\cos nt\,dt, \end{aligned}\right\} \tag{4.73}$$

since $b_n = 0$.

Similarly, if $f(t)$ is an odd function, then we have

$$\left.\begin{aligned} f(t) &= \sum_{n=1}^{\infty} b_n \sin nt, \\ \text{where} \qquad & \\ b_n &= \frac{2}{\pi}\int_0^{\pi} f(t)\sin nt\, dt. \end{aligned}\right\} \tag{4.74}$$

since $a_0 = a_n = 0$.

The formulae (4.68)–(4.71) can be expressed in a different way. For this, the well-known relations are used:

$$\cos nt = \frac{e^{int} + e^{-int}}{2} \quad \text{and} \quad \sin nt = \frac{e^{int} - e^{-int}}{2i}. \tag{4.75}$$

Using Eq. (4.75), Eqs. (4.68)–(4.71) can be expressed as

$$f(t) = \sum_{p=-\infty}^{\infty} A_p e^{2\pi ipt/T}, \tag{4.76}$$

where

$$A_p = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\ e^{-2\pi ipt/T} dt, \qquad p = 0, 1, 2, \ldots \tag{4.77}$$

These formulae directly lead us to the discussion of Fourier transforms but, before this, we consider an illustrative example on Fourier series.

Example 4.15 Find the Fourier series of the function defined by

$$f(t) = \begin{cases} -1, & -\pi < t < 0 \\ 0, & t = 0 \\ 1, & 0 < t < \pi. \end{cases}$$

The graph of the given function is shown in Fig. 4.2

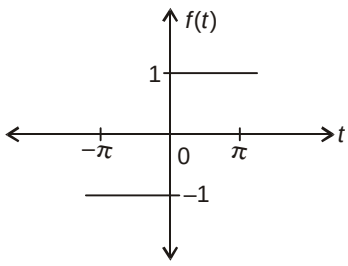


Figure 4.2

From the graph, it can be seen that $f(t)$ is an odd function. Hence the Fourier series for $f(t)$ contains only the coefficients b_n.

We, therefore, have

$$f(t) = \sum_{n=1}^{\infty} b_n \sin nt,$$

where

$$\begin{aligned} b_n &= \frac{2}{\pi}\int_0^{\pi} f(t)\sin nt\, dt \\ &= \frac{2}{\pi}\int_0^{\pi} \sin nt\, dt, \qquad \text{since } f(t) = 1 \\ &= \frac{2}{\pi}\left[-\frac{1}{n}\cos nt\right]_0^{\pi} \\ &= \frac{2}{n\pi}[1-(-1)^n] \\ &= \frac{4}{n\pi}, \qquad n = 1, 3, 5, \ldots \end{aligned}$$

It follows that

$$f(t) = \sum_{n=1,3,5,\ldots}^{\infty} \frac{4}{n\pi}\sin nt = \frac{4}{\pi}\left(\sin t + \frac{1}{3}\sin 3t + \frac{1}{5}\sin 5t + \cdots\right).$$

4.6.1 Fourier Transform

In the preceding section, we considered the Fourier series for periodic functions. There exist, however, several functions which are not periodic. Similarly, we come across, in nature, many phenomena (for example, lightning) which are *aperiodic*. The study of such phenomena is of great importance to the engineer. In such cases, the Fourier transform is the applicable tool and this can be derived, from Eqs. (4.76) and (4.77), by making T approach infinity so that the function becomes aperiodic. When $T \to \infty$, Eq. (4.77) can be written in the form

$$F(i\omega_0) = \int_{-\infty}^{\infty} f(t)\, e^{-i\omega_0 t} dt, \tag{4.78}$$

and is called the *Fourier transform* of $f(t)$. Similarly, Eq. (4.76) is written as

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(i\omega_0) e^{i\omega_0 t} d\omega_0, \tag{4.79}$$

and is called the *inverse Fourier transform* of $f(t)$. Equations (4.78) and (4.79) enable us to transform from time domain to frequency domain and from frequency to time domain, respectively. Physically, $F(i\omega_0)$ represents the frequency content of the signal. In Eq. (4.78), the function $f(t)$ is given in the continuous form which is rarely the case with a signal. In fact, the function

$f(t)$ is available only in a discrete form. In such a case, discrete analogues of both the integrals are used to compute the transforms. These equations, called the *Discrete Fourier Transforms,* are discussed below.

4.6.2 Discrete Fourier Transform (DFT)

Let $f(t)$ be specified at the points t_i, $i = 0, 1, 2, \ldots, N-1$, and $\Delta t = \dfrac{T}{N}$. If f_k denotes the value of $f(t)$ at $t = t_k$, then the discrete Fourier transform (DFT) and the *inverse discrete Fourier transform* (IDFT) are defined by

$$F_p = \sum_{k=0}^{N-1} f_k \,.\, e^{-2\pi ikp/N}, \qquad p = 0, 1, 2, \ldots, N-1 \tag{4.80}$$

and

$$f_k = \frac{1}{N}\sum_{p=0}^{N-1} F_p \,.\, e^{2\pi ikp/N}, \qquad k = 0, 1, 2, \ldots, N-1 \tag{4.81}$$

Denoting

$$W_N = e^{-2\pi i/N}, \tag{4.82}$$

Equations (4.80) and (4.81) become

$$F_p = \sum_{k=0}^{N-1} f_k \,.\, W_N^{kp}, \qquad p = 0, 1, 2, \ldots, N-1 \tag{4.83}$$

and

$$f_k = \frac{1}{N}\sum_{p=0}^{N-1} F_p \cdot W_N^{-kp}, \qquad k = 0, 1, 2, \ldots, N-1 \tag{4.84}$$

The above equations are, respectively, called the discrete Fourier transform (DFT) and the inverse DFT. The coefficients $|F_p|$ form a periodic sequence when extended outside of the range $p = 0, 1, 2, \ldots, N-1$, and we have

$$F_{p+N} = F_p \tag{4.85}$$

From Eq. (4.83), it may be seen that to compute each point of the DFT, we have to perform N complex multiplications and $(N-1)$ complex additions. Hence the N–point DFT requires N^2 complex multiplications and $N(N-1)$ complex additions.

Properties of W_N

(i) Symmetric property

$$W_N^{k+\frac{N}{2}} = W_N^k \cdot W_N^{N/2}$$

$$= -W_N^k, \text{ since } W_N^{N/2} = e^{-2\pi i \frac{N}{2N}} = e^{-\pi i} = -1.$$

(ii) Periodic property

$$W_N^{k+N} = W_N^k \cdot W_N^N$$
$$= W_N^k, \text{ since } W_N^N = e^{-2\pi iN/N} = e^{-2\pi i} = 1.$$

(iii) Another useful property

$$W_N^2 = e^{-2\pi i 2/N} = e^{-4\pi i/N}$$

and

$$W_{N/2} = e^{-2\pi i/N/2} = e^{-4\pi i/N}$$

Hence

$$W_{N/2} = W_N^2 .$$

A useful analysis that is important in the practical applications of Fourier transform (such as *smoothing of noisy data)* is called the *power spectrum* which is a plot of the power versus frequency. If $f(t)$ is a discrete time signal with period N, then the *power* P is defined by the relation

$$P = \frac{1}{N}\sum_{k=0}^{N-1} |F_k|^2 = \sum_{k=0}^{N-1} |f_k|^2 \tag{4.86}$$

Therefore, the sequence

$$P_k = |f_k|^2, \quad k = 0, 1, 2, \ldots, N-1 \tag{4.87}$$

is the distribution of power as a function of frequency and is called the *power density spectrum* of the periodic signal. Since F_k is a periodic sequence with period N, it follows that the spectrum of F_k (k = 0, 1, 2, …, N – 1) is also a periodic sequence with period N.

Matrix Representations of Equations (4.83) *and* (4.84).
We have

$$F(p) = \sum_{k=0}^{N-1} f_k W_N^{kp}, \quad p = 0, 1, 2, \ldots, N-1.$$
$$= f_0 + f_1 W_N^p + f_2 W_N^{2p} + \cdots + f_{N-1} W_N^{(N-1)p}, \quad p = 0, 1, \ldots, N-1.$$

Putting p = 0, 1, 2, …, N – 1, in the above equation, we obtain

$$F_0 = f_0 + f_1 W_N^0 + f_2 W_N^0 + \cdots + f_{N-1} W_N^0$$
$$F_1 = f_0 + f_1 W_N^1 + f_2 W_N^2 + \cdots + f_{N-1} W_N^{N-1}$$
$$\vdots \qquad\qquad \vdots$$
$$F_{N-1} = f_0 + f_1 W_N^{N-1} + f_2 W_N^{2(N-1)} + \cdots + F_{N-1} W_N^{(N-1)(N-1)}$$

The preceding can be expressed in the matrix form:

$$\begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_{N-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)(N-1)} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{bmatrix} \tag{4.88}$$

or

$$[F] = [W_N][f]$$

In a similar way, Eq. (4.84) can be expressed as

$$[f] = \frac{1}{N}\left[W_N^*\right][F], \tag{4.89}$$

where W_N^* is the complex conjugate of W_N.

Example 4.16 Using matrices, find the DFT of the sequence

$$f_k = \{1,\ 2,\ 3,\ 4\}.$$

We have

$$F_p = \sum_{k=0}^{3} W_4^{kp} \cdot f_k,\ \ p = 0, 1, 2, 3.$$

The matrix representation is

$$\begin{bmatrix} F_0 \\ F_1 \\ F_2 \\ F_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & W_4^1 & W_4^2 & W_4^3 \\ 1 & W_4^2 & W_4^4 & W_4^6 \\ 1 & W_4^3 & W_4^6 & W_4^9 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 10 \\ -2+2i \\ -2 \\ -2-2i \end{bmatrix}$$

Example 4.17 Find the inverse DFT of the sequence

$$F_p = \{1,\ 1-i,\ -1,\ 1+i\}.$$

We have

$$[f] = \frac{1}{N}\left[W_N^*\right][F]$$

For $N = 4$,

$$[W_n] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}$$

Therefore,

$$[W_N^*] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}$$

Hence

$$[f] = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}\begin{bmatrix} 1 \\ 1-i \\ -1 \\ 1+i \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 \\ 4 \\ -2 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 \\ 1.0 \\ -0.5 \\ 0 \end{bmatrix}$$

Example 4.18 Find the DFT of the sequence

$f_k = \{2, 2, 2, \ldots, 2\}$, for $k = 0, 1, 2, \ldots, N-1$.

We have

$$F_p = \sum_{k=0}^{N-1} f_k\, e^{-2\pi ipk/N}, \qquad p = 0, 1, 2, \ldots, N-1$$

$$= 2\sum_{k=0}^{N-1} e^{-2\pi ikp/N}, \qquad \text{since } f(k) = 2 \text{ for all } k.$$

$$= 2\left[1 + e^{-2\pi ip/N} + e^{-2\pi i2p/N} + \cdots + e^{-2\pi ip(N-1)/N}\right],$$

which is a geometric progression with $r = e^{-2\pi ip/N}$.

$$= 2\cdot\frac{1-r^N}{1-r}.$$

i.e.,

$$F_p = 2\cdot\left[\frac{1-e^{-2\pi ip}}{1-e^{-2\pi ip/N}}\right], \qquad p = 0, 1, 2, \ldots, N-1.$$

For $p = 1, 2, 3, \ldots, N-1$,

$$F_p = 0.$$

For $p = 0$, F_p is of the form $\frac{0}{0}$. By L'Hospital's rule, we obtain

$$F_0 = 2 \lim_{p \to 0} \frac{1 - e^{-2\pi ip}}{1 - e^{-2\pi ip/N}} = 2N.$$

Hence

$$F_p = \begin{cases} 2N, & p = 0 \\ 0 & p = 1, 2, 3, \ldots, N-1 \end{cases}$$

4.6.3 Fast Fourier Transform (FFT)

The computation of DFT by formula (4.83) requires N^2 complex multiplications and $N(N-1)$ complex additions. It also requires memory to store the values of $f(t)$ and W_N^{kp}. Besides, it does not make use of the periodic and symmetric properties of W_N^{kp}. As N increases, the computation of DFT demands very high memory requirements and becomes a time–consuming process.

The *Fast Fourier Transform* (FFT) algorithms make use of the symmetric and periodic properties of W_N^{kp} and compute the DFT in an economic fashion. It requires $N \log_2 N$ operations, which means that in terms of computing time and memory requirements, the FFT is far superior to the DFT. For $N = 50$, for example, the FFT requires about 250 complex operations compared to about 2500 complex operations required by the direct use of Eq. (4.83).

There exist several FFT algorithms and the basic idea behind all these is that a DFT of length N is *decimated* (or split) into successive smaller DFT's. One class of algorithms, called *radix*-2 algorithms, assume that N is a power of 2. The decimation is carried out in either the time domain or frequency domain. Accordingly, we have two types of algorithms in this class, namely, (a) decimation-in-time (DIT) and (b) Decimation-in-frequency (DIF). The Cooley–Tukey algorithm belongs to the type (a), whereas, the Sande–Tukey algorithm to the type (b). Both the algorithms require $N \log_2 N$ operations and the Cooley–Tukey algorithm is discussed in the next section.

4.6.4 Cooley–Tukey Algorithm

This algorithm assumes that N is an integral power of 2, i.e., $N = 2^m$, where m is an integer. The basic idea of this algorithm is to decompose the N-point DFT into two $N/2$-point DFT's then decompose each of the $N/2$-point DFT's into two $N/4$-point DFT's and continuing this process until we obtain $N/2$ two-point DFT's. It is clear that we require m steps to achieve this.

To describe the algorithm, we first consider the case $N = 4$, i.e., $m = 2$. Let f_0, f_1, f_2, f_3 be the sequence of values of $f(t)$. The DFT for f_k is given by

$$F_p = \sum_{k=0}^{3} f_k W_4^{pk}, \quad p = 0, 1, 2, 3. \tag{4.90}$$

where

$$W_4 = e^{-2\pi i/4} = -i$$

We split the sum on the right side of Eq. (4.90) into two equal parts of length 2, one containing the even-indexed values of $f(t)$ and the other, the odd-indexed values. We, therefore, write

$$F_p = \sum_{k=0,2} f_k W_4^{kp} + \sum_{k=1,3} f_k W_4^{kp} \tag{4.91}$$

Putting $k = 2r$ in the first sum and $k = 2r + 1$ in the second sum of Eq. (4.91), we obtain

$$F_p = \sum_{r=0}^{1} f_{2r} W_4^{2rp} + \sum_{r=0}^{1} f_{2r+1} W_4^{(2r+1)p}$$

$$= \sum_{r=0}^{1} f_{2r} W_2^{rp} + W_4^p \sum_{r=0}^{1} f_{2r+1} W_2^{rp}, \quad p = 0, 1$$

It may be seen that the first sum consists of f_0 and f_2 (even-indexed values) and the second sum f_1 and f_3 (odd-indexed values). A convenient notation is to use the superscripts e and o for the first and second sums, respectively*. We therefore write

$$F_p = F_p^e + W_4^p F_p^o, \quad p = 0, 1 \tag{4.92}$$

where

$$\left.\begin{aligned} F_p^e &= \sum_{r=0}^{1} f_{2r} W_2^{rp} \\ \text{and} \quad F_p^o &= \sum_{r=0}^{1} f_{2r+1} W_2^{rp} \end{aligned}\right\} p = 0, 1 \tag{4.93}$$

Since F_p is periodic with period 2, we have

$$F_{p+2} = F_p^e + W_4^{p+2} F_p^o, \quad p = 0, 1$$

$$= F_p^e - W_4^p F_p^o, \quad p = 0, 1 \tag{4.94}$$

* Numerical Recipes in FORTRAN, CUP, Indian edition [1994].

since

$$W_4^2 = -1.$$

From Eqs. (4.92) and (4.94), we obtain

$$\left.\begin{aligned} F_0 &= f_0 + f_2 + f_1 + f_3 \\ F_1 &= f_0 - f_2 - i\,(f_1 - f_3) \\ F_2 &= f_0 + f_2 - (f_1 + f_3) \\ F_3 &= f_0 - f_2 + i\,(f_1 - f_3) \end{aligned}\right\} \tag{4.95}$$

Results of Example 4.16 follow easily from Eqs. (4.95).

The computations involving Eqs. (4.92) and (4.94) of the 4-point DIT–FFT are shown in Fig. 4.3 called *flow-graph.*

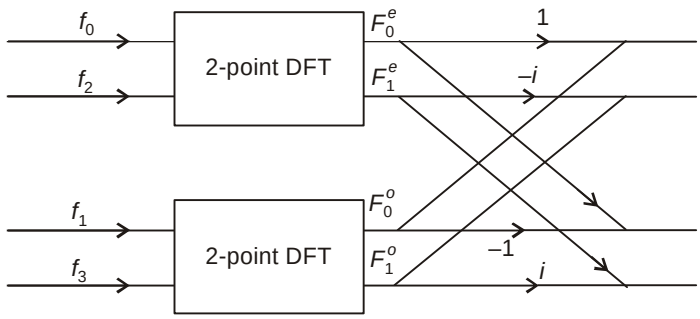


Figure 4.3 Flow-graph for DIT–FFT, $N = 4$.

We next consider the case $N = 8$. Let the sequence of values of $f(t)$ be

$$f_k = \{f_0, f_1, f_2, f_3, \ldots, f_7\}.$$

The DFT for f_k is

$$F_p = \sum_{k=0}^{7} f_k W_8^{pk}, \quad p = 0, 1, 2, \ldots, 7,$$

where

$$W_8 = e^{-2\pi i/8}$$

Splitting the 8-point DFT into two equal parts of length 4, one containing the even-indexed f_k and the other of the odd-indexed f_k, we write

$$F_p = F_p^e + W_8^p F_p^o \tag{4.96}$$

where

$$\left.\begin{aligned} F_p^e &= \sum_{r=0}^{3} f_{2r} W_4^{rp} \\ \text{and} \qquad F_p^o &= \sum_{r=0}^{3} f_{2r+1} W_4^{rp} \end{aligned}\right\} \; p = 0,\ 1,\ 2,\ 3 \tag{4.97}$$

Using the periodic properties, we also have

$$\begin{aligned} F_{p+4} &= F_p^e + W_8^{p+4} F_p^o, \qquad p = 0, 1, 2, 3 \\ &= F_p^e + W_8^p \cdot W_8^4 F_p^o \\ &= F_p^e - W_8^p F_p^o, \qquad p = 0, 1, 2, 3 \end{aligned} \tag{4.98}$$

since

$$W_8^4 = W_2^1 = -1.$$

It may be seen that F_p^e and F_p^o are both 4-point DFT's. This completes the first stage of the decimation process and the computations are shown in the flow-graph in Fig. 4.4.

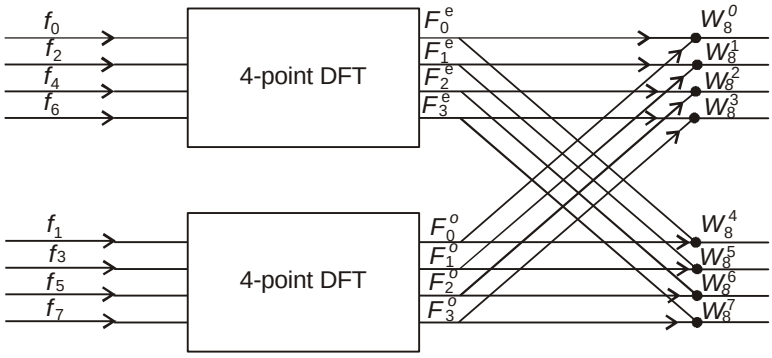


Figure 4.4 Flow-graph of first stage of 8-point DIT–FFT.

In Fig. 4.4, values of the factor W_8^{kp} are given below.

$$\left.\begin{aligned} &W_8^0 = 1, \quad W_8^1 = \frac{1-i}{\sqrt{2}}, \quad W_8^2 = -i, \quad W_8^3 = -\frac{1+i}{\sqrt{2}}, \\ &W_8^4 = -1, \quad W_8^5 = \frac{-1+i}{\sqrt{2}}, \quad W_8^6 = i, \quad W_8^7 = \frac{1+i}{\sqrt{2}} \end{aligned}\right\} \tag{4.99}$$

In the second stage of decimation, each of the 4-point DFT's in Fig. 4.4 is split into two 2-point DFT's. We then write

$$\begin{aligned} F_p^e &= \sum_{r=0}^{3} f_{2r} W_4^{pr} \\ &= \sum_{s=0}^{1} f_{4s} W_2^{sp} + W_4^p \sum_{s=0}^{1} f_{4s+2} W_2^{sp} \\ &= F_p^{ee} + W_4^p \cdot F_p^{eo}, \end{aligned} \tag{4.100}$$

where

$$\left.\begin{aligned} F_p^{ee} &= \sum_{s=0}^{1} f_{4s}\, W_2^{sp} \\ \text{and} \quad F_p^{eo} &= \sum_{s=0}^{1} f_{4s+2}\, W_2^{sp} \end{aligned}\right\} \tag{4.101}$$

Similarly, we obtain

$$F_p^{o} = F_p^{oe} + W_4^{p} F_p^{oo} \tag{4.102}$$

where

$$\left.\begin{aligned} F_p^{oe} &= \sum_{l=0}^{1} f_{4l+1}\, W_2^{lp} \\ \text{and} \quad F_p^{oo} &= \sum_{l=0}^{1} f_{4l+3}\, W_2^{lp} \end{aligned}\right\} \quad p = 0,\ 1 \tag{4.103}$$

Using the periodic properties, we also have

$$F_{p+2}^{e} = F_p^{ee} + W_4^{p+2} F_p^{eo}, \qquad p = 0, 1. \tag{4.104}$$

and

$$F_{p+2}^{o} = F_p^{oe} + W_4^{p+2} F_p^{oo}, \qquad p = 0, 1. \tag{4.105}$$

This completes the second stage of decimation where each of the 4-point transforms is broken into two 2-point transforms. The flow-graph of the second stage is shown in Fig. 4.5.

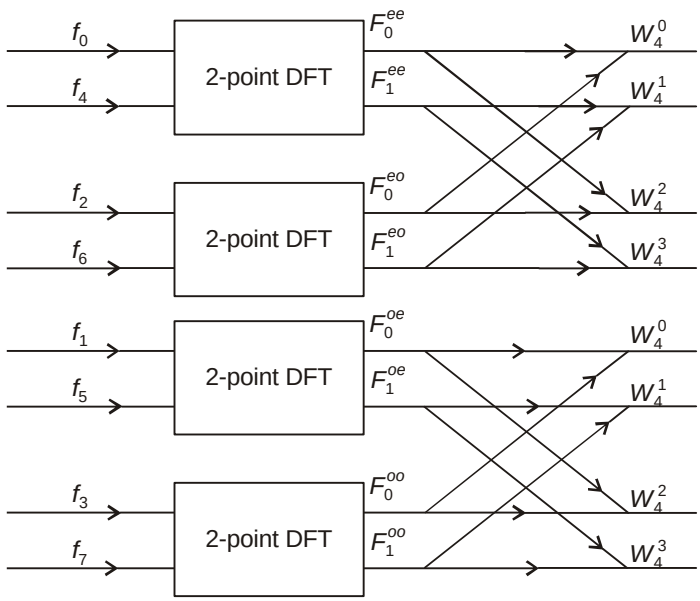


Figure 4.5 Second stage of the decomposition.

From Eq. (4.101), we find

$$F_p^{ee} = \sum_{s=0}^{1} f_{4s} W_2^{sp}, \qquad p = 0,1$$

$$= f_0 + f_4 W_2^p$$

and

$$F_p^{eo} = \sum_{s=0}^{1} f_{4s+2} W_2^{sp}, \qquad p = 0, 1$$

$$= f_2 + f_6 W_2^p, \qquad p = 0, 1$$

It follows that at the third stage of decimation, we obtain

$$F_p^{eee} = f_0, F_p^{eeo} = f_4, \; F_p^{eoe} = f_2 \text{ and } F_p^{eoo} = f_6.$$

Thus, for the 8-point DFT, we start with the input sequence f_0, f_4, f_2, f_6, f_1, f_5, f_3 and f_7, and obtain the output in the natural order, i.e., F_0, F_1, F_2, F_3, F_4, F_5, F_6 and F_7.

The three stages of computation can be shown in a single flow-graph (Fig. 4.6).

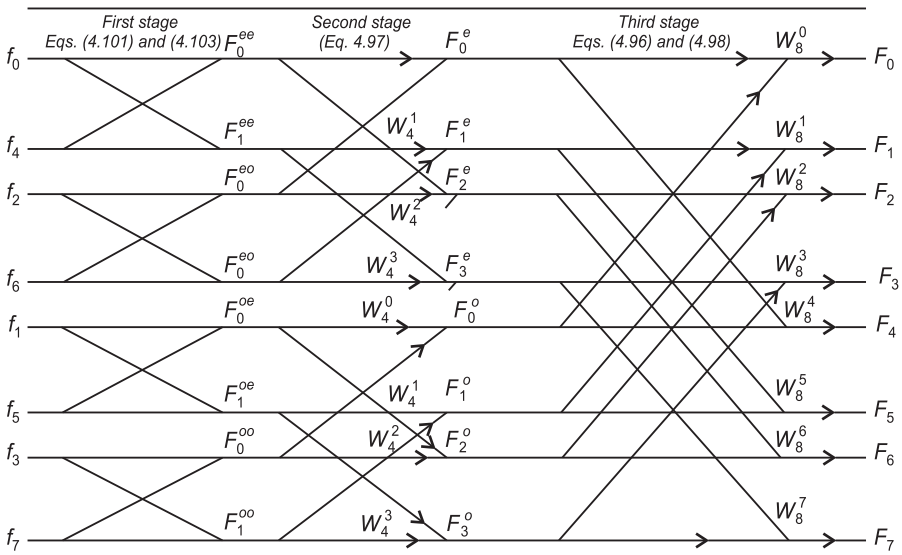


Figure 4.6 Flow-graph of an 8-point DIT–FFT.

From Fig. 4.6, the following observations can be made

(i) The input is in the *bit-reversed*. order, namely f_0, f_4, f_2, f_6, f_1, f_5, f_3, f_7, as shown in Table 4.2.

Table 4.2 Input Data in the Reversed Bits

Input position	*Binary digits*	*Reversed bits*	*Index of the sequence*
0	000	000	0
1	001	100	4
2	010	010	2
3	011	110	6
4	100	001	1
5	101	101	5
6	110	011	3
7	111	111	7

(ii) The output for the Fourier coefficients F_k is in the natural order.
(iii) Computations are carried out in terms of what is called a butterfly. A typical butterfly is shown in Fig. 4.7.

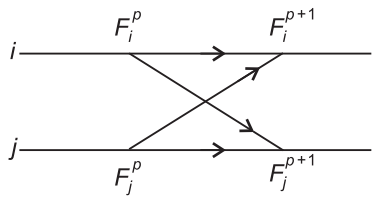


Figure 4.7 A typical butterfly.

Example 4.19 Apply Cooley–Tukey algorithm to compute the DFT of the sequence

$$f_k = \{1, 2, 3, 4\}.$$

The key equations are

$$F_p = F_p^e + W_4^p F_p^o \quad \text{and} \quad F_{p+2} = F_p^e - W_4^p F_p^o, \qquad p = 0, 1 \tag{i}$$

Also

$$\left.\begin{aligned} F_p^e &= f_0 + f_2 W_2^p, \quad p = 0, 1 \\ \text{and} \quad F_p^o &= f_1 + f_3 W_2^p, \quad p = 0, 1 \end{aligned}\right\} \tag{ii}$$

We, therefore, obtain

$$F_0^e = f_0 + f_2 = 4, \qquad F_1^e = f_0 - f_2 = -2$$

$$F_0^o = f_1 + f_3 = 6, \qquad F_1^o = f_1 - f_3 = -2$$

Equation (i) now gives

$$F_0 = 4 + 6 = 10, \qquad F_1 = -2 + W_4^1(-2) = -2 + 2i$$

$$F_2 = 4 - 6 = -2, \qquad F_3 = -2 - W_4^1(-2) = -2 - 2i$$

The flow-graph for this computation is shown in Fig. 4.8.

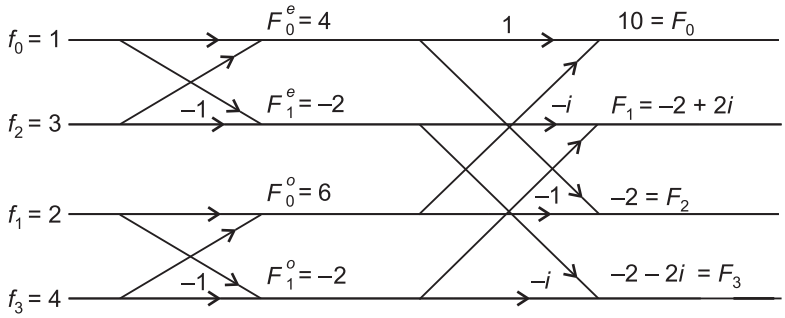


Figure 4.8 Flow-graph for Example 4.19.

Example 4.20 Use the Cooley–Tukey algorithm to find the DFT of the sequence

$$f_k = \{1, 2, 3, 4, 4, 3, 2, 1\}.$$

First stage: The key-equations are

$$F_p^{ee} = f_0 + W_2^p f_4, \quad F_p^{eo} = f_2 + W_2^p f_6$$

$$F_p^{oe} = f_1 + W_2^p f_5, \quad F_p^{oo} = f_3 + W_2^p f_7.$$

From these equations, we obtain

$$F_0^{ee} = f_0 + f_4 = 5, \quad F_1^{ee} = f_0 - f_4 = -3,$$

$$F_0^{eo} = f_2 + f_6 = 5, \quad F_1^{eo} = f_2 - f_6 = 1,$$

$$F_0^{oe} = f_1 + f_5 = 5, \quad F_1^{oe} = f_1 - f_5 = -1,$$

$$F_0^{oo} = f_3 + f_7 = 5, \quad F_1^{oo} = f_3 - f_7 = 3.$$

Second stage: The key-equations are

$$F_p^e = F_p^{ee} + W_4^p F_p^{eo}, \quad F_p^o = F_p^{oe} + W_4^p F_p^{oo}, \quad p = 0, 1$$

$$F_{p+2}^e = F_p^{ee} - W_4^p F_p^{eo}, \quad F_{p+2}^o = F_p^{oe} - W_4^p F_p^{oo}, \quad p = 0, 1$$

From these equations, we obtain

$$F_0^e = 5 + 5 = 10, \; F_1^e = -3 - i, \; F_2^e = 5 - 5 = 0, \; F_3^e = -3 + i,$$

$$F_0^o = 5 + 5 = 10, \; F_1^o = -1 - 3i, \; F_2^o = 0, \; F_3^o = -1 + 3i$$

For the third stage, we have

$$F_p = F_p^e + W_8^p F_p^o, \quad p = 0, 1, 2, 3$$

and

$$F_{p+4} = F_p^e \; ? \; W_8^p F_p^o, \quad p = 0, 1, 2, 3$$

where,

$$W_8^0 = 1,\quad W_8^1 = \frac{1-i}{\sqrt{2}},\quad W_8^2 = -i,\quad W_8^3 = -\frac{1+i}{\sqrt{2}},$$

$$W_8^4 = -1,\quad W_8^5 = -\frac{1-i}{\sqrt{2}},\quad W_8^6 = i,\quad W_8^7 = \frac{1+i}{\sqrt{2}}.$$

From these equations, we obtain

$$\begin{aligned} F_0 = F_0^e + F_0^o = 20,\quad F_1 &= F_1^e + W_8^1 F_1^o \\ &= -3 - i + \frac{1-i}{\sqrt{2}}(-1-3i) \\ &= -5.828 - i(2.414), \end{aligned}$$

$$\begin{aligned} F_2 = 0,\quad F_3 &= -3 + i - \frac{(1+i)}{\sqrt{2}}(-1+3i) \\ &= -0.172 - i(0.414), \end{aligned}$$

$$\begin{aligned} F_4 = 0,\quad F_5 &= -3 - i - \frac{(1-i)}{\sqrt{2}}(-1-3i) \\ &= -0.172 + i(0.414) \end{aligned}$$

$$\begin{aligned} F_6 = 0,\quad F_7 &= -3 + i + \frac{(1+i)}{\sqrt{2}}(-1+3i) \\ &= -5.828 + i(2.414). \end{aligned}$$

The flow-graph for this computation is given in Fig. 4.9.

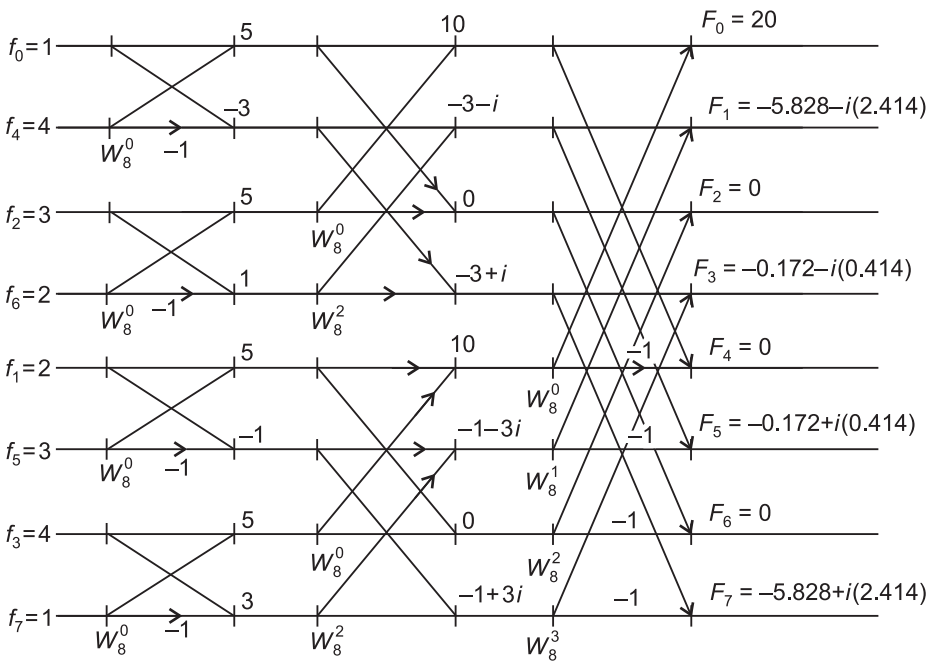


Figure 4.9 Flow-graph for Example 4.20.

4.6.5 Sande–Tukey Algorithm (DIF–FFT)

This alternative approach is a member of the class of algorithms called *decimation-in-frequency* techniques. It is the reverse of the Cooley–Tukey algorithm described in the previous section. In this case, the output is in the *bit-reversed order.*

To start with we take $N = 4$, i.e.,

$$f_k = \{f_0, f_1, f_2, f_3\}.$$

Then we have

$$F_p = \sum_{k=0}^{3} f_k W_4^{pk}, \qquad p = 0, 1, 2, 3$$

In the method, the above sum is divided in terms of the first two and last two points as:

$$\begin{aligned} F_p &= \sum_{k=0,1} f_k W_4^{pk} + \sum_{k=2,3} f_k W_4^{pk} \\ &= \sum_{k=0}^{1} f_k W_4^{pk} + \sum_{k=0}^{1} f_{k+2} W_4^{p(k+2)}; \qquad p = 0, 1, 2, 3 \end{aligned} \tag{4.106}$$

Now,

$$W_4^{p(k+2)} = W_4^{pk} \cdot W_4^{2p} = W_4^{pk}(-1)^p,$$

since

$$W_4^{2p} = W_2^p = (-1)^p.$$

Hence, Eq. (4.106) becomes

$$F_p = \sum_{k=0}^{1} \left[f_k + (-1)^p \cdot f_{k+2} \right] W_4^{pk}, \quad p = 0, 1, 2, 3 \tag{4.107}$$

which consists of both even and odd components.
Let

$$F_p = F_{2r} + F_{2r+1},$$

where

$$\left.\begin{aligned} F_{2r} &= \sum_{k=0}^{1} (f_k + f_{k+2}) W_4^{2rk} \\ &= \sum_{k=0}^{1} (f_k + f_{k+2}) W_2^{rk} \\ \text{and} \quad F_{2r+1} &= \sum_{k=0}^{1} (f_k - f_{k+2}) W_4^{k(2r+1)} \\ &= \sum_{k=0}^{1} (f_k - f_{k+2}) W_4^{2kr} \cdot W_4^k \\ &= \sum_{k=0}^{1} (f_k - f_{k+2}) W_2^{kr} \cdot W_4^k, \quad r = 0, 1 \end{aligned}\right\} \tag{4.108}$$

We define

$$\left.\begin{aligned} g_k &= f_k + f_{k+2} \\ \text{and} \quad h_k &= (f_k - f_{k+2})W_4^k \end{aligned}\right\} \tag{4.109}$$

Then

$$\left.\begin{aligned} F_{2r} &= \sum_{k=0}^{1} g_k W_2^{rk}, \quad r = 0, 1 \\ \text{and} \quad F_{2r+1} &= \sum_{k=0}^{1} h_k W_2^{rk}, \quad r = 0, 1 \end{aligned}\right\} \tag{4.110}$$

It can be seen that the output is in the order F_0, F_2, F_1 and F_3. We now consider the case $N = 8$. Let

$$f_k = \{f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7\}.$$

Then

$$F_p = \sum_{k=0}^{7} f_k W_8^{pk}, \quad p = 0, 1, \ldots, 7.$$

We split the above sum in terms of the first four and last four points as

$$\begin{aligned} F_p &= \sum_{k=0,1,2,3} f_k W_8^{pk} + \sum_{k=4,5,6,7} f_k W_8^{pk} \\ &= \sum_{k=0,1,2,3} f_k W_8^{pk} + \sum_{k=0,1,2,3} f_{k+4} W_8^{p(k+4)} \end{aligned} \tag{4.111}$$

Now,

$$W_8^{p(k+4)} = W_8^{pk} \cdot W_8^{4p} = (-1)^p W_8^{pk}$$

Hence Eq. (4.111) becomes

$$F_p = \sum_{k=0}^{3} \left[f_k + (-1)^p f_{k+4} \right] W_8^{pk}, \quad p = 0, 1, \ldots, 7. \tag{4.112}$$

Since the right side of Eq. (4.112) consists of both even and odd Fourier coefficients, we write

$$F_p = F_{2r} + F_{2r+1},$$

where

$$\left.\begin{aligned} F_{2r} &= \sum_{k=0}^{3}[f_k + f_{k+4}]W_8^{2rk} \\ &= \sum_{k=0}^{3}[f_k + f_{k+4}]W_4^{rk}, \quad r = 0, 1, 2, 3 \\ \text{and } F_{2r+1} &= \sum_{k=0}^{3}[f_k - f_{k+4}]W_8^{k(2r+1)}, \\ &= \sum_{k=0}^{3}[f_k - f_{k+4}]W_4^{rk} \cdot W_8^{k}, \quad r = 0, 1, 2, 3 \end{aligned}\right\} \tag{4.113}$$

For brevity, let

$$\left.\begin{aligned} f_k + f_{k+4} &= g_k \\ \text{and } W_8^k(f_k - f_{k+4}) &= h_k \end{aligned}\right\}, \quad k = 0, 1, 2, 3 \tag{4.114}$$

Then Eq. (4.113) becomes

$$\left.\begin{aligned} F_{2r} &= \sum_{k=0}^{3} g_k W_4^{rk} \\ \text{and } F_{2r+1} &= \sum_{k=0}^{3} h_k W_4^{rk} \end{aligned}\right\}, \quad r = 0, 1, 2, 3 \tag{4.115}$$

The flow-graph for the first stage of this algorithm is given in Fig. 4.10.

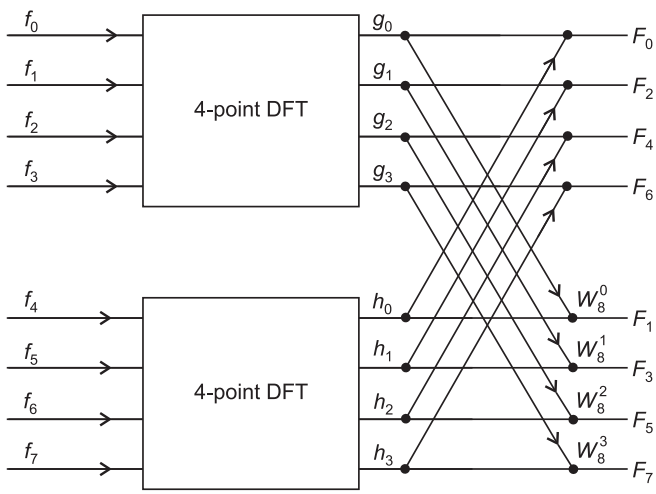


Figure 4.10 Flow-graph for the first stage of Sande–Tukey algorithm.

Clearly, this approach can be repeated at the second stage to split each of the 4-point DFT's into two 2-point transforms. Flow-graph for this computation is shown in Fig. 4.11.

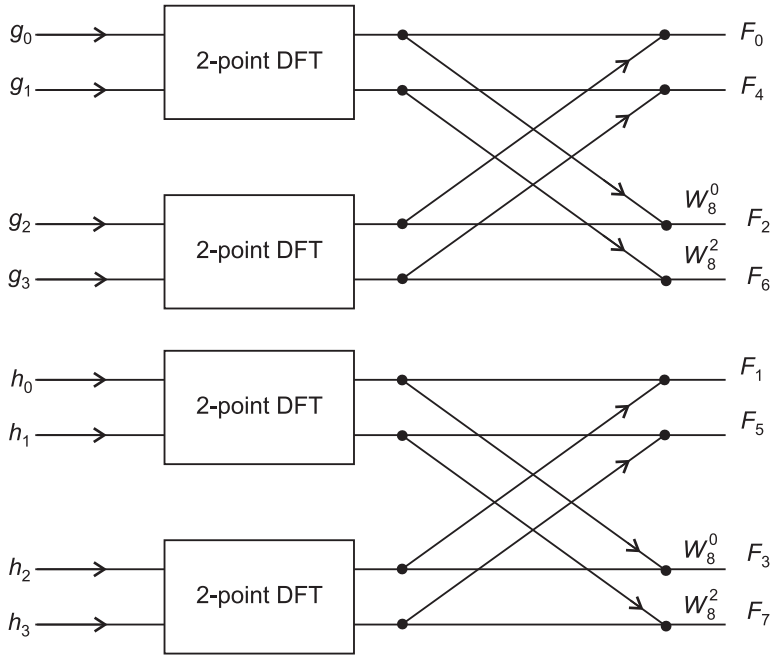


Figure 4.11 Second stage for Sande–Tukey algorithm.

In the general case, the final result is obtained after $\log_2 N$ stages. Figure 4.12 shows the flow-graph for the 8-point decimation in frequency FFT.

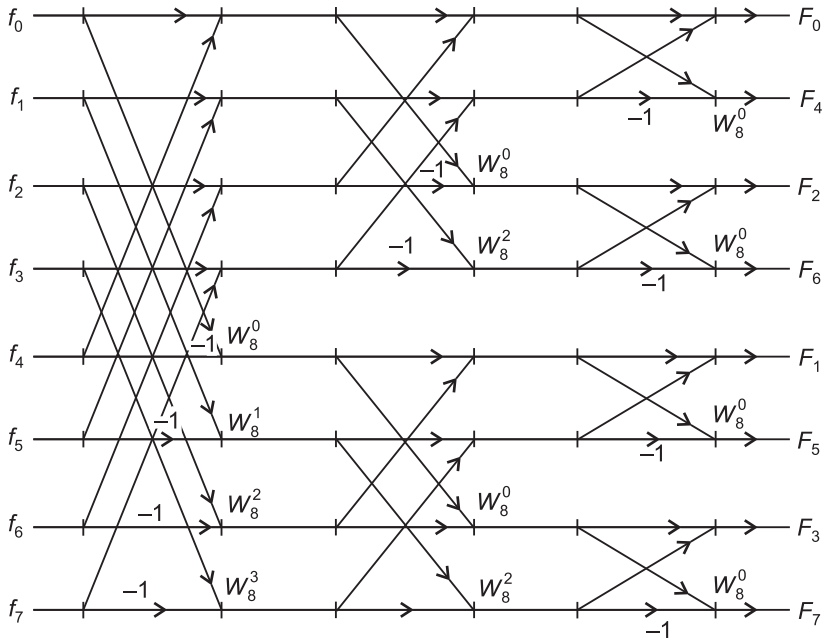


Figure 4.12 8-point flow-graph for Sande–Tukey algorithm.

We observe that the input is in the natural order, whereas the output for the frequency components is in the bit-reversed order.

4.6.6 Computation of the Inverse DFT

The inverse DFT is defined by

$$f_k = \frac{1}{N}\sum_{p=0}^{N-1} F_p W_N^{-kp}, \qquad k = 0, 1, \ldots, N-1.$$

Comparison with DFT shows that the factors W_N^{kp} have changed signs, the input and output have interchanged, and that the final output is divided by N. Hence the flow-graph for the calculation of DFT can also be adopted for the computation of inverse DFT after making the above changes.

Example 4.21 Using Sande–Tukey algorithm, find the DFT of the sequence

$$f_k = \{1,\ 2,\ 3,\ 4\}.$$

We have the key-equations

$$\left.\begin{aligned} F_{2r} &= \sum_{k=0}^{1} g_k W_2^{rk} \\ \text{and} \quad F_{2r+1} &= \sum_{k=0}^{1} h_k W_2^{rk} \end{aligned}\right\} \quad r = 0,\ 1$$

where

$$g_k = f_k + f_{k+2} \text{ and } h_k = (f_k - f_{k+2})\ W_4^k$$

With $f_k = \{1,\ 2,\ 3,\ 4\}$, we obtain

$$g_0 = f_0 + f_2 = 4, \quad g_1 = f_1 + f_3 = 6;$$

$$h_0 = f_0 - f_2 = -2, \quad h_1 = (f_1 - f_3)\ W_4^1 = -2(-i) = 2i$$

Hence

$$F_0 = g_0 + g_1 = 10, \qquad F_2 = g_0 + g_1 W_2^1 = g_0 - g_1 = -2$$

$$F_1 = \sum_{k=0}^{1} h_k W_2^0 = h_0 + h_1 = -2 + 2i,$$

$$F_3 = \sum_{k=0}^{1} h_k W_2^k = h_0 + h_1 W_2^1 = h_0 - h_1 = -2 - 2i.$$

Hence

$$F_k = \{10, -2 + 2i, -2, -2 - 2i\}.$$

The flow-graph for this computation is shown in Fig. 4.13.

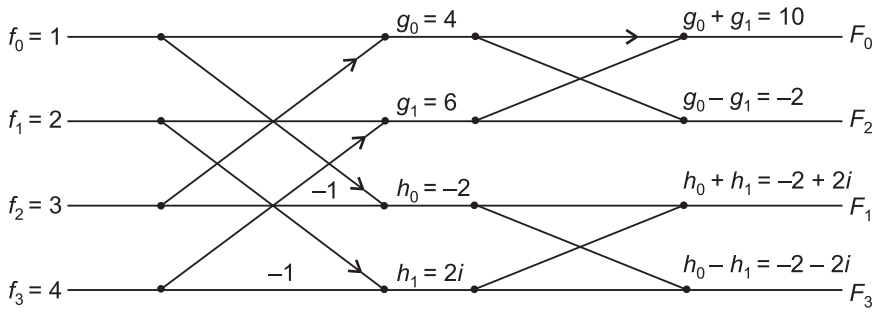


Figure 4.13 Flow-graph for Example 4.21.

Example 4.22 Using Sande–Tukey algorithm, find the DFT of the sequence

$$f_k = \{1,\ 2,\ 3,\ 4,\ 4,\ 3,\ 2,\ 1\}$$

We have

$$F_p = \sum_{k=0}^{7} f_k W_8^{pk}, \qquad p = 0, 1, \ldots, 7$$

First stage:

$$F_{2r} = \sum_{k=0}^{3} g_k W_4^{rk},\ F_{2r+1} = \sum_{k=0}^{3} h_k W_4^{rk}, \qquad r = 0, 1, 2, 3.$$

$$g_k = f_k + f_{k+4},\ \ h_k = (f_k - f_{k+4}) W_8^k$$

Then

$$g_0 = 5,\ \ g_1 = 5,\ \ g_2 = 5,\ \ g_3 = 5,$$

$$h_0 = -3,\ \ h_1 = -W_8^1,\ \ h_2 = W_8^2,\ \ h_3 = 3W_8^3.$$

Second stage:

$$F_{2r} = F_{4s} + F_{4s+2},\ \ F_{2r+1} = F_{4t+1} + F_{4t+3},$$

$$F_{4s} = \sum_{k=0}^{1} p_k \cdot W_2^{sk},\ \ F_{4s+2} = \sum_{k=0}^{1} q_k \cdot W_2^{sk}, \quad s = 0, 1.$$

$$p_k = g_k + g_{k+2},\ q_k = (g_k - g_{k+2}) W_4^k, \quad k = 0, 1.$$

$$F_{4t+1} = \sum_{k=0}^{1} u_k \cdot W_2^{sk}, \quad F_{4t+3} = \sum_{k=0}^{1} v_k \cdot W_2^{sk},$$

$$u_k = h_k + h_{k+2}, \qquad v_k = (h_k - h_{k+2})\ W_4^k$$

We obtain,

$$p_0 = 10,\ p_1 = 10,\ q_0 = 0,\ \ q_1 = 0;$$

$$u_0 = -3 + W_8^2 = -3 - i, \qquad u_1 = h_1 + h_3 = -2\sqrt{2} - i\sqrt{2},$$

$$v_0 = -3 + i, \qquad v_1 = 2\sqrt{2} - i\sqrt{2}.$$

Finally,

$$F_0 = p_0 + p_1 = 20, \quad F_4 = p_0 - p_1 = 0,$$
$$F_2 = q_0 + q_1 = 0, \quad F_6 = q_0 - q_1 = 0,$$
$$F_1 = -3 - i - 2\sqrt{2} - i\sqrt{2} = -5.828 - i(2.414)$$
$$F_5 = -3 - i + 2\sqrt{2} + i\sqrt{2} = -0.172 + i(0.414)$$
$$F_3 = v_0 + v_1 = -0.172 - i(0.414)$$
$$F_7 = v_0 - v_1 = -5.828 + i(2.414)$$

EXERCISES

4.1 Explain the method of least squares to fit a straight line of the form $Y = a_0 + a_1x$ to the data (x_i, y_i):

x	1	2	3	4	5	6
y	2.4	3.1	3.5	4.2	5.0	6.0

4.2 Find the values of a_0 and a_1 so that $Y = a_0 + a_1x$ fits the data given in the table:

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

4.3 If the straight line $Y = a_0 + a_1x$ is the best fit to the set of data points (x_1, y_1), (x_2, y_2), …, (x_n, y_n), show that

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix}, \quad i = 1, 2, \ldots, n.$$

4.4 Use the method of least squares to fit the straight line $Y = a + bx$ to the data:

x	0	1	2	3
y	2	5	8	11
w	1	1	1	1

4.5 Find the values of a, b, c so that $Y = a + bx + cx^2$ is the best fit to the data:

x	0	1	2	3	4
y	1	0	3	10	21

4.6 Fit a least squares parabola $Y = a + bx + cx^2$ to the data:

x	0	1	2	3	4	5	6
y	71	89	67	43	31	18	9

4.7 Determine the normal equations if the cubic polynomial $Y = a_0 + a_1x + a_2x^2 + a_3x^3$ is fitted to the data (x_i, y_i), $i = 1, 2, \ldots, m$.

4.8 Determine the constants a and b, by the method of least squares, such that the curve $y = ae^{bx}$ fits the data:

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

4.9 Fit a function of the form $y = ax^b$ for the following data:

x	61	26	7	2.6
y	350	400	500	600

4.10 Using the method of least squares, fit a curve of the form $Y = \dfrac{a}{x} + bx$ to the following data (x, y):

(1, 5.43), (2, 6.28), (4, 10.32), (6, 14.86), (8, 19.51).

4.11 Fit a function of the form $y = A_1e^{\lambda_1 x} + A_2e^{\lambda_2 x}$ to the data given by

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
y	1.54	1.67	1.81	1.97	2.15	2.35	2.58	2.83	3.11

4.12 Write an algorithm to fit a linear least squares approximation to the data (x_i, y_i), $i = 1, 2, \ldots, N$.

4.13 If the functions $f_1(x) = 1$, $f_2(x) = x$ are orthogonal on the interval $[-1, 1]$, find the values of a and b so that the function $f_3 = 1 + ax + bx^2$ is orthogonal to both f_1 and f_2 on $[-1, 1]$.

4.14 Define an orthogonal set of functions and show that the set

$$f(x) = \sin \frac{n\pi x}{l}, \quad n = 1, 2, \ldots$$

is orthogonal on $[0, l]$.

4.15 Explain the difference between Fourier series and Fourier transform. Find the Fourier series for the function defined by

$$f(x) = \begin{cases} x, & -1 < x \le 0 \\ x + 2, & 0 < x \le 1 \end{cases}$$

4.16 Define discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT).

Find the DFT of each of the following sequences using matrix method (Problems 4.17–4.20):

4.17 {0, 1, 0, –1}

4.18 $\left\{1, \frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}\right\}$

4.19 {1, 2, 3, 1}

4.20 {2, 2, 4, 3}

Using the definition

$$F_p = \sum_{k=0}^{N-1} f_p e^{-2\pi ikp/N}, \quad p = 0, 1, 2, \ldots, N-1,$$

find the DFT of each of the following sequences (Problems 4.21– 4.23):

4.21 {1, 2, 3, 4}

4.22 {1, 1, …, 1}, N values.

4.23 {1, –1, 2, –2, 3, –3}

4.24 Write the computational steps for computing the DFT of a sequence $f(t)$ by the above formula and test it on Problem (4.21).

4.25 List any two properties of DFT and find the IDFT of the sequence {0, 1, 1, 0}.

4.26 State the key equations in the Cooley–Tukey algorithm for computing the DFT of the sequence f_p, p = 0, 1, 2, 3. Draw the flow-graph for its computation.

Use the Cooley–Tukey algorithum to compute the DFT of each of the following sequences (Problems 4.27–4.30):

4.27 {1, 1, 0, 0}

4.28 {1, 0, 1, 0}

4.29 {1, –1, 1, –1}

4.30 {1, 2, 1, 2}

4.31 Draw a flow-chart to implement the bit reversal procedure for the Cooley–Tukey algorithm.

4.32 Write down the key equations in the Cooley–Tukey algorithm for computing the 8-point DFT of the sequence f_p, p = 0, 1, 2, …, 7, and draw the flow-graph for its computation.

Use the Cooley–Tukey algorithm to compute the DFT of each of the following sequences (Problems 4.33–4.35):

4.33 {1, –1, 1, –1, 1, –1, 1, –1}

4.34 {1, 1, 1, 1, 1, 1, 1, 1}

4.35 {1, 2, 3, 4, 5, 6, 7, 8}

4.36 Explain the difference between Cooley–Tukey and Sande–Tukey algorithms for an 8-point computation of the DFT of a sequence. Write down the key equations of Sande–Tukey algorithm for computing the DFT of the sequence f_k, $k = 0, 1, 2, \ldots, 7$.

Use Sande–Tukey algorithm to compute the DFT of each of the following sequences (Problems 4.37–4.40):

4.37 {1, 1, 0, 0}

4.38 {1, 0, 1, 0}

4.39 {1, 2, 3, 4, 5, 6, 7, 8}

4.40 {0, 1, 0, 1, 0, 1, 0, 1}

4.41 Show that $T_n(x) = \cos(n \cos^{-1}x)$ is a polynomial in x of degree n.

4.42 Show that the coefficient of x^n in $T_n(x)$ is 2^{n-1}.

4.43 Economize the series given by

$$\sinh x = x + \frac{x^3}{6} + \frac{x^5}{120} + \frac{x^7}{5040} + \cdots$$

on the interval [–1, 1], allowing for a tolerance of 0.0005.

Answers to Exercises

4.1 $a_1 = 0.503$, $a_0 = 2.021$.

4.2 $a_1 = 2.0$, $a_0 = 0.8$.

4.4 $a = 2$, $b = 3$.

4.5 $a = 1$, $b = -3$ and $c = 2$.

4.6 $a = 81.93$, $b = -8.28$, $c = -0.78$

4.8 $a = 1.5$, $b = 0.5 = a_1$

4.9 $a = 702$, $b = -0.17$.

4.10 $a = 3.02$, $b = 2.39$.

4.11 $\lambda_1 = 0.99$, $\lambda_2 = -0.96$, $A_1 = 0.499$, $A_2 = 0.491$.

4.13 $1 - 3x^2$.

4.15 $a_0 = 2$, $a_n = 0$, $b_n = \dfrac{2}{n\pi}[1 - 2(-1)^n]$.

4.17 $[0, -2i, 0, 2i]$

4.18 $[1, 1 - i\sqrt{2}, 1, 1 + i\sqrt{2}]$

4.19 $[7, -2 - i, 1, -2 + i]$

4.20 $[11, -2 + i, 1, -2 - i]$

4.21 $F_0 = 10$, $F_1 = -2 + 2i$, $F_2 = -2$, $F_3 = -2 - 2i$.

4.22 $F_p = \begin{cases} N, & p = 0 \\ 0, & p = 1, 2, \ldots, N-1 \end{cases}$

4.23 $\left[-1.5 - \frac{\sqrt{3}}{2}i, -1.5 - \frac{3\sqrt{3}}{2}i, 12, -1.5 + \frac{3\sqrt{3}}{2}i \; -1.5 + \frac{\sqrt{3}}{2}i\right]$

4.24 $[10, -2 + 2i, -2, -2 - 2i]$

4.25 $f_p = [0.5, -0.25 + 0.25i, 0, -0.25 - 0.25i]$

4.26 $F_0^e = f_0 + f_2$, $F_1^e = f_0 - f_2$, $F_0^o = f_1 + f_3$, $F_1^o = f_1 - f_3$.

4.27 $F_0 = 2$, $F_1 = 1 - i$, $F_2 = 0$, $F_3 = 1 + i$.

4.28 $f_k = \{1, 0, 1, 0\}$, $F_p = \{2, 0, 2, 0\}$

4.29 $f_k = \{1, -1, 1, -1\}$, $F_p = \{0, 0, 4, 0\}$

4.30 $f_k = \{1, 2, 1, 2\}$, $F_p = \{6, 0, 0, 0\}$

4.33 Final stage: $F_0 = 0$, $F_1 = 0$, $F_2 = 0$, $F_3 = 0$; $F_4 = 8$, $F_5 = 0$, $F_6 = 0$, $F_7 = 0$.

4.34 $f_k = \{1, 1, 1, 1, 1, 1, 1, 1\}$, $F_p = \{8, 0, 0, 0, 0, 0, 0, 0\}$.

4.35 $36, -4 + i(9.66), -4 + 4i, -4 + i(1.66), 4, -4 - i(1.66), -4 - 4i, -4 - i(9.66)$.

4.37 $F_p = \{2, 1 - i, 0, 1 + i\}$

4.38 $F_0 = 2$, $F_2 = 2$, $F_1 = 0$, $F_3 = 0$.

4.39 $F_p = \{36, -4 + 9.66i, -4 + 4i, -4 + 1.66i, -4, -4 - 1.66i, -4 - 4i, -4 - 9.66i\}$

4.40 $F_p = \{4, 0, 0, 0, -4, 0, 0, 0\}$

4.41 Hint: Use mathematical induction

4.42 $T_1(x) = 2^{1-1}x$, $T_2(x) = 2^{2-1}x^2 - 1$,
$T_3(x) = 2^{3-1}x^3 - 2x$, $T_4(x) = 2^{4-1}x^4 - 8x^2 + 1$.

4.43 $\sinh x = \frac{383}{384}x + \frac{17}{96}x^3$.

Chapter 5

Spline Functions

5.1 INTRODUCTION

In Chapter 3, we have discussed the methods of finding an nth order polynomial passing through $(n + 1)$ given data points. In certain cases, these polynomials tend to give erroneous results due to round-off and other errors. Further, it was found that a low-order polynomial approximation in each *subinterval* provides a better approximation to the tabulated function than fitting a single high-order polynomial to the entire interval. Such an interpolation is called *piecewise polynomial interpolation* and the *spline functions* are such piecewise connecting polynomials.

The name *spline* has been adopted following the draftsman's device of using a thin flexible strip (called a *spline*) to draw a smooth curve through given points. The points at which two connecting splines meet are called *knots.* The connecting polynomials could be of any degree and, therefore, we have different types of splines: linear, quadratic, cubic, quintic, etc. Of these, the cubic spline (spline of degree three, or order four) has been found to be the most popular in engineering applications. Before discussing about cubic splines, we shall start with linear and quadratic splines since such a discussion will eventually justify the development of cubic splines. In Section 5.2, we derive the governing equations of a cubic spline and consider different end conditions. Errors in the cubic spline and cubic spline derivatives are important from the stand point of applications. These derivatives will be discussed in Section 5.2. Surface fitting cubic splines will be considered in Section 5.3, while cubic B-splines and their computation will be introduced in Section 5.4.

Applications of cubic splines to numerical differentiation, integration, numerical solution of differential equations, etc. will be considered in subsequent chapters.

5.1.1 Linear Splines

Let the given data points be

$$(x_i, y_i), \qquad i = 0, 1, 2, \ldots, n, \tag{5.1}$$

where

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

and let

$$h_i = x_i - x_{i-1}, \qquad i = 1, 2, \ldots, n. \tag{5.2}$$

Further, let $s_i(x)$ be the spline of degree one defined in the interval $[x_{i-1}, x_i]$. Obviously, $s_i(x)$ represents a straight line joining the points (x_{i-1}, y_{i-1}) and (x_i, y_i). Hence, we write

$$s_i(x) = y_{i-1} + m_i(x - x_{i-1}), \tag{5.3}$$

where

$$m_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}. \tag{5.4}$$

Setting $i = 1, 2, \ldots, n$ successively in Eq. (5.3), we obtain different splines of degree one valid in the subintervals 1 to n, respectively. It is easily seen that $s_i(x)$ is continuous at both the end points.

Example 5.1 Given the set of data points (1, –8) (2, –1) and (3,18) satisfying the function $y = f(x)$, find the linear splines satisfying the given data. Determine the approximate values of $y(2.5)$ and $y'(2.0)$.

Let the given points be $A(1, -8)$, $B(2, -1)$ and $C(3, 18)$. Equation of AB is

$$s_1(x) = -8 + (x - 1)7 = 7x - 15,$$

and equation of BC is

$$s_2(x) = -1 + (x - 2)19 = 19x - 39,$$

Since $x = 2.5$ belongs to the interval [2, 3], we have

$$y(2.5) \approx s_2(2.5) = 19(2.5) - 39 = 8.5,$$

and

$$y'(2.0) \approx m_1 = 19.$$

It is easy to check that the splines $s_i(x)$ are continuous in $[1, 3]$ but their slopes are discontinuous. This is clearly a *drawback of linear splines* and therefore we next discuss quadratic splines which assume the continuity of the slopes in addition to that of the function.

5.1.2 Quadratic Splines

With reference to the data points given in Eq. (5.1), let $s_i(x)$ be the quadratic spline approximating the function $y = f(x)$ in the interval $[x_{i-1}, x_i]$, where $x_i - x_{i-1} = h_i$. Let $s_i(x)$ and $s_i'(x)$ be continuous in $[x_0, x_n]$ and let

$$s_i(x_i) = y_i, \qquad i = 0, 1, 2, \ldots, n. \tag{5.5}$$

Since $s_i(x)$ is a quadratic in $[x_{i-1}, x_i]$, it follows that $s_i'(x)$ is a linear function and therefore we write

$$s_i'(x) = \frac{1}{h_i}\left[(x_i - x)m_{i-1} + (x - x_{i-1})m_i\right], \tag{5.6}$$

where

$$m_i = s_i'(x_i). \tag{5.7}$$

Integrating Eq. (5.6) with respect to x, we obtain

$$s_i(x) = \frac{1}{h_i}\left[-\frac{(x_i - x)^2}{2}m_{i-1} + \frac{(x - x_{i-1})^2}{2}m_i\right] + c_i, \tag{5.8}$$

where c_i are constants to be determined. Putting $x = x_{i-1}$ in Eq. (5.8), we get

$$c_i = y_{i-1} + \frac{1}{h_i}\frac{h_i^2}{2}m_{i-1} = y_{i-1} + \frac{h_i}{2}m_{i-1}.$$

Hence Eq. (5.8) becomes:

$$s_i(x) = \frac{1}{h_i}\left[-\frac{(x_i - x)^2}{2}m_{i-1} + \frac{(x - x_{i-1})^2}{2}m_i\right] + y_{i-1} + \frac{h_i}{2}m_{i-1}. \tag{5.9}$$

In Eq. (5.9), the m_i are still unknown. To determine the m_i, we use the condition of continuity of the function since the first derivatives are already continuous. For the continuity of the function $s_i(x)$ at $x = x_i$, we must have

$$s_i(x_i-) = s_{i+1}(x_i+) \tag{5.10}$$

From Eq. (5.9), we obtain

$$\begin{aligned} s_i(x_i-) &= \frac{h_i}{2}m_i + y_{i-1} + \frac{h_i}{2}m_{i-1} \\ &= \frac{h_i}{2}(m_{i-1} + m_i) + y_{i-1}. \end{aligned} \tag{5.11}$$

Further,

$$s_{i+1}(x) = \frac{1}{h_{i+1}}\left[-\frac{(x_{i+1} - x)^2}{2}m_i + \frac{(x - x_i)^2}{2}m_{i+1}\right] + y_i + \frac{h_{i+1}}{2}m_i,$$

and therefore

$$s_{i+1}(x_i+) = -\frac{h_{i+1}}{2}m_i + y_i + \frac{h_{i+1}}{2}m_i = y_i. \tag{5.12}$$

Equality of Eqs. (5.11) and (5.12) produces the recurrence relation

$$m_{i-1} + m_i = \frac{2}{h_i}(y_i - y_{i-1}), \qquad i = 1, 2, \ldots, n \tag{5.13}$$

for the spline first derivatives m_i. Equations (5.13) constitute n equations in $(n + 1)$ unknowns, viz, $m_0, m_1, \ldots, m_n$. Hence, we require one more condition to determine the m_i uniquely. There are several ways of choosing this condition. One natural way is to choose $s_1''(x_1) = 0$, since the mechanical spline straightens out in the end intervals. Such a spline is called a *natural spline.* Differentiating Eq. (5.9) twice with respect to x, we obtain

$$s_i''(x) = \frac{1}{h_i}(-m_{i-1} + m_i),$$

or

$$s_1''(x_1) = \frac{1}{h_1}(m_1 - m_0).$$

Hence, we have the additional condition as

$$m_0 = m_1. \tag{5.14}$$

Therefore, Eqs. (5.13) and (5.14) can be solved for m_i, which when substituted in Eq. (5.9) gives the required quadratic spline.

Example 5.2 Determine the quadratic splines satisfying the data given in Example 5.1. Find also approximate values of $y(2.5)$ and $y'(2.0)$.

We have $n = 2$ and $h = 1$. Equation (5.13) gives

$$m_0 + m_1 = 14 \quad \text{and} \quad m_1 + m_2 = 38.$$

Since $m_0 = m_1$, we obtain $m_0 = m_1 = 7$, and $m_2 = 31$.
Hence, Eq. (5.9) gives:

$$s_2(x) = -\frac{(x_2 - x)^2}{2}(7) + \frac{(x - x_1)^2}{2}(31) - 1 + \frac{7}{2}$$

$$= -\frac{(3 - x)^2}{2}(7) + \frac{31}{2}(x - 2)^2 + \frac{5}{2}$$

$$= 12x^2 - 41x + 33,$$

which is the spline in the interval [2, 3].
Hence,

$$y(2.5) \approx s_2(2.5) = 5.5 \quad \text{and} \quad y'(2.0) \approx 7.0.$$

The quadratic spline $s_1(x)$ in the interval [1, 2] can be determined in a similar way. A straightforward way of deriving the quadratic splines is as follows:

Since $s_i(x)$ is a quadratic in (x_{i-1}, x_i), we can write

$$s_i(x) = a_i + b_i x + c_i x^2, \tag{5.15}$$

where a_i, b_i and c_i are constants to be determined. Clearly, there are $3n$ constants and therefore we require $3n$ conditions to determine them. These conditions are obtained by using the properties of the quadratic spline. Firstly, we use the condition that the spline passes through the interior points. This means

$$s_i(x_i) = a_i + b_i x_i + c_i x_i^2 \qquad i = 1, 2, \ldots, n-1. \tag{5.16}$$

Next, $s_i(x)$ is continuous at $x = x_i$. This condition requires

$$s_i(x_i-) = s_{i+1}(x_i+). \tag{5.17}$$

Hence, we must have

$$a_i + b_i x_i + c_i x_i^2 = a_{i+1} + b_{i+1} x_i + c_{i+1} x_i^2, \quad i = 1, 2, \ldots, n-1. \tag{5.18}$$

Again, $s'_i(x)$ is continuous at $x = x_i$. This gives

$$b_i + 2c_i x_i = b_{i+1} + 2c_{i+1} x_i, \quad i = 1, 2, \ldots, n-1. \tag{5.19}$$

We thus have $3n-3$ conditions and we require three more conditions. Since the spline passes through the end points also, we must have

$$y_0 = a_1 + b_1 x_0 + c_1 x_0^2 \tag{5.20}$$

and

$$y_n = a_n + b_n x_n + c_n x_n^2. \tag{5.21}$$

Finally, for the natural spline, we have

$$s''_1(x_0) = 0, \tag{5.22}$$

and this gives

$$c_1 = 0. \tag{5.23}$$

We have, thus, a completed system of $3n$ equations in $3n$ unknowns. Although this system can certainly be solved, it is obviously more expensive and, therefore, this method is less preferred to the previous one.

The discontinuity in the second derivatives is an obvious disadvantage of the quadratic splines and this drawback is removed in the cubic splines discussed below.

5.2 CUBIC SPLINES

We consider the same set of data points, viz., the data defined in Eq. (5.1), and let $s_i(x)$ be the cubic spline defined in the interval $[x_{i-1}, x_i]$. The conditions for the *natural* cubic spline are

(i) $s_i(x)$ is atmost a cubic in each subinterval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$,

(ii) $s_i(x_i) = y_i$, $i = 0, 1, 2, \ldots, n$,

(iii) $s_i(x), s_i'(x)$ and $s_i''(x)$ are continuous in $[x_0, x_n]$, and

(iv) $s_i''(x_0) = s_i''(x_n) = 0$.

To derive the governing equations of the cubic spline, we observe that the spline second derivatives must be linear. Hence, we have in $[x_{i-1}, x_i]$:

$$s_i''(x) = \frac{1}{h_i}\left[(x_i - x)M_{i-1} + (x - x_{i-1})M_i\right], \tag{5.24}$$

where $h_i = x_i - x_{i-1}$ and $s_i''(x_i) = M_i$ for all i. Obviously, the spline second derivatives are continuous. Integrating Eq. (5.24) twice with respect to x, we get

$$s_i(x) = \frac{1}{h_i}\left[\frac{(x_i - x)^3}{6}M_{i-1} + \frac{(x - x_{i-1})^3}{6}M_i\right] + c_i(x_i - x) + d_i(x - x_{i-1}), \tag{5.25}$$

where c_i and d_i are constants to be determined.

Using conditions $s_i(x_{i-1}) = y_{i-1}$ and $s_i(x_i) = y_i$, we immediately obtain

$$c_i = \frac{1}{h_i}\left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) \quad \text{and} \quad d_i = \frac{1}{h_i}\left(y_i - \frac{h_i^2}{6}M_i\right). \tag{5.26}$$

Substituting for c_i and d_i in Eq. (5.25), we obtain

$$s_i(x) = \frac{1}{h_i}\left[\frac{(x_i - x)^3}{6}M_{i-1} + \frac{(x - x_{i-1})^3}{6}M_i + \left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right)(x_i - x)\right.$$
$$\left. + \left(y_i - \frac{h_i^2}{6}M_i\right)(x - x_{i-1})\right]. \tag{5.27}$$

In Eq. (5.27), the spline second derivatives, M_i, are still not known. To determine them, we use the condition of continuity of $s'_i(x)$. From Eq. (5.27), we obtain by differentiation:

$$s_i'(x) = \frac{1}{h_i}\left[\frac{-3(x_i - x)^2}{6}M_{i-1} + \frac{3(x - x_{i-1})^2}{6}M_i\right.$$
$$\left. - \left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) + \left(y_i - \frac{h_i^2}{6}M_i\right)\right]$$

Setting $x = x_i$ in the above, we obtain the left-hand derivative

$$s_i'(x_i-) = \frac{h_i}{2}M_i - \frac{1}{h_i}\left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right) + \frac{1}{h_i}\left(y_i - \frac{h_i^2}{6}M_i\right)$$

$$= \frac{1}{h_i}(y_i - y_{i-1}) + \frac{h_i}{6}M_{i-1} + \frac{h_i}{3}M_i \qquad (i = 1, 2, \ldots, n). \tag{5.28}$$

To obtain the right-hand derivative, we need first to write down the equation of the cubic spline in the subinterval (x_i, x_{i+1}). We do this by setting $i = i + 1$ in Eq. (5.27)

$$s_{i+1}(x) = \frac{1}{h_{i+1}}\left[\frac{(x_{i+1} - x)^3}{6}M_i + \frac{(x - x_i)^3}{6}M_{i+1} + \left(y_i - \frac{h_{i+1}^2}{6}M_i\right)(x_{i+1} - x)\right.$$

$$\left. + \left(y_{i+1} - \frac{h_{i+1}^2}{6}M_{i+1}\right)(x - x_i)\right], \tag{5.29}$$

where $h_{i+1} = x_{i+1} - x_i$. Differentiating Eq. (5.29) and setting $x = x_i$, we obtain the right-hand derivative at $x = x_i$

$$s_{i+1}'(x_i+) = \frac{1}{h_{i+1}}(y_{i+1} - y_i) - \frac{h_{i+1}}{3}M_i - \frac{h_{i+1}}{6}M_{i+1} \qquad (i = 0, 1, \ldots, n-1). \tag{5.30}$$

Equality of Eqs. (5.28) and (5.30) produces the recurrence relation

$$\frac{h_i}{6}M_{i-1} + \frac{1}{3}(h_i + h_{i+1})M_i + \frac{h_{i+1}}{6}M_{i+1}$$

$$= \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \qquad (i = 1, 2, \ldots, n-1). \tag{5.31}$$

For equal intervals, we have $h_i = h_{i+1} = h$ and Eq. (5.31) simplifies to

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(y_{i+1} - 2y_i + y_{i-1}), \qquad (i = 1, 2, \ldots, n-1). \tag{5.32}$$

The system of Eq. (5.31) has some special significance. If M_0 and M_n are known, then the system can be written as

$$\left.\begin{aligned}
2(h_1+h_2)M_1+h_2M_2 &= 6\left(\frac{y_2-y_1}{h_2}-\frac{y_1-y_0}{h_1}\right)-h_1M_0\\
h_2M_1+2(h_2+h_3)M_2+h_3M_3 &= 6\left(\frac{y_3-y_2}{h_3}-\frac{y_2-y_1}{h_2}\right)\\
h_3M_2+2(h_3+h_4)M_3+h_4M_4 &= 6\left(\frac{y_4-y_3}{h_4}-\frac{y_3-y_2}{h_3}\right)\\
&\vdots\\
h_{n-1}M_{n-2}+2(h_{n-1}+h_n)M_{n-1} &= 6\left(\frac{y_n-y_{n-1}}{h_n}-\frac{y_{n-1}-y_{n-2}}{h_{n-1}}\right)-h_nM_n.
\end{aligned}\right\} \quad (5.33)$$

Equations (5.31) or (5.32) constitute a system of $(n-1)$ equations and with the two conditions in (iv) for the *natural* spline, we have a complete system which can be solved for the M_i. Systems of the form (5.33) are called *tridiagonal* systems and in Chapter 7, we shall describe an efficient and accurate method for solving them. When the M_i are known, Eq. (5.27) then gives the required cubic spline in the subinterval $[x_{i-1}, x_i]$. Also, the y_i' can be obtained from Eqs. (5.28) and (5.30).

Example 5.3 Determine the cubic splines satisfying the data of Example 5.1. Find also the approximate values of $y(2.5)$ and $y'(2.0)$.

We have $n = 2$ and $M_0 = M_2 = 0$. Hence, the recurrence relation (5.32) gives $M_1 = 18$. If $s_1(x)$ and $s_2(x)$ are, respectively, the cubic splines in the intervals $1 \le x \le 2$ and $2 \le x \le 3$, we obtain

$$s_1(x) = 3(x-1)^3 - 8(2-x) - 4(x-1)$$

and

$$s_2(x) = 3(3-x)^3 + 22x - 48.$$

We, therefore, have

$$y(2.5) \approx s_2(2.5) = \frac{3}{8} + 7 = 7.375$$

and

$$y'(2.0) \approx s_2'(2.0) = 13.0.$$

It should be noted that the tabulated function is $y = x^3 - 9$ and hence the exact values of $y(2.5)$ and $y'(2.0)$ are, respectively, 6.625 and 12.0. The convergence to the actual values, with the increase in the order of the spline, is clearly seen from Examples 5.1, 5.2 and 5.3.

In many applications, it will be convenient to work with the spline first derivatives. Denoting $s_i'(x_i) = m_i$ *and taking suitable combinations of Eqs.* (5.28) *and* (5.30), *we can derive the following relationship for the* m_i :

$$\frac{1}{h_i}m_{i-1} + 2\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right)m_i + \frac{1}{h_{i+1}}m_{i+1}$$

$$= \frac{3}{h_{i+1}^2}(y_{i+1} - y_i) + \frac{3}{h_i^2}(y_i - y_{i-1}), \qquad i = 1, 2, \ldots, n-1. \qquad (5.34)$$

The cubic spline in (x_{i-1}, x_i) in terms of the m_i is then given by

$$s_i(x) = \frac{1}{h_i^2}\{m_{i-1}(x_i - x)^2(x - x_{i-1}) - m_i(x - x_{i-1})^2(x_i - x)\}$$

$$+ \frac{1}{h_i^3}\{y_{i-1}(x_i - x)^2[2(x - x_{i-1}) + h_i] + y_i(x - x_{i-1})^2[2(x_i - x) + h_i]\}. \quad (5.35)$$

The above result can easily be derived using the Hermite interpolation formula given in Section 3.9.3.

For equally spaced knots, Eq. (5.34) assume the simpler form:

$$m_{i-1} + 4m_i + m_{i+1} = \frac{3}{h}(y_{i+1} - y_{i-1}), \qquad i = 1, 2, \ldots, n-1. \qquad (5.36)$$

Equations (5.32) or (5.36) constitute $(n-1)$ equations in $(n+1)$ unknowns, viz., $m_0, m_1, \ldots, m_n$. Clearly, two further relations are required in order that a unique interpolating spline may be found. These conditions are called the *end conditions* and are discussed in detail in Kershaw [1971, 1972].

Specifically, we mention three types of end conditions:

(i) *Natural cubic spline:* $M_0 = M_n = 0$
(ii) D_1 *spline:* $s_i'(x_0) = y'(x_0)$, $s_i'(x_n) = y'(x_n)$,
(iii) D_2 *spline:* $s_i''(x_0) = M_0 = y''(x_0)$ and $s_i''(x_n) = M_n = y''(x_n)$.

The following example demonstrates the improvement in accuracy of the cubic spline interpolates with successive interval halving.

Example 5.4 Given the points (0, 0), (π/2,1) and (π, 0) satisfying the function $y = \sin x$ $(0 \le x \le \pi)$, determine the value of $y(\pi/6)$ using the cubic spline approximation.

We have $n = 2$ and $h = \pi/2$. The recurrence relation for the spline second derivatives gives:

$$M_0 + 4M_1 + M_2 = \frac{6 \times 4}{\pi^2}(0 - 2 + 0) = -\frac{48}{\pi^2}.$$

For the natural spline, we have $M_0 = M_2 = 0$. Hence, we have

$$M_1 = -\frac{12}{\pi^2}$$

In the interval [0, $\pi/2$], the *natural cubic spline* is given by

$$s_1(x) = \frac{2}{\pi}\left(-\frac{2x^3}{\pi^2} + \frac{3x}{2}\right).$$

Hence

$$y\left(\frac{\pi}{6}\right) \approx s_1\left(\frac{\pi}{6}\right) = \frac{2}{\pi}\left(-\frac{\pi}{108} + \frac{\pi}{4}\right) = 0.4815.$$

We next take $h = \pi/4$, i.e., the data points are (0, 0), $(\pi/4, 1/\sqrt{2})$, $(\pi/2, 1)$, $(3\pi/4, 1/\sqrt{2})$ and $(\pi, 0)$. In this case, the recurrence relation gives:

$$\left.\begin{aligned} 4M_1 + M_2 &= -4.029 \\ M_1 + 4M_2 + M_3 &= -5.699 \\ M_2 + 4M_3 &= -4.029 \end{aligned}\right\} \qquad \text{(i)}$$

since $M_0 = M_4 = 0$. Solving Eq. (i), we obtain

$$M_1 = -0.7440, \qquad M_2 = -1.053, \qquad M_3 = -0.7440.$$

In $0 \le x \le \pi/4$, the cubic spline is given by

$$s_1(x) = \frac{4}{\pi}[-0.1240(x^3) + 0.7836(x)].$$

Hence,

$$y\left(\frac{\pi}{6}\right) \approx s_1\left(\frac{\pi}{6}\right) = 0.4998.$$

This result shows that the cubic spline has produced a better approximation when the interval is halved. We finally consider values of $y = \sin x$ in intervals of 10° from $x = 0$ to π and then interpolate for x = 5°, 15°, 25°, 35° and 45°, using the natural cubic spline. The cubic spline values together with the exact values are given in the following table:

	$y = \sin x$	
x (in degrees)	*Cubic spline values*	*Exact values*
5	0.087155743	0.087155530
15	0.258819045	0.258818415
25	0.422618262	0.422617233
35	0.573576436	0.573575040
45	0.707106781	0.707105059

Example 5.5 Given the points (1, 6), (2, 18), and (3, 42), satisfying the function $y = x^3 + 5x$, determine the cubic spline in the interval [1, 2] using the end conditions $y'(1) = 8$ and $y'(3) = 32$.

We have $h = 1$ and $n = 2$. The recurrence relation is

$$m_0 + 4m_1 + m_2 = 3(y_2 - y_0)$$

$$\Rightarrow 40 + 4m_1 = 3(42 - 6) = 108$$

$$\Rightarrow m_1 = 17.$$

In [1, 2], the cubic spline is given by

$$s_1(x) = m_0(x_1 - x)^2 (x - x_0) - m_1(x - x_0)^2 (x_1 - x)$$
$$+ y_0(x_1 - x)^2 [2(x - x_0) + 1] + y_1(x - x_0)^2[2(x_1 - x) + 1]$$

Substituting the values of x_i, y_i and m_i, we obtain

$$s_1(x) = x^3 + 5x,$$

which is the tabulated function itself. In this case, the spline interpolation is exact because the two end conditions prescribed are exact and the tabulated function is a cubic.

5.2.1 Minimizing Property of Cubic Splines

We prove this property for the natural cubic spline. Let $s(x)$ be the natural cubic spline interpolating the set of data points (x_i, y_i), $i = 0, 1, 2, \ldots, n$, where it is assumed that $a = x_0 < x_1 < x_2 < \cdots < x_n = b$. Since $s(x)$ is the natural cubic spline, we have $s(x_i) = y_i$ for all i and also $s''(x_0) = s''(x_n) = 0$.

Let $z(x)$ be a function such that $z(x_i) = y_i$ for all i, and $z(x)$, $z'(x)$, $z''(x)$ are continuous in $[a, b]$. Then the integral defined by

$$I = \int_a^b [z''(x)]^2 dx \tag{5.37}$$

will be minimum if and only if $z(x) \equiv s(x)$. This means that $s(x)$ is the *smoothest* function interpolating to the set of data points defined above, since the second derivative is a good approximation to the *curvature* of a curve. We write

$$\int_a^b [z''(x)]^2 dx = \int_a^b [s''(x) + z''(x) - s''(x)]^2 dx$$

$$= \int_a^b [s''(x)^2 dx + 2\int_a^b s''(x)[z''(x) - s''(x)] dx$$

$$+ \int_a^b [z''(x) - s''(x)]^2 dx. \tag{5.38}$$

Now,

$$\int_a^b s''(x)[z''(x)-s''(x)]\,dx = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} s''(x)[z''(x)-s''(x)]\,dx$$

$$= \sum_{i=0}^{n-1}\{s''(x)[z'(x)-s'(x)]\}_{x_i}^{x_{i+1}}$$

$$-\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} s'''(x)[z'(x)-s'(x)]\,dx. \tag{5.39}$$

The first term in Eq. (5.39) simplifies to

$$s''(x_n)[z'(x_n)-s'(x_n)]-s''(x_0)[z'(x_0)-s'(x_0)].$$

Since $s''(x_n)=s''(x_0)=0,$ the above expression vanishes. Similarly, the second term in Eq. (5.39) is zero since $s'''(x)$ has a constant value in each interval and $s(x_i) = z(x_i) = y_i$, for all i. Hence, Eq. (5.38) becomes

$$\int_a^b [z''(x)]^2\,dx = \int_a^b [s''(x)]^2\,dx + \int_a^b [z''(x)-s''(x)]^2\,dx \tag{5.40}$$

or

$$\int_a^b [z''(x)]^2\,dx \ge \int_a^b [s''(x)]^2\,dx. \tag{5.41}$$

It follows that the integral

$$I = \int_a^b [z''(x)]^2\,dx$$

will be minimum if and only if

$$\int_b^a [z''(x)-s''(x)]^2\,dx = 0, \tag{5.42}$$

which means that $z''(x) = s''(x)$. Hence $z(x) - s(x)$ is a polynomial in x of degree at most three in $[a, b]$. But the difference $z(x) - s(x)$ vanishes at the points $i = 0, 1, 2, \ldots, n$. It, therefore, follows that

$$z(x) = s(x), \qquad a \le x \le b.$$

5.2.2 Error in the Cubic Spline and Its Derivatives

The errors in the natural and D_1 splines are of the order of h^2 and h^4, respectively. The errors in the derivatives are given by

$$s'(x_i) = y_i' - \frac{1}{180}h^4 y_i^{\mathrm{v}} + O(h^6) \tag{5.43}$$

$$s''(x_i) = y''(x_i) - \frac{1}{12}h^2 y^{\text{iv}}(x_i) + \frac{1}{360}h^4 y^{\text{vi}}(x_i) + \text{O}(h^6) \tag{5.44}$$

$$\frac{1}{2}\left[s'''(x_i+) + s'''(x_i-)\right] = y'''(x_i) + \frac{1}{12}h^2 y^{\text{v}}(x_i) + \text{O}(h^4) \tag{5.45}$$

$$s'''(x_i+) - s'''(x_i-) = hy^{\text{iv}}(x_i) - \frac{1}{720}h^5 y^{\text{viii}}(x_i) + \text{O}(h^7) \tag{5.46}$$

From the above equations, we obtain

$$y'(x_i) = s'(x_i) + \text{O}(h^4) \tag{5.47}$$

$$y''(x_i) = s''(x_i) + \frac{1}{12}h^2 y^{\text{iv}}(x_i) + \text{O}(h^4) \tag{5.48}$$

$$y'''(x_i) = \frac{1}{2}\left[s'''(x_i+) + s'''(x_i-)\right] + \text{O}(h^2) \tag{5.49}$$

$$y^{\text{iv}}(x_i) = \frac{1}{h}\left[s'''(x_i+) - s'''(x_i-)\right] + \text{O}(h^4) \tag{5.50}$$

These relations can be established* by using the spline recurrence relations together with the operators E and D.

5.3 SURFACE FITTING BY CUBIC SPLINES

The cubic splines derived in the previous section can be extended to functions of two or more variables. We derive the formulae for functions of two variables, the extension to higher dimensions being straightforward.** Let $L_i(x)$ be *natural cubic splines* which satisfy

$$\left.\begin{aligned} L_i(x_j) = \delta_{ij} &= 1, \quad j = i \\ &= 0, \quad j \neq i. \end{aligned}\right\} \tag{5.51}$$

These splines bear the same relation to the general cubic spline as the Lagrange polynomials bear to the Lagrange interpolation polynomial. Due to this reason, we call them *cardinal splines*. Let $s(x)$ be the natural cubic spline, in $x_{j-1} \le x \le x_j$, corresponding to the set of data points (x_j, y_j), $j = 0, 1, 2, \ldots, n$. Then, $L_i(x)$ are the cardinal splines corresponding to the set of data points $(x_j, \delta_{i,j})$, where $\delta_{i,j}$ is the *Kronecker delta* defined above.

*See, Curtis and Powell [1967].

**See, Ichida and Kiyono [1974].

The cardinal splines are given by

$$L_i(x) = \frac{1}{h}\left[\frac{(x_j - x)^3}{3!} M_{i,\,j-1} + \frac{(x - x_{j-1})^3}{3!} M_{i,\,j} + (x_j - x)\left(\delta_{i,\,j-1} - \frac{h^2}{3!} M_{i,\,j-1}\right)\right.$$
$$\left. + (x - x_{j-1})\left(\delta_{i,\,j} - \frac{h^2}{3!} M_{i,\,j}\right)\right], \tag{5.52}$$

where $M_{i,\,j} = L_i''(x_j)$. It is easy to verify that Eq. (5.52) satisfies conditions (5.51). As in the case of general splines, the condition of continuity of the first derivatives leads to the recurrence relation

$$M_{i,\,j-1} + 4M_{i,\,j} + M_{i,\,j+1} = \frac{6}{h^2}(\delta_{i,\,j-1} - 2\delta_{i,\,j} + \delta_{i,\,j+1}). \tag{5.53}$$

In terms of the cardinal splines $L_i(x)$, the general spline $s(x)$, in the interval $x_{j-1} \le x \le x_j$, can be written as

$$s(x) = \sum_{i=0}^{n} L_i(x)\, y_i, \tag{5.54}$$

where $L_i(x)$ are given by Eq. (5.52).

Extension to functions of two variables is now quite straightforward. Let the values

$$z(x_i,\ y_i), \qquad i = 0,\ 1,\ 2,\ \ldots,\ n$$

of a function of two variables, $z = f(x,\ y)$, be given at the n^2 data points arranged at the intersections of a rectangular mesh. The interpolation problem now is to determine the value of z at an arbitrary point in the rectangular region. The cubic spline formula is given by

$$s(x,\ y) = \sum_{i=0}^{n} \sum_{j=0}^{n} L_i(x)\, L_j(y)\, z_{i,\,j}, \tag{5.55}$$

where $L_i(x)$ and $L_j(y)$ are given by formulae of the type given in Eq. (5.52). The spline second derivatives, M_{ij}, are calculated from the recurrence relation (5.53) by imposing the natural end conditions, $M_{i,0} = M_{i,n} = 0$.

The following examples demonstrate the use of the formulae derived above.

Example 5.6 Using the data of Example 5.1, viz., (1, –8), (2, –1) and (3, 18), find the *cardinal splines* $L_i(x)$ and hence determine the *general natural cubic spline* in the interval $1 \le x \le 2$.

For the interval $1 \le x \le 2$ we have $j = 1$. With $h = 1$, and $j = 1$, Eq. (5.52) gives:

$$L_i(x) = \frac{(2-x)^3}{6} M_{i,0} + \frac{(x-1)^3}{6} M_{i,1} + (2-x)\left(\delta_{i,0} - \frac{1}{6} M_{i,0}\right)$$
$$+ (x-1)\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right)$$
$$= \frac{(x-1)^3}{6} M_{i,1} + (2-x)\,\delta_{i,0} + (x-1)\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right), \qquad \text{(i)}$$

since $M_{i,0} = 0$ for the *natural cubic spline*.
Similarly, the recurrence relation (5.53), becomes:

$$4M_{i,1} = 6(\delta_{i,0} - 2\delta_{i,1} + \delta_{i,2}),$$

from which we obtain

$$M_{0,1} = \frac{3}{2}, \qquad M_{1,1} = -3, \qquad M_{2,1} = \frac{3}{2}.$$

Hence, (i) gives:

$$L_0(x) = \frac{(x-1)^3}{6}\left(\frac{3}{2}\right) + (2-x) + (x-1)\left(-\frac{1}{4}\right) = \frac{1}{4}(x-1)^3 - \frac{5}{4}x + \frac{9}{4}, \qquad \text{(ii)}$$

$$L_1(x) = -\frac{1}{2}(x-1)^3 + \frac{3}{2}(x-1), \qquad \text{(iii)}$$

$$L_2(x) = \frac{1}{4}(x-1)^3 - \frac{1}{4}(x-1). \qquad \text{(iv)}$$

Hence, in $1 \le x \le 2$, the general natural cubic spline is given by

$$s(x) = \sum_{i=0}^{2} y_i L_i(x)$$
$$= \left[\frac{1}{4}(x-1)^3 - \frac{5}{4}x + \frac{9}{4}\right](-8) + \left[\frac{3}{2}(x-1) - \frac{1}{2}(x-1)^3\right](-1)$$
$$+ \left[\frac{1}{4}(x-1)^3 - \frac{1}{4}(x-1)\right](18)$$
$$= 3(x-1)^3 + 4x - 12,$$

which is the same as that obtained in Example 5.3. The next example demonstrates the use of *cardinal splines* in surface fitting.

Example 5.7 The function $z = f(x, y)$ satisfies the following data for $0 \le x, y \le 2$. Determine the *natural cubic spline* $s(x, y)$ which approximates the above data and hence find the approximate value of $z(0.5, 0.5)$.

	x		
y	0	1	2
0	1	2	9
1	2	3	10
2	9	10	17

For determining $z(0.5, 0.5)$, we need to obtain the *natural cubic spline* for the interval $0 \le x, y \le 1$.

With $h = 1$, $j = 1$, we have

$$L_i(x) = \frac{(1-x)^3}{6} M_{i,0} + \frac{x^3}{6} M_{i,1} + (1-x)\left(\delta_{i,0} - \frac{1}{6} M_{i,0}\right) + x\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right)$$

$$= \frac{x^3}{6} M_{i,1} + (1-x)\,\delta_{i,0} + x\left(\delta_{i,1} - \frac{1}{6} M_{i,1}\right), \qquad \text{(i)}$$

since $M_{i,0} = 0$ for the *natural cubic spline*. Also,

$$M_{i,1} = \frac{3}{2}(\delta_{i,0} - 2\delta_{i,1} + \delta_{i,2}).$$

Hence, we obtain

$$M_{0,1} = \frac{3}{2}, \qquad M_{1,1} = -3, \qquad M_{2,1} = \frac{3}{2}.$$

From Eq. (i), we then obtain

$$L_0(x) = \frac{x^3}{4} - \frac{5x}{4} + 1,$$

$$L_1(x) = -\frac{1}{2}x^3 + \frac{3}{2}x,$$

$$L_2(x) = \frac{1}{4}x^3 - \frac{1}{4}x.$$

Hence, in $0 \le x,\ y \le 1$, we have

$$s(x, y) = \sum_{i=0}^{2} \sum_{j=0}^{2} L_i(x)\, L_j(y)\, z_{i,j}$$

$$= L_0(x)\,[L_0(y)z_{0,0} + L_1(y)z_{0,1} + L_2(y)z_{0,2}]$$

$$+ L_1(x)\,[L_0(y)z_{1,0} + L_1(y)z_{1,1} + L_2(y)z_{1,2}]$$

$$+ L_2(x)\,[L_0(y)z_{2,0} + L_1(y)z_{2,1} + L_2(y)z_{2,2}].$$

Since $x = y = 0.5$, the preceding equation gives:

$$
\begin{aligned}
z(0.5,\ 0.5) &\approx s(0.5,\ 0.5) \\
&= \frac{13}{32}\left(\frac{13}{32}\times 1+\frac{11}{16}\times 2-\frac{3}{32}\times 9\right)+\frac{11}{16}\left(\frac{13}{32}\times 2+\frac{11}{16}\times 3-\frac{3}{32}\times 10\right) \\
&\quad -\frac{3}{32}\left(\frac{13}{32}\times 9+\frac{11}{16}\times 10-\frac{3}{32}\times 17\right) \\
&= 0.875.
\end{aligned}
$$

The tabulated function is $z = x^3 + y^3 + 1$ and therefore the exact value of z (0.5, 0.5) is 1.25, which means that the above interpolated value has an error of 30%.

5.4 CUBIC B-SPLINES

The cubic spline formulae derived in the preceding section are global in nature, which means that they do not permit any local changes in the given data. Hence, efforts were made to derive spline formulae which are 'nonglobal'. We require *basis functions* which allow the degree of the resulting curve to be changed without any change in the data. The B-splines are such *basis functions* which have important applications in computer graphics. The B-splines can be of any degree but those of second or third degree are usually found to be sufficient.

The theory of B-splines was first suggested by Schoenberg [1946], but recurrence formulae for their computation were discovered independently by Cox [1972] and de Boor [1972].

A B-spline of *order n* (or *degree n* – 1), denoted by $s_{n,i}(x)$, is a spline of degree $(n - 1)$ with knots $k_{i-n}, k_{i-n+1}, \ldots, k_i$, which is zero everywhere except in the interval $k_{i-n} < x < k_i$. The cubic B-spline $s_{4,i}(x)$ with knots $k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i$, is such that

(i) on each interval, $s_{4,i}(x)$ is a polynomial in x of degree 3 or less,
(ii) $s_{4,i}(x)$, $s'_{4,i}(x)$ and $s''_{4,i}(x)$ are continuous, and
(iii) $s_{4,i}(x) > 0$ inside $[k_{i-4}, k_i]$ and is identically zero outside $[k_{i-4}, k_i]$.

A typical cubic B-spline with knots $k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i$, is shown in Fig. 5.1.

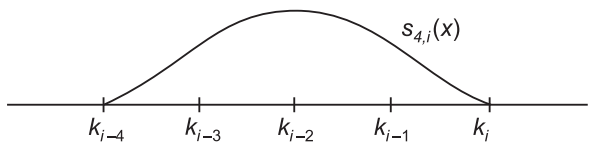


Figure 5.1 A typical cubic B-spline.

5.4.1 Representations of B-splines

To represent the cubic B-spline $s_{4,i}(x)$ at $x = k_i$, we consider the five knots

$$k_{i-4},\ k_{i-3},\ k_{i-2},\ k_{i-1},\ k_i$$

which are such that

$$k_{i-4} < k_{i-3} < k_{i-2} < k_{i-1} < k_i \tag{5.56}$$

We also define the function

$$P_+^3 = \begin{cases} P^3, & \text{when } P \geq 0 \\ 0, & \text{when } P \leq 0 \end{cases} \tag{5.57}$$

Then, a unique representation of the cubic B-spline with knots $k_{i-4}, \ldots, k_i$, is given by Greville [1968]:

$$s_{4,i}(x) = \sum_{r=0}^{3} \alpha_r x^r + \sum_{m=i-4}^{i} \beta_m (x - k_m)_+^3 \tag{5.58}$$

Another representation based on the concept of divided differences is given below.

$$s_{4,i}(x) = [k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i] \tag{5.59}$$

$$= \frac{(k_{i-4} - x)_+^3}{(k_{i-4} - k_{i-3})(k_{i-4} - k_{i-2})(k_{i-4} - k_{i-1})(k_{i-4} - k_i)}$$

$$+ \frac{(k_{i-3} - x)_+^3}{(k_{i-3} - k_{i-4})(k_{i-3} - k_{i-2})(k_{i-3} - k_{i-1})(k_{i-3} - k_i)}$$

$$+ \cdots + \frac{(k_i - x)_+^3}{(k_i - k_{i-4})(k_i - k_{i-3})(k_i - k_{i-2})(k_i - k_{i-1})} \tag{5.60}$$

Denoting

$$\Pi_{4,i}(x) = (x - k_{i-4})(x - k_{i-3})(x - k_{i-2})(x - k_{i-1})(x - k_i) \tag{5.61}$$

We obtain

$$\Pi'_{4,i}(k_i) = (k_i - k_{i-4})(k_i - k_{i-3})(k_i - k_{i-2})(k_i - k_{i-1}) \tag{5.62}$$

Hence, Eq. (5.60) can be written as

$$s_{4,i}(x) = \frac{(k_{i-4} - x)_+^3}{\Pi'_{4,i}(k_{i-4})} + \frac{(k_{i-3} - x)_+^3}{\Pi'_{4,i}(k_{i-3})} + \frac{(k_{i-2} - x)_+^3}{\Pi'_{4,i}(k_{i-2})}$$

$$+ \frac{(k_{i-1} - x)_+^3}{\Pi'_{4,i}(k_{i-1})} + \frac{(k_i - x)_+^3}{\Pi'_{4,i}(k_i)}$$

$$= \sum_{m=i-4}^{i} \frac{(k_m - x)_+^3}{\Pi'_{4,i}(k_m)} \tag{5.63}$$

More generally, a B-spline of order n (degree $n-1$) is defined by

$$s_{n,i}(x) = [k_{i-n}, k_{i-n+1}, \ldots, k_i] \tag{5.64}$$

$$= \sum_{m=i-n}^{i} \frac{(k_m - x)_+^3}{\Pi'_{n,i}(k_m)} \tag{5.65}$$

where

$$\Pi_{n,i}(x) = (x - k_{i-n})(x - k_{i-n+1}) \ \ldots \ (x - k_i) \tag{5.66}$$

and $\Pi_{n,i}(x)$ denoting its derivative with respect to x. From the definition of a divided difference, we have

$$\begin{aligned}\left[k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}, k_i\right] &= \frac{[k_{i-3}, k_{i-2}, k_{i-1}, k_i] - [k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1}]}{k_i - k_{i-4}} \\ &= \frac{s_{3,i}(x) - s_{3,i-1}(x)}{k_i - k_{i-4}}\end{aligned} \tag{5.67}$$

Thus,

$$s_{4,i}(x) = \frac{s_{3,i}(x) - s_{3,i-1}(x)}{k_i - k_{i-4}}, \tag{5.68}$$

which is a recurrence relation. Similarly, for B-splines of order n, we have the recurrence relation

$$s_{n,i}(x) = \frac{s_{n-1,i}(x) - s_{n-1,i-1}(x)}{k_i - k_{i-n}} \tag{5.69}$$

In practice, however, formulae given in Eqs. (5.58), (5.68) and (5.69) are not used since they are found to be computationally inefficient. Instead, a recurrence formula discovered independently by Cox [1972] and de Boor [1972] is found to be efficient. This formula is

$$s_{n,i}(x) = \frac{(x - k_{i-n})s_{n-1,i-1}(x) + (k_i - x)s_{n-1,i}(x)}{k_i - k_{i-n}}, \tag{5.70}$$

for all x. For proof of Eq. (5.70), see Cox [1972].

From Eq. (5.70), it can be seen that the computation of $s_{n,i}(x)$ depends on $s_{n-1,i-1}(x)$ and $s_{n-1,i}(x)$. Thus, if the knots are k_{i-4}, k_{i-3}, k_{i-2}, k_{i-1} and k_i, then the cubic B-spline $S_{4,i}(x)$ can be computed, from left to right, as in the following Fig. 5.2.

$$
\begin{array}{cccc}
s_{1,i-3} & & & \\
 & s_{2,i-2} & & \\
s_{1,i-2} & & s_{3,i-1} & \\
 & s_{2,i-1} & & s_{4,i} \\
s_{1,i-1} & & s_{3,i} & \\
 & s_{2,i} & & \\
s_{1,i} & & &
\end{array}
$$

Figure 5.2 Array of elements for computing $s_{4,i}$.

Further, simplification in the computation of the above array may be made by using the property

$$
s_{1,j} = \begin{cases} \dfrac{1}{k_j - k_{j-1}} & \text{if } k_{j-1} \le x < k_j \\ 0, & \text{otherwise.} \end{cases} \tag{5.71}
$$

For example, if $k_{i-4} \le x < k_{i-3}$, the array in Fig. 5.2 simplifies to the following Fig. 5.3.

$$
\begin{array}{cccc}
s_{1,i-3} & & & \\
 & s_{2,i-2} & & \\
0 & & s_{3,i-1} & \\
 & 0 & & s_{4,i} \\
0 & & 0 & \\
 & 0 & & \\
0 & & &
\end{array}
$$

Figure 5.3 Simplified array of Fig. 5.2.

The numerical computation of B-splines will now become more simpler. The following examples demonstrate the use of formulae (5.64) and (5.70) for the computation of B-splines.

Example 5.8 With respect to the knots 0, 1, 2, 3, 4, compute cubic B-splines at $x = 1$ and $x = 2$ using formula (5.64).

We have

$$
s_{4,i}(x) = \sum_{m=i-4}^{i} \frac{(k_m - x)_+^3}{\Pi'_{4,i}(k_m)} \tag{i}
$$

(a) $x = 1$

Here $i = 4$, $k_0 = 0$, $k_1 = 1$, $k_2 = 2$, $k_3 = 3$ and $k_4 = 4$.

Therefore,

$$s_{4,4}(1) = \sum_{m=0}^{4} \frac{(k_m - x)_+^3}{\Pi'_{4,4}(k_m)}$$

$$= \frac{(k_0 - 1)_+^3}{\Pi'_{4,4}(k_0)} + \frac{(k_1 - 1)_+^3}{\Pi'_{4,4}(k_1)} + \frac{(k_2 - 1)_+^3}{\Pi'_{4,4}(k_2)} + \frac{(k_3 - 1)_+^3}{\Pi'_{4,4}(k_3)} + \frac{(k_4 - 1)_+^3}{\Pi'_{4,4}(k_4)}$$

$$= 0 + 0 + \frac{1}{\Pi'_{4,4}(2)} + \frac{8}{\Pi'_{4,4}(3)} + \frac{27}{\Pi'_{4,4}(4)} \qquad \text{(ii)}$$

Now,

$$\Pi'_{4,4}(2) = (2 - 0)(2 - 1)(2 - 3)(2 - 4) = 4,$$

$$\Pi'_{4,4}(3) = (3 - 0)(3 - 1)(3 - 2)(3 - 4) = -6,$$

and

$$\Pi'_{4,4}(4) = (4 - 0)(4 - 1)(4 - 2)(4 - 3) = 24.$$

Hence, (ii) gives

$$s_{4,4}(1) = \frac{1}{4} - \frac{8}{6} + \frac{27}{24}$$

$$= \frac{1}{24}.$$

(b) $x = 2$

We have

$$s_{4,4}(2) = \frac{(3-2)_+^3}{\Pi'_{4,4}(3)} + \frac{(4-2)_+^3}{\Pi'_{4,4}(4)}$$

$$= \frac{1}{-6} + \frac{8}{24} = \frac{1}{6}.$$

Example 5.9 With the same data as in Example 5.8, compute cubic B-splines at $x = 1$ and $x = 2$, using formula (5.70).

The formula is

$$s_{4,i}(x) = \frac{(x - k_{i-4})s_{3,i-1}(x) + (k_i - x)s_{3,i}}{k_i - k_{i-4}} \qquad \text{(iii)}$$

(a) $x = 1$

We have $1 \le x < 2$. Then $s_{1,2} = \dfrac{1}{k_2 - k_1} = 1.$

Hence we need to compute the elements in the following array

$$\begin{array}{cccc}
0 & & & \\
 & s_{2,2} & & \\
s_{1,2} & & s_{3,3} & \\
 & s_{2,3} & & s_{4,4} \\
0 & & s_{3,4} & \\
 & 0 & & \\
0 & & &
\end{array}$$

Now,

$$s_{2,2} = \frac{(1-k_0)s_{1,1} + (k_2 - 1)s_{1,2}}{k_2 - k_0} = \frac{1}{2},$$

$$s_{2,3} = \frac{(1-k_1)s_{1,2} + (k_3 - 1)s_{1,3}}{k_3 - k_1} = 0,$$

$$s_{3,3} = \frac{(1-k_0)s_{2,2} + (k_3 - 1)s_{2,3}}{k_3 - k_0} = \frac{\frac{1}{2}}{3} = \frac{1}{6},$$

$$s_{3,4} = \frac{(1-k_1)s_{2,3} + (k_4 - 1)s_{2,4}}{k_4 - k_1} = 0,$$

$$s_{4,4} = \frac{(1-k_0)s_{3,3} + (k_4 - 1)s_{3,4}}{k_4 - k_0} = \frac{1}{24}.$$

(b) $x = 2$

We have $2 \le x < 3$

Then

$$s_{1,3} = \frac{1}{k_3 - k_2} = 1.$$

In this case, the array to be computed is given below.

$$\begin{array}{cccc}
0 & & & \\
 & 0 & & \\
0 & & s_{3,3} & \\
 & s_{2,3} & & s_{4,4} \\
s_{1,3} & & s_{3,4} & \\
 & s_{2,4} & & \\
0 & & &
\end{array}$$

$$s_{2,3} = \frac{(2-k_1)s_{1,2} + (3-2)s_{1,3}}{3 - k_1} = \frac{1}{2},$$

$$s_{2,4} = \frac{(2-k_2)s_{1,3} + (4-2)s_{1,4}}{2} = 0,$$

$$s_{3,3} = \frac{(2-0)s_{2,2} + (3-2)s_{2,3}}{3} = \frac{1}{6}$$

$$s_{3,4} = \frac{(2-1)s_{2,3} + (4-2)s_{2,4}}{3} = \frac{1}{6},$$

and

$$s_{4,4} = \frac{(2-0)s_{3,3} + (4-2)s_{3,4}}{4} = \frac{1}{6}.$$

5.4.2 Least Squares Solution

Let the set of data points be (x_i, y_i), $i = 1, 2, \ldots, m$, and $a \le x \le b$. Let $s(x)$ be the cubic spline with knots $k_1, k_2, \ldots, k_p$, where $a < k_1 < k_2 < \cdots < k_p < b$. To define the full set of B-splines, it is necessary to introduce eight additional knots, namely, k_3, k_2, k_1, k_0, k_{p+1}, k_{p+2}, k_{p+3} and k_{p+4}. These knots are chosen such that

$$k_3 < k_2 < k_1 < k_0 = a \quad \text{and} \quad b = k_{p+1} < k_{p+2} < k_{p+3} < k_{p+4} \tag{5.72}$$

We now have $(p + 4)$ B-splines (of order 4) in the range $a \le x \le b$, and then the general cubic spline $s(x)$ with knots $k_1, k_2, \ldots, k_p$ has a unique representation, in the range $a \le x \le b$, of the form

$$s(x) = \sum_{i=1}^{p+4} \alpha_i s_{4,i}(x), \quad a \le x \le b, \tag{5.73}$$

where α_i are constants to be determined.

To determine the constants α_i in Eq. (5.73), we substitute $x = x_r$ and obtain

$$s(x_r) = y_r = \sum_{i=1}^{p+4} \alpha_i\, s_{4,i}(x_r), \quad r = 1, 2, \ldots, m \tag{5.74}$$

where $m \gg p + 4$. In matrix notation, Eq. (5.74) can be written as

$$A\boldsymbol{\alpha} = \mathbf{y} \tag{5.75}$$

where A is an $m \times (p + 4)$ band matrix and $\boldsymbol{\alpha}$, $\mathbf{y}$ are column vectors. The required solution is obtained by solving the normal equations

$$A^{\mathrm{T}}A\boldsymbol{\alpha} = A^{\mathrm{T}}\mathbf{y} \tag{5.76}$$

5.4.3 Applications of B-splines

An important application of the B-spline theory is in *digital signal processing* where the B-splines are used for noise reduction in wavelet domain. See, for example, Poornachandra et al. [2004], where B-splines are adopted to obtain noisefree signals.

EXERCISES

5.1 The function $y = x^3 + 9$ is tabulated below.

$$(3, 36), (4, 73), (5, 134)$$

Predict the value of $y(4.5)$ using quadratic and natural cubic splines and state the absolute error in each case.

5.2 Given the points $(0, 0)$, $(\pi/2, 1)$ and $(\pi, 0)$ satisfying the function $y = \sin x$, $0 \le x \le \pi$, determine the value of $y\ (\pi/6)$ using quadratic spline approximation. Compare your result with that obtained in Example 5.5.

5.3 Determine the cubic spline $s(x)$ valid in the interval $[x_0, x_1]$ for the following data, given that

$$s''(x_0) = y''(x_0) \text{ and } s''(x_2) = y''(x_2).$$

Find $y(6.3)$.

x	6.2	6.4	6.6
$y = x \ln x$	11.3119	11.8803	12.4549

5.4 Fit (a) a natural cubic spline and (b) a D_1 cubic spline to the following data:

x	0.10	0.20	0.30
$y = e^x$	1.1052	1.2214	1.3499

Find $y(0.15)$ in each case and state which of these is the best fit.

5.5 Derive the cubic spline formula with D_1 end conditions. Show that

$$s'(x_i) = y_i' - \frac{1}{180} h^4 y_i^{\text{v}} + O(h^6)$$

Given that the data

$$(-1, 1), (0, 0) \text{ and } (1, 1)$$

satisfies the function $y = f(x)$ in $-1 \le x \le 1$, find the cubic spline approximation of $y\,(-0.5)$ if $y'(-1) = -4$ and $y'(1) = 4$.

5.6 Determine the cubic spline in $0 \le x \le 1$ representing the data points and the end conditions given in Question 5.5 above.

5.7 Write an algorithm to implement formula (5.27) for computing a cubic spline in the interval $[x_{i-1}, x_i]$.

5.8 The data points

$$(-2, 16), (0, 0), (2, 16)$$

satisfy the relation $y = f(x)$ in the interval $-2 \le x \le 2$. Determine the cubic spline valid in this interval and satisfying the end conditions $y'(-2) = -32$ and $y'(2) = 32$.

5.9 Show that the error in the spline second derivative is given by

$$s''(x_i) = y_i'' - \frac{1}{12}h^2 y_i^{\text{iv}} + O(h^4).$$

5.10 Prove the minimization property of Section 5.2.1 for the D_1 spline.

5.11 If the function $y(x)$ is periodic with period $(x_n - x_0)$, the spline is called a *periodic spline.* In this case, we have

$$s_0^{(r)}(x) = s_n^{(r)}(x), \quad r = 0, 1, 2.$$

Using these conditions, obtain the recurrence relation in the matrix form.

5.12 If the end conditions are taken as

$$M_1 = \frac{M_0 + M_2}{2} \quad \text{and} \quad M_{n-1} = \frac{M_n + M_{n-2}}{2},$$

then the spline curve is the same as the single cubic which is fitted throughout to the given data.

For the data points

$$(-2, -12), (-1, -8), (0, -3.9), (1, -0.1),$$

the single cubic which fits these points is given by

$$y = -\frac{1}{15}x^3 - \frac{3}{20}x^2 + \frac{241}{60}x - 3.9.$$

Find the cubic spline which fits this data with the above end conditions.

5.13 The following table gives the values of $z = f(x, y)$ for different values of x and y. Use the method of Section 5.3 to find z when $x = 1.5$ and $y = 1.5$. Compare your result with the actual value obtained from $z = f(x, y) = x^2 + y^2 + y$.

$y \backslash x$	1	2	3
1	3	6	11
2	7	10	15
3	13	16	21

5.14 Using the data

x	0	1	2
$y = x^3 - 5$	–5	–4	3

find the cardinal splines $L_i(x)$ and hence determine the natural cubic spline valid in the interval $0 \le x \le 2$.

5.15 Using the formula (5.58), determine the cubic B-spline $s(x)$ with support [0, 4] on the knots 0, 1, 2, 3, 4. State whether this is a unique representation.

5.16 With respect to the knots 0, 1, 2, 3, 4, 5, 6, compute B-splines of order 6 at $x = 1$ using Cox–de Boor recurrence formula.

Answers to Exercises

5.1 106.75 (error = 6.625); 103 (error = 2.875)

5.2 0.3333

5.3 11.5953

5.4 1.1622; 1.1618

5.5 $s_1(-0.5) = 0$

5.6 $s_2(x) = 2x^3 - x^2,\ 0 \le x \le 1$

5.8 $-4x^2 - 4x^3;\ -4x^2 + 4x^3$

5.12 $s_1(x) = -\dfrac{1}{15}x^3 - \dfrac{3}{20}x^2 + \dfrac{241}{60}x - 3.9$

5.13 6.125

5.14 $s(x) = \dfrac{3}{2}x^3 - \dfrac{1}{2}x - 5,\ 0 \le x \le 1$

5.15 $s(x) = c_3\left[x^3 - 4(x-1)_+^3 + 6(x-2)_+^3 - 4(x-3)_+^3\right]$

5.16 $s_{1,3} = 1, \quad s_{2,3} = \dfrac{1}{2}, \quad s_{3,4} = \dfrac{1}{6},$

$s_{4,5} = \dfrac{1}{24}, \quad s_{4,6} = 0 \quad s_{5,6} = \dfrac{1}{120},$

$s_{5,7} = 0 \quad s_{6,7} = \dfrac{1}{720}.$

Chapter 6

Numerical Differentiation and Integration

6.1 INTRODUCTION

In Chapter 3, we were concerned with the general problem of interpolation, viz., given the set of values (x_0, y_0), (x_1, y_1), …, (x_n, y_n) of x and y, to find a polynomial $\phi(x)$ of the lowest degree such that $y(x)$ and $\phi(x)$ agree at the set of tabulated points. In the present chapter, we shall be concerned with the problems of numerical differentiation and integration. That is to say, given the set of values of x and y, as above, we shall derive formulae to compute:

(i) $\dfrac{dy}{dx}, \dfrac{d^2y}{dx^2}, \cdots$ for any value of x in $[x_0, x_n]$, and

(ii) $\displaystyle\int_{x_0}^{x_n} y\, dx.$

6.2 NUMERICAL DIFFERENTIATION

The general method for deriving the numerical differentiation formulae is to differentiate the interpolating polynomial. Hence, corresponding to each of the formulae derived in Chapter 3, we may derive a formula for the derivative. We illustrate the derivation with Newton's forward difference formula only, the method of derivation being the same with regard to the other formulae.

Consider Newton's forward difference formula:

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \cdots, \tag{6.1}$$

where

$$x = x_0 + uh. \tag{6.2}$$

Then

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = \frac{1}{h}\left(\Delta y_0 + \frac{2u-1}{2}\Delta^2 y_0 + \frac{3u^2-6u+2}{6}\Delta^3 y_0 + \cdots\right). \tag{6.3}$$

This formula can be used for computing the value of *dy*/*dx* for *non-tabular values* of *x*. For tabular values of *x*, the formula takes a simpler form, for by setting $x = x_0$ we obtain $u = 0$ from Eq. (6.2), and hence Eq. (6.3) gives

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left(\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \cdots\right). \tag{6.4}$$

Differentiating Eq. (6.3) once again, we obtain

$$\frac{d^2y}{dx^2} = \frac{1}{h^2}\left(\Delta^2 y_0 + \frac{6u-6}{6}\Delta^3 y_0 + \frac{12u^2-36u+22}{24}\Delta^4 y_0 + \cdots\right), \tag{6.5}$$

from which we obtain

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left(\Delta^2 y_0 - \Delta^3 y_0 + \frac{11}{12}\Delta^4 y_0 + \cdots\right). \tag{6.6}$$

Formulae for computing higher derivatives may be obtained by successive differentiation. In a similar way, different formulae can be derived by starting with other interpolation formuale. Thus,

(a) Newton's backward difference formula gives

$$\left[\frac{dy}{dx}\right]_{x=x_n} = \frac{1}{h}\left(\nabla y_n + \frac{1}{2}\nabla^2 y_n + \frac{1}{3}\nabla^3 y_n + \cdots\right) \tag{6.7}$$

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_n} = \frac{1}{h^2}\left(\nabla^2 y_n + \nabla^3 y_n + \frac{11}{12}\nabla^4 y_n + \frac{5}{6}\nabla^5 y_n + \cdots\right). \tag{6.8}$$

(b) Stirling's formula gives

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{1}{h}\left(\frac{\Delta y_{-1} + \Delta y_0}{2} - \frac{1}{6}\frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2} + \frac{1}{30}\frac{\Delta^5 y_{-3} + \Delta^5 y_{-2}}{2} + \cdots\right) \tag{6.9}$$

and

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\left(\Delta^2 y_{-1} - \frac{1}{12}\Delta^4 y_{-2} + \frac{1}{90}\Delta^6 y_{-3} - \cdots\right). \tag{6.10}$$

If a derivative is required near the end of a table, one of the following formulae may be used to obtain better accuracy

$$hy_0' = \left(\Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \frac{1}{4}\Delta^4 + \frac{1}{5}\Delta^5 - \frac{1}{6}\Delta^6 + \cdots\right)y_0 \tag{6.11}$$

$$= \left(\Delta + \frac{1}{2}\Delta^2 - \frac{1}{6}\Delta^3 + \frac{1}{12}\Delta^4 - \frac{1}{20}\Delta^5 + \frac{1}{30}\Delta^6 - \cdots\right)y_{-1} \tag{6.12}$$

$$h^2 y_0'' = \left(\Delta^2 - \Delta^3 + \frac{11}{12}\Delta^4 - \frac{5}{6}\Delta^5 + \frac{137}{180}\Delta^6 - \frac{7}{10}\Delta^7 + \frac{363}{560}\Delta^8 - \cdots\right)y_0 \tag{6.13}$$

$$= \left(\Delta^2 - \frac{1}{12}\Delta^4 + \frac{1}{12}\Delta^5 - \frac{13}{180}\Delta^6 + \frac{11}{180}\Delta^7 - \frac{29}{560}\Delta^8 + \cdots\right)y_{-1} \tag{6.14}$$

$$hy_n' = \left(\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \frac{1}{4}\nabla^4 + \frac{1}{5}\nabla^5 + \frac{1}{6}\nabla^6 + \frac{1}{7}\nabla^7 + \frac{1}{8}\nabla^8 + \cdots\right)y_n \tag{6.15}$$

$$= \left(\nabla - \frac{1}{2}\nabla^2 - \frac{1}{6}\nabla^3 - \frac{1}{12}\nabla^4 - \frac{1}{20}\nabla^5 - \frac{1}{30}\nabla^6 - \frac{1}{42}\nabla^7 - \frac{1}{56}\nabla^8 - \cdots\right)y_{n+1} \tag{6.16}$$

$$h^2 y_n'' = \left(\nabla^2 + \nabla^3 + \frac{11}{12}\nabla^4 + \frac{5}{6}\nabla^5 + \frac{137}{180}\nabla^6 + \frac{7}{10}\nabla^7 + \frac{363}{560}\nabla^8 + \cdots\right)y_n \tag{6.17}$$

$$= \left(\nabla^2 - \frac{1}{12}\nabla^4 - \frac{1}{12}\nabla^5 - \frac{13}{180}\nabla^6 - \frac{11}{180}\nabla^7 - \frac{29}{560}\nabla^8 - \cdots\right)y_{n+1}. \tag{6.18}$$

For more details, the reader is referred to Interpolation and Allied Tables. The following examples illustrate the use of the formulae stated above.

Example 6.1 From the following table of values of x and y, obtain dy/dx and d^2y/dx^2 for $x = 1.2$:

x	y	x	y
1.0	2.7183	1.8	6.0496
1.2	3.3201	2.0	7.3891
1.4	4.0552	2.2	9.0250
1.6	4.9530		

The difference table is

x	y	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
1.0	2.7183						
		0.6018					
1.2	3.3201		0.1333				
		0.7351		0.0294			
1.4	4.0552		0.1627		0.0067		
		0.8978		0.0361		0.0013	
1.6	4.9530		0.1988		0.0080		0.0001
		1.0966		0.0441		0.0014	
1.8	6.0496		0.2429		0.0094		
		1.3395		0.0535			
2.0	7.3891		0.2964				
		1.6359					
2.2	9.0250						

Here $x_0 = 1.2$, $y_0 = 3.3201$ and $h = 0.2$. Hence Eq. (6.11) gives

$$\left[\frac{dy}{dx}\right]_{x=1.2} = \frac{1}{0.2}\left[0.7351 - \frac{1}{2}(0.1627) + \frac{1}{3}(0.0361) - \frac{1}{4}(0.0080) + \frac{1}{5}(0.0014)\right]$$

$$= 3.3205.$$

If we use formula (6.12), then we should use the differences diagonally downwards from 0.6018 and this gives

$$\left[\frac{dy}{dx}\right]_{x=1.2} = \frac{1}{0.2}\left[0.6018 + \frac{1}{2}(0.1333) - \frac{1}{6}(0.0294) + \frac{1}{12}(0.0067) - \frac{1}{20}(0.0013\right]$$

$$= 3.3205, \text{ as before.}$$

Similarly, formula (6.13) gives

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.2} = \frac{1}{0.04}\left[0.1627 - 0.0361 + \frac{11}{12}(0.0080) - \frac{5}{6}(0.0014)\right] = 3.318.$$

Using formula (6.14), we obtain

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.2} = \frac{1}{0.04}\left[0.1333 - \frac{1}{12}(0.0067) + \frac{1}{12}(0.0013)\right] = 3.32.$$

Example 6.2 Calculate the first and second derivatives of the function tabulated in the preceding example at the point $x = 2.2$ and also dy/dx at $x = 2.0$.

We use the table of differences of Example 6.1. Here $x_n = 2.2$, $y_n = 9.0250$ and $h = 0.2$. Hence formula (6.15) gives

$$\left[\frac{dy}{dx}\right]_{x=2.2} = \frac{1}{0.2}\left[1.6359 + \frac{1}{2}(0.2964) + \frac{1}{3}(0.0535) + \frac{1}{4}(0.0094) + \frac{1}{5}(0.0014)\right]$$
$$= 9.0228.$$

$$\left[\frac{d^2y}{dx^2}\right]_{x=2.2} = \frac{1}{0.04}\left[0.2964 + 0.0535 + \frac{11}{12}(0.0094) + \frac{5}{6}(0.0014)\right] = 8.992.$$

To find dy/dx at $x = 2.0$, we can use either (6.15) or (6.16). Formula (6.15) gives

$$\left[\frac{dy}{dx}\right]_{x=2.0} = \frac{1}{0.2}\left[1.3395 + \frac{1}{2}(0.2429) + \frac{1}{3}(0.0441) + \frac{1}{4}(0.0080)\right.$$
$$\left. + \frac{1}{5}(0.0013) + \frac{1}{6}(0.0001)\right]$$
$$= 7.3896.$$

whereas from formula (6.16), we obtain

$$\left[\frac{dy}{dx}\right]_{x=2.0} = \frac{1}{0.2}\left[1.6359 - \frac{1}{2}(0.2964) - \frac{1}{6}(0.0535) - \frac{1}{12}(0.0094) - \frac{1}{20}(0.0014)\right]$$
$$= 7.3896.$$

Example 6.3 Find dy/dx and d^2y/dx^2 at $x = 1.6$ for the tabulated function of Example 6.1.

Choosing $x_0 = 1.6$, formula (6.9) gives

$$\left[\frac{dy}{dx}\right]_{x=1.6} = \frac{1}{0.2}\left(\frac{0.8978 + 1.0966}{2} - \frac{1}{2}\,\frac{0.0361 + 0.0441}{2} + \frac{1}{30}\,\frac{0.0013 + 0.0014}{2}\right)$$
$$= 4.9530.$$

Similarly, formula (6.10) yields

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.6} = \frac{1}{0.04}\left[0.1988 - \frac{1}{12}(0.0080) + \frac{1}{90}(0.0001)\right] = 4.9525.$$

In the preceding examples, the tabulated function is e^x and hence it is easy to see that the error is considerably more in the case of the second derivatives. This is due to the reason that although the tabulated function and its approximating polynomial would agree at the set of data points, *their slopes at these points may vary considerably*. Numerical differentiation, is, therefore, an unsatisfactory process and should be used only in 'rare cases.' The next section will be devoted to a discussion of errors in the numerical differentiation formulae.

6.2.1 Errors in Numerical Differentiation

The numerical computation of derivatives involves two types of errors, viz. *truncation errors* and *rounding errors*. These are discussed below.

The truncation error is caused by replacing the tabulated function by means of an interpolating polynomial. This error can usually be estimated by formula (3.7). As noted earlier, this formula is of theoretical interest only, since, in practical computations, we usually do not have any information about the derivative $y^{(n+1)}(\xi)$. However, the truncation error in any numerical differentiation formula can easily be estimated in the following manner. Suppose that the tabulated function is such that its differences of a certain order are small and that the tabulated function is well approximated by the polynomial. (This means that the tabulated function does not have any rapidly varying components.) We know that 2ε is the total absolute error in the values of Δy_i, 4ε in the values of $\Delta^2 y_i$, etc., where ε is the absolute error in the values of y_i. Consider now, for example, Stirling's formula (6.9). This can be written in the form

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{\Delta y_{-1} + \Delta y_0}{2h} + T_1 = \frac{y_1 - y_{-1}}{2h} + T_1, \tag{6.19}$$

where T_1, the truncation error, is given by

$$T_1 = \frac{1}{6h}\left|\frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2}\right|. \tag{6.20}$$

Similarly, formula (6.10) can be written as

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{1}{h^2}\Delta^2 y_{-1} + T_2, \tag{6.21}$$

where

$$T_2 = \frac{1}{12h^2}|\Delta^4 y_{-2}|. \tag{6.22}$$

The *rounding error*, on the other hand, is inversely proportional to h in the case of first derivatives, inversely proportional to h^2 in the case of second

derivatives, and so on. Thus, *rounding error* increases as h decreases. Considering again Stirling's formula in the form of Eq. (6.19), the rounding error does not exceed $2\varepsilon/2h = \varepsilon/h$, where ε is the maximum error in the value of y_i. On the other hand, the formula

$$\left[\frac{dy}{dx}\right]_{x=x_0} = \frac{\Delta y_{-1} + \Delta y_0}{2h} - \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{12h} + \cdots$$

$$= \frac{y_{-2} - 8y_{-1} + 8y_1 - y_2}{12h} + \cdots \tag{6.23}$$

has the maximum rounding error

$$\frac{18\varepsilon}{12h} = \frac{3\varepsilon}{2h}.$$

Finally, the formula

$$\left[\frac{d^2y}{dx^2}\right]_{x=x_0} = \frac{\Delta^2 y_{-1}}{h^2} + \cdots = \frac{y_{-1} - 2y_0 + y_1}{h^2} + \cdots \tag{6.24}$$

has the maximum rounding error $4\varepsilon/h^2$. It is clear that in the case of higher derivatives, the rounding error increases rather rapidly.

Example 6.4 Assuming that the function values given in the table of Example 6.1 are correct to the accuracy given, estimate the errors in the values of dy/dx and d^2y/dx^2 at $x = 1.6$.

Since the values are correct to 4D, it follows that $\varepsilon < 0.00005 = 0.5 \times 10^{-4}$.

Value of dy/dx at x = 1.6:

$$\textit{Truncation error} = \frac{1}{6h}\left|\frac{\Delta^3 y_{-1} + \Delta^3 y_0}{2}\right|, \quad \text{from (6.20)}$$

$$= \frac{1}{6(0.2)}\,\frac{0.0361 + 0.0441}{2}$$

$$= 0.03342$$

and

$$\textit{Rounding error} = \frac{3\varepsilon}{2h}, \quad \text{from (6.23)}$$

$$= \frac{3(0.5)\,10^{-4}}{0.4}$$

$$= 0.00038.$$

Hence,

$$\text{Total error} = 0.03342 + 0.00038 = 0.0338.$$

Using Stirling's formula from Eq. (6.19), with the first differences, we obtain

$$\left(\frac{dy}{dx}\right)_{x=1.6} = \frac{\Delta y_{-1} + \Delta y_0}{2h} = \frac{0.8978 + 1.0966}{0.4} = \frac{1.9944}{0.4} = 4.9860.$$

The *exact value* is 4.9530 so that the error in the above solution is (4.9860 – 4.9530), i.e., 0.0330, which agrees with the total error obtained above.

Value of d^2y/dx^2 *at* $x = 1.6$: Using Eq. (6.24), we obtain

$$\left[\frac{d^2y}{dx^2}\right]_{x=1.6} = \frac{\Delta^2 y_{-1}}{h^2} = \frac{0.1988}{0.04} = 4.9700$$

so that the error = 4.9700 – 4.9530 = 0.0170.

Also,

$$\textit{Truncation error} = \frac{1}{12h^2}|\Delta^4 y_{-2}| = \frac{1}{12(0.04)} \times 0.0080 = 0.01667$$

and

$$\textit{Rounding error} = \frac{4\varepsilon}{h^2} = \frac{4 \times 0.5 \times 10^{-4}}{0.04} = 0.0050.$$

Hence

$$\text{Total error in} \left[\frac{d^2y}{dx^2}\right]_{x=1.6} = 0.0167 + 0.0050 = 0.0217.$$

6.2.2 Cubic Spline Method

The cubic spline derived in Section 5.2 can conveniently be used to compute the *first* and *second* derivatives of a function. For a natural cubic spline, the recurrence formulae (5.31) or (5.32) may be used to compute the spline second derivatives depending upon the choice of the subdivisions. Then Eq. (5.29) gives the spline in the interval of interest, from which the first derivatives can be computed. For the first derivatives at the tabular points, it would, of course, be easier to use formulae (5.28) and (5.30) directly. If, on the other hand, end conditions involving the first derivatives are given, then recurrence formulae (5.34) or (5.36) may be used to compute the remaining first derivatives.

The following examples illustrate the use of the spline formulae in numerical differentiation.

Example 6.5 We consider the function $y(x) = \sin x$ in $[0, \pi]$.

Here $M_0 = M_N = 0$. Let $N = 2$, i.e., $h = \pi/2$. Then

$$y_0 = y_2 = 0, \quad y_1 = 1 \quad \text{and} \quad M_0 = M_2 = 0.$$

Using formulae (5.32), we obtain

$$M_0 + 4M_1 + M_2 = \frac{6}{h^2}(y_0 - 2y_1 + y_2)$$

or

$$M_1 = -\frac{12}{\pi^2}.$$

Formula (5.29) now gives the spline in each interval. Thus, in $0 \le x \le \pi/2$, we obtain

$$s(x) = \frac{2}{\pi}\left(\frac{-2x^3}{\pi^2} + \frac{3x}{2}\right),$$

which gives

$$s'(x) = \frac{2}{\pi}\left[-\frac{2}{\pi^2}(3x^2) + \frac{3}{2}\right]. \qquad \text{(i)}$$

Hence

$$s'\left(\frac{\pi}{4}\right) = \frac{2}{\pi}\left(-\frac{6}{\pi^2}\frac{\pi^2}{16} + \frac{3}{2}\right) = \frac{9}{4\pi} = 0.71619725.$$

Exact value of $s'(\pi/4) = \cos \pi/4 = 1/\sqrt{2} = 0.70710681$. The percentage error in the computed value of $s'(\pi/4)$ is 1.28%. From (i),

$$s''(x) = -\frac{24}{\pi^3}x$$

and hence

$$s''\left(\frac{\pi}{4}\right) = -\frac{24}{\pi^3}\frac{\pi}{4} = -\frac{6}{\pi^2} = -0.60792710.$$

Since the exact value is $-1/\sqrt{2}$, the percentage error in this result is 14.03 %. We now consider values of $y = \sin x$ in intervals of 10° from $x = 0$ to π. To obtain the spline second derivatives we used a computer and the results are given in the following table (up to $x = 90°$).

	$y''(x)$	
x (in degrees)	*Exact*	*Cubic spline*
10	−0.173 648 178	−0.174 089 426
20	−0.342 020 143	−0.342 889 233
30	−0.500 000 000	−0.501 270 524
40	−0.642 787 610	−0.644 420 964
50	−0.766 044 443	−0.767 990 999
60	−0.866 025 404	−0.868 226 016
70	−0.939 692 621	−0.942 080 425
80	−0.984 807 753	−0.987 310 197
90	−1.000 000 000	−1.002 541 048

It is seen that there is greater inaccuracy in the values of the spline second derivatives.

Example 6.6 From the following data for $y(x)$, find $y'(1.0)$.

x	–2	–1	2	3
$y(x)$	–12	–8	3	5

The function from which the above data was calculated is given by $y = -\frac{1}{15}x^3 - \frac{3}{20}x^2 + \frac{241}{60}x - 3.9$. Hence, the exact value of $y'(1)$ is 3.51667.

To apply the cubic spline formula (5.31), we observe that $h_1 = 1$, $h_2 = 3$ and $h_3 = 1$.

For $i = 1, 2$, the recurrence relation gives:

$$8M_1 + 3M_2 = -2$$

and

$$3M_1 + 8M_2 = -10,$$

since $M_0 = M_3 = 0$. We obtain $M_1 = \frac{14}{55}$ and $M_2 = -\frac{74}{55}$. In $-1 \le x \le 2$, we have

$$s_2(x) = \frac{1}{3}\left[\frac{(2-x)^3}{6}\cdot\frac{14}{55} + \frac{(x+1)^3}{6}\left(-\frac{74}{55}\right)\right]$$

$$+\frac{1}{3}\left[-8 - \frac{21}{55}\right](2-x) + \frac{1}{3}\left[3 - \frac{9}{6}\left(-\frac{74}{55}\right)\right](x+1)$$

Differentiating the above and putting $x = 1$, we obtain

$$y'(1) \approx s_2'(1.0) = \frac{1}{3}\left[-\frac{7}{55} - \frac{148}{55} + \frac{461}{55} + \frac{276}{55}\right]$$

$$= 3.52727, \text{ on simplification.}$$

6.2.3 Differentiation Formulae with Function Values

In Section 6.2, we developed forward, backward and central difference approximations of derivatives in terms of finite differences. From the computational point of view, it would be convenient to express the numerical differentiation formulae in terms of function values. We list below some differentiation formulae for use in numerical computations.

(i) *Forward Differences*

$$y'(x_i) = \frac{y_{i+1} - y_i}{h}; \qquad y'(x_i) = \frac{-y_{i+2} + 4y_{i+1} - 3y_i}{2h} + O(h^2)$$

$$y''(x_i) = \frac{y_i - 2y_{i+1} + y_{i+2}}{h^2}; \qquad y''(x_i) = \frac{-y_{i+3} + 4y_{i+2} - 5y_{i+1} + 2y_i}{h^2};$$

(ii) *Backward Differences*

$$y'(x_i) = \frac{y_i - y_{i-1}}{h}; \qquad y'(x_i) = \frac{3y_i - 4y_{i-1} + y_{i-2}}{2h};$$

$$y''(x_i) = \frac{y_i - 2y_{i-1} + y_{i-2}}{h^2}; \qquad y''(x_i) = \frac{2y_i - 5y_{i-1} + 4y_{i-2} - y_{i-3}}{h^2};$$

(iii) *Central Differences*

$$y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h}; \qquad y'(x_i) = \frac{-y_{i+2} + 8y_{i+1} - 8y_{i-1} + y_{i-2}}{12h};$$

$$y''(x_i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2};$$

$$y''(x_i) = \frac{-y_{i+2} + 16y_{i+1} - 30y_i + 16y_{i-1} - y_{i-2}}{12h^2}$$

These formulae can be derived by using Taylor series expansion of the functions.

6.3 MAXIMUM AND MINIMUM VALUES OF A TABULATED FUNCTION

It is known that the maximum and minimum values of a function can be found by equating the first derivative to zero and solving for the variable. The same procedure can be applied to determine the maxima and minima of a tabulated function.

Consider Newton's forward difference formula

$$y = y_0 + p\Delta y_0 + \frac{p(p-1)}{2}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{6}\Delta^3 y_0 + \cdots$$

Differentiating this with respect to p, we obtain

$$\frac{dy}{dp} = \Delta y_0 + \frac{2p-1}{2}\Delta^2 y_0 + \frac{3p^2 - 3p + 2}{6}\Delta^3 y_0 + \cdots \tag{6.25}$$

For maxima or minima $dy/dp = 0$. Hence, terminating the right-hand side, for simplicity, after the third difference and equating it to zero, we obtain the quadratic for p

$$c_0 + c_1 p + c_2 p^2 = 0, \tag{6.26}$$

where

$$\left.\begin{aligned} c_0 &= \Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 \\ c_1 &= \Delta^2 y_0 - \Delta^3 y_0 \\ \text{and} \quad c_2 &= \frac{1}{2}\Delta^3 y_0. \end{aligned}\right\} \tag{6.27}$$

Values of x can then be found from the relation $x = x_0 + ph$.

Example 6.7 From the following table, find x, correct to two decimal places, for which y is maximum and find this value of y.

x	y
1.2	0.9320
1.3	0.9636
1.4	0.9855
1.5	0.9975
1.6	0.9996

The table of differences is

x	y	Δ	Δ^2
1.2	0.9320		
		0.0316	
1.3	0.9636		–0.0097
		0.0219	
1.4	0.9855		–0.0099
		0.0120	
1.5	0.9975		–0.0099
		0.0021	
1.6	0.9996		

Let x_0 = 1.2. Then formula (6.25), terminated after second differences, gives

$$0 = 0.0316 + \frac{2p-1}{2}(-0.0097)$$

from which we obtain p = 3.8. Hence

$$x = x_0 + ph = 1.2 + (3.8)\,(0.1) = 1.58.$$

For this value of x, Newton's backward difference formula at $x_n = 1.6$ gives

$$\begin{aligned} y(1.58) &= 0.9996 - 0.2\,(0.0021) + \frac{-0.2\,(-0.2+1)}{2}\,(-0.0099) \\ &= 0.9996 - 0.0004 + 0.0008 \\ &= 1.0. \end{aligned}$$

6.4 NUMERICAL INTEGRATION

The general problem of numerical integration may be stated as follows. Given a set of data points (x_0, y_0), (x_1, y_1), …, (x_n, y_n) of a function $y = f(x)$, where $f(x)$ is not known explicitly, it is required to compute the value of the definite integral

$$I = \int_a^b y\,dx. \tag{6.28}$$

As in the case of numerical differentiation, one replaces $f(x)$ by an interpolating polynomial $\phi(x)$ and obtains, on integration, an approximate value of the definite integral. Thus, different integration formulae can be obtained depending upon the type of the interpolation formula used. We derive in this section a general formula for numerical integration using Newton's forward difference formula.

Let the interval $[a, b]$ be divided into n equal subintervals such that $a = x_0 < x_1 < x_2 < \cdots x_n = b$. Clearly, $x_n = x_0 + nh$. Hence the integral becomes

$$I = \int_{x_0}^{x_n} y\, dx.$$

Approximating y by Newton's forward difference formula, we obtain

$$I = \int_{x_0}^{x_n} \left[y_0 + p\Delta y_0 + \frac{p(p-1)}{2}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{6}\Delta^3 y_0 + \cdots \right] dx.$$

Since $x = x_0 + ph, dx = h\, dp$ and hence the above integral becomes

$$I = h\int_0^n \left[y_0 + p\Delta y_0 + \frac{p(p-1)}{2}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{6}\Delta^3 y_0 + \cdots \right] dp,$$

which gives on simplification

$$\int_{x_0}^{x_n} y\, dx = nh\left[y_0 + \frac{n}{2}\Delta y_0 + \frac{n(2n-3)}{12}\Delta^2 y_0 + \frac{n(n-2)^2}{24}\Delta^3 y_0 + \cdots \right]. \quad (6.29)$$

From this *general formula*, we can obtain different integration formulae by putting n = 1, 2, 3, …, etc. We derive here a few of these formulae but it should be remarked that the trapezoidal and Simpson's 1/3-rules are found to give sufficient accuracy for use in practical problems.

6.4.1 Trapezoidal Rule

Setting n = 1 in the general formula (6.29), all differences higher than the first will become zero and we obtain

$$\int_{x_0}^{x_1} y\, dx = h\left(y_0 + \frac{1}{2}\Delta y_0 \right) = h\left[y_0 + \frac{1}{2}(y_1 - y_0) \right] = \frac{h}{2}(y_0 + y_1). \quad (6.30)$$

For the next interval $[x_1, x_2]$, we deduce similarly

$$\int_{x_1}^{x_2} y\, dx = \frac{h}{2}(y_1 + y_2) \quad (6.31)$$

and so on. For the last interval $[x_{n-1}, x_n]$, we have

$$\int_{x_{n-1}}^{x_n} y\,dx = \frac{h}{2}(y_{n-1} + y_n). \tag{6.32}$$

Combining all these expressions, we obtain the rule

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{2}[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n], \tag{6.33}$$

which is known as the *trapezoidal rule*.

The geometrical significance of this rule is that the curve $y = f(x)$ is replaced by n straight lines joining the points (x_0, y_0) and (x_1, y_1); (x_1, y_1) and (x_2, y_2), …, (x_{n-1}, y_{n-1}) and (x_n, y_n). The area bounded by the curve $y = f(x)$, the ordinates $x = x_0$ and $x = x_n$, and the x-axis is then approximately equivalent to the sum of the areas of the n trapeziums obtained.

The error of the trapezoidal formula can be obtained in the following way. Let $y = f(x)$ be continuous, well-behaved, and possess continuous derivatives in $[x_0, x_n]$. Expanding y in a Taylor's series around $x = x_0$, we obtain

$$\int_{x_0}^{x_1} y\,dx = \int_{x_0}^{x_1} \left[y_0 + (x - x_0)y_0' + \frac{(x - x_0)^2}{2} y_0'' + \cdots \right] dx$$

$$= hy_0 + \frac{h^2}{2} y_0' + \frac{h^3}{6} y_0'' + \cdots \tag{6.34}$$

Similarly,

$$\frac{h}{2}(y_0 + y_1) = \frac{h}{2}\left(y_0 + y_0 + hy_0' + \frac{h^2}{2} y_0'' + \frac{h^3}{6} y_0''' + \cdots \right)$$

$$= hy_0 + \frac{h^2}{2} y_0' + \frac{h^3}{4} y_0'' + \cdots \tag{6.35}$$

From Eqs. (6.34) and (6.35), we obtain

$$\int_{x_0}^{x_1} y\,dx - \frac{h}{2}(y_0 + y_1) = -\frac{1}{12} h^3 y_0'' + \cdots, \tag{6.36}$$

which is the error in the interval $[x_0, x_1]$. Proceeding in a similar manner we obtain the errors in the remaining subintervals, viz., $[x_1, x_2]$, $[x_2, x_3]$, … and $[x_{n-1}, x_n]$. We thus have

$$E = -\frac{1}{12}h^3(y_0'' + y_1'' + \cdots + y_{n-1}''), \tag{6.37}$$

where E is the *total error*. Assuming that $y''(\bar{x})$ is the largest value of the n quantities on the right-hand side of Eq. (6.37), we obtain

$$E = -\frac{1}{12}h^3 n y''(\bar{x}) = -\frac{b-a}{12}h^2 y''(\bar{x}) \tag{6.38}$$

since $nh = b - a$.

6.4.2 Simpson's 1/3-Rule

This rule is obtained by putting $n = 2$ in Eq. (6.29), i.e. by replacing the curve by $n/2$ arcs of second-degree polynomials or parabolas. We have then

$$\int_{x_0}^{x_2} y\,dx = 2h\left(y_0 + \Delta y_0 + \frac{1}{6}\Delta^2 y_0\right) = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

Similarly,

$$\int_{x_2}^{x_4} y\,dx = \frac{h}{3}(y_2 + 4y_3 + y_4)$$

$$\vdots$$

and finally

$$\int_{x_{n-2}}^{x_n} y\,dx = \frac{h}{3}(y_{n-2} + 4y_{n-1} + y_n).$$

Summing up, we obtain

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + \cdots + y_{n-1}) + 2(y_2 + y_4 + y_6 + \cdots + y_{n-2}) + y_n], \tag{6.39}$$

which is known as *Simpson's* 1/3-*rule*, or simply Simpson's rule. It should be noted that this rule requires the division of the whole range into an even number of subintervals of width h.

Following the method outlined in Section 6.4.1, it can be shown that the error in Simpson's rule is given by

$$\int_a^b y\,dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + \cdots + y_{n-1}) + 2(y_2 + y_4 + y_6 + \cdots + y_{n-2}) + y_n]$$

$$= -\frac{b-a}{180}h^4 y^{\text{iv}}(\bar{x}), \tag{6.40}$$

where $y^{\text{iv}}(\bar{x})$ is the largest value of the fourth derivatives.

6.4.3 Simpson's 3/8-Rule

Setting $n = 3$ in Eq. (6.29), we observe that all the differences higher than the third will become zero and we obtain

$$\int_{x_0}^{x_3} y\,dx = 3h\left(y_0 + \frac{3}{2}\Delta y_0 + \frac{3}{4}\Delta^2 y_0 + \frac{1}{8}\Delta^3 y_0\right)$$

$$= 3h\left[y_0 + \frac{3}{2}(y_1 - y_0) + \frac{3}{4}(y_2 - 2y_1 + y_0) + \frac{1}{8}(y_3 - 3y_2 + 3y_1 - y_0)\right]$$

$$= \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + y_3).$$

Similarly

$$\int_{x_3}^{x_6} y\,dx = \frac{3h}{8}(y_3 + 3y_4 + 3y_5 + y_6)$$

and so on. Summing up all these, we obtain

$$\int_{x_0}^{x_n} y\,dx = \frac{3h}{8}[(y_0 + 3y_1 + 3y_2 + y_3) + (y_3 + 3y_4 + 3y_5 + y_6) + \cdots$$

$$+ (y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n)]$$

$$= \frac{3h}{8}(y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + \cdots$$

$$+ 2y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n) \tag{6.41}$$

This rule, called Simpson's (3/8)-rule, is not so accurate as Simpson's rule, the dominant term in the error of this formula being $-(3/80)\,h^5 y^{\text{iv}}(\bar{x})$.

6.4.4 Boole's and Weddle's Rules

If we wish to retain differences up to those of the fourth order, we should integrate between x_0 and x_4 and obtain Boole's formula

$$\int_{x_0}^{x_4} y\,dx = \frac{2h}{45}(7y_0 + 32y_1 + 12y_2 + 32y_3 + 7y_4) \tag{6.42}$$

The leading term in the error of this formula can be shown to be

$$-\frac{8h^7}{945} y^{\text{vi}}(\bar{x}).$$

If, on the other hand, we integrate between x_0 and x_6 retaining differences up to those of the sixth order, we obtain Weddle's rule

$$\int_{x_0}^{x_6} y\,dx = \frac{3h}{10}(y_0 + 5y_1 + y_2 + 6y_3 + y_4 + 5y_5 + y_6), \tag{6.43}$$

the error in which is given by $-(h^7/140)y^{\text{vi}}(\bar{x})$.

These two formulae can also be generalized as in the previous cases. It should, however, be noted that the number of strips will have to be a multiple of four in the case of Boole's rule and a multiple of six for Weddle's rule.

6.4.5 Use of Cubic Splines

If $s(x)$ is the cubic spline in the interval (x_{i-1}, x_i), then we have

$$I = \int_{x_0}^{x_n} y\,dx \approx \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} s(x)\,dx$$

$$= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left\{ \frac{1}{6h}[(x_i - x)^3 M_{i-1} + (x - x_{i-1})^3 M_i] \right.$$

$$\left. + \frac{1}{h}(x_i - x)\left(y_{i-1} - \frac{h^2}{6} M_{i-1} \right) + \frac{1}{h}(x - x_{i-1})\left(y_i - \frac{h^2}{6} M_i \right) \right\} dx,$$

using Eq. (5.27). On carrying out the integration and simplifying, we obtain

$$I = \sum_{i=1}^{n} \left[\frac{h}{2}(y_{i-1} + y_i) - \frac{h^3}{24}(M_{i-1} + M_i) \right], \tag{6.44}$$

where M_i, the spline second-derivatives, are calculated from the recurrence relation

$$M_{i-1} + 4M_i + M_{i+1} = \frac{6}{h^2}(y_{i-1} - 2y_i + y_{i+1}), \qquad i = 1, 2, \ldots, n-1.$$

The use of the cubic spline method is demonstrated in Example 6.13.

6.4.6 Romberg Integration

This method can often be used to improve the approximate results obtained by the finite-difference methods. Its application to the numerical evaluation of definite integrals, for example in the use of trapezoidal rule, can be described, as follows. We consider the definite integral

$$I = \int_a^b y\,dx$$

and evaluate it by the trapezoidal rule (6.33) with two different subintervals of widths h_1 and h_2 to obtain the approximate values I_1 and I_2, respectively. Then Eq. (6.38) gives the errors E_1 and E_2 as

$$E_1 = -\frac{1}{12}(b-a)h_1^2 y''(\bar{x}) \tag{6.45}$$

and

$$E_2 = -\frac{1}{12}(b-a)h_2^2 y''(\bar{\bar{x}}). \tag{6.46}$$

Since the term $y''(\bar{\bar{x}})$ in Eq. (6.46) is also the largest value of $y''(x)$, it is reasonable to assume that the quantities $y''(\bar{x})$ and $y''(\bar{\bar{x}})$ are very nearly the same. We therefore have

$$\frac{E_1}{E_2} = \frac{h_1^2}{h_2^2}$$

and hence

$$\frac{E_2}{E_2 - E_1} = \frac{h_2^2}{h_2^2 - h_1^2}.$$

Since $E_2 - E_1 = I_2 - I_1$, this gives

$$E_2 = \frac{h_2^2}{h_2^2 - h_1^2}(I_2 - I_1). \tag{6.47}$$

We therefore obtain a new approximation I_3 defined by

$$I_3 = I_2 - E_2 = \frac{I_1\, h_2^2 - I_2\, h_1^2}{h_2^2 - h_1^2}, \tag{6.48}$$

which, in general, would be closer to the actual value—provided that the errors decrease monotonically and are of the same sign.

If we now set

$$h_2 = \frac{1}{2}h_1 = \frac{1}{2}h$$

Equation (6.48) can be written in the more convenient form

$$I\left(h, \frac{1}{2}h\right) = \frac{1}{3}\left[4I\left(\frac{1}{2}h\right) - I(h)\right], \tag{6.49}$$

where $I(h) = I_1$,

$$I\left(\frac{1}{2}h\right) = I_2 \quad \text{and} \quad I\left(h, \frac{1}{2}h\right) = I_3.$$

With this notation the following table can be formed

$I(h)$			
	$I\left(h, \frac{1}{2}h\right)$		
		$I\left(h, \frac{1}{2}h, \frac{1}{4}h\right)$	
$I\left(\frac{1}{2}h\right)$			$I\left(h, \frac{1}{2}h, \frac{1}{4}h, \frac{1}{8}h\right)$
	$I\left(\frac{1}{2}h, \frac{1}{4}h\right)$		
$I\left(\frac{1}{4}h\right)$		$I\left(\frac{1}{2}h, \frac{1}{4}h, \frac{1}{8}h\right)$	
	$I\left(\frac{1}{4}h, \frac{1}{8}h\right)$		
$I\left(\frac{1}{8}h\right)$			

The computations can be stopped when two successive values are sufficiently close to each other. This method, due to L.F. Richardson, is called the *deferred approach to the limit* and the systematic tabulation of this is called *Romberg Integration*.

6.4.7 Newton–Cotes Integration Formulae

Let the interpolation points, x_i, be equally spaced, i.e. let $x_i = x_0 + ih$, $i = 0, 1, 2, \ldots, n$, and let the end points of the interval of integration be placed such that

$$x_0 = a. \quad x_n = b, \quad h = \frac{b-a}{n}.$$

Then the definite integral

$$I = \int_a^b y\, dx \tag{6.50}$$

is evaluated by an integration formula of the type

$$I_n = \sum_{i=0}^{n} C_i y_i, \tag{6.51}$$

where the coefficients C_i are determined completely by the abscissae x_i. Integration formulae of the type (6.51) are called *Newton–Cotes closed integration formulae*. They are 'closed' since the end points a and b are the extreme abscissae in the formulae. It is easily seen that the integration formulae derived in Eqs. (6.47)–(6.50) are the simplest Newton–Cotes closed formulae.

On the other hand, formulae which do not employ the end points are called Newton–Cotes, *open* integration formulae. We give below the five simplest Newton–Cotes open integration formulae

(a) $$\int_{x_0}^{x_2} y\,dx = 2hy_1 + \frac{h^3}{3} y''(\bar{x}),\ (x_0 < \bar{x} < x_2) \tag{6.52}$$

(b) $$\int_{x_0}^{x_3} y\,dx = \frac{3h}{2}(y_1 + y_2) + \frac{3h^3}{4} y''(\bar{x}),\ (x_0 < \bar{x} < x_3) \tag{6.53}$$

(c) $$\int_{x_0}^{x_4} y\,dx = \frac{4h}{3}(2y_1 - y_2 + 2y_3) + \frac{14}{45} h^5 y^{\text{iv}}(\bar{x}),\ (x_0 < \bar{x} < x_4) \tag{6.54}$$

(d) $$\int_{x_0}^{x_5} y\,dx = \frac{5h}{24}(11y_1 + y_2 + y_3 + 11y_4) + \frac{95}{144} h^5 y^{\text{iv}}(\bar{x}),\ (x_0 < \bar{x} < x_5) \tag{6.55}$$

(e) $$\int_{x_0}^{x_6} y\,dx = \frac{6h}{20}(11y_1 - 14y_2 + 26y_3 - 14y_4 + 11y_5) + \frac{41}{140} h^7 y^{\text{vi}}(\bar{x}),\quad (x_0 < \bar{x} < x_6). \tag{6.56}$$

A convenient method for determining the coefficients in the Newton–Cotes formulae is the method of undetermined coefficients. This is demonstrated in Example 6.14.

Example 6.8 Find, from the following table, the area bounded by the curve and the x-axis from $x = 7.47$ to $x = 7.52$

x	$f(x)$	x	$f(x)$
7.47	1.93	7.50	2.01
7.48	1.95	7.51	2.03
7.49	1.98	7.52	2.06

We know that

$$\text{Area} = \int_{7.47}^{7.52} f(x)\, dx$$

with h = 0.01, the trapezoidal rule given in Eq. (6.32) gives

$$\text{Area} = \frac{0.01}{2}[1.93 + 2(1.95 + 1.98 + 2.01 + 2.03) + 2.06] = 0.0996.$$

Example 6.9 A solid of revolution is formed by rotating about the x-axis the area between the x-axis, the lines $x = 0$ and $x = 1$, and a curve through the points with the following coordinates:

x	y
0.00	1.0000
0.25	0.9896
0.50	0.9589
0.75	0.9089
1.00	0.8415

Estimate the volume of the solid formed, giving the answer to three decimal places.

If V is the volume of the solid formed, then we know that

$$V = \pi \int_0^1 y^2 dx$$

Hence we need the values of y^2 and these are tabulated below, correct to four decimal places

x	y^2
0.00	1.0000
0.25	0.9793
0.50	0.9195
0.75	0.8261
1.00	0.7081

With h = 0.25, Simpson's rule gives

$$V = \frac{\pi(0.25)}{3}[1.0000 + 4(0.9793 + 0.8261) + 2(0.9195) + 0.7081]$$

$$= 2.8192.$$

Example 6.10 Evaluate

$$I = \int_0^1 \frac{1}{1+x}\,dx,$$

correct to three decimal places.

We solve this example by both the trapezoidal and Simpson's rules with $h = 0.5, 0.25$ and 0.125 respectively.

(i) $h = 0.5$: The values of x and y are tabulated below:

x	y
0.0	1.0000
0.5	0.6667
1.0	0.5000

(a) Trapezoidal rule gives

$$I = \frac{1}{4}[1.0000 + 2(0.6667) + 0.5] = 0.70835.$$

(b) Simpson's rule gives

$$I = \frac{1}{6}[1.0000 + 4(0.6667) + 0.5] = 0.6945.$$

(ii) $h = 0.25$: The tabulated values of x and y are given below:

x	y
0.00	1.0000
0.25	0.8000
0.50	0.6667
0.75	0.5714
1.00	0.5000

(a) Trapezoidal rule gives

$$I = \frac{1}{8}[1.0 + 2(0.8000 + 0.6667 + 0.5714) + 0.5] = 0.6970.$$

(b) Simpson's rule gives

$$I = \frac{1}{12}[1.0 + 4(0.8000 + 0.5714) + 2(0.6667) + 0.5] = 0.6932.$$

(iii) Finally, we take $h = 0.125$: The tabulated values of x and y are

x	y	x	y
0	1.0	0.625	0.6154
0.125	0.8889	0.750	0.5714
0.250	0.8000	0.875	0.5333
0.375	0.7273	1.0	0.5
0.5	0.6667		

(a) Trapezoidal rule gives

$$I = \frac{1}{16}[1.0 + 2(0.8889 + 0.8000 + 0.7273 + 0.6667) + 0.6154 + 0.5714 + 0.5333) + 0.5]$$

$$= 0.6941.$$

(b) Simpson's rule gives

$$I = \frac{1}{24}[1.0 + 4(0.8889 + 0.7273 + 0.6154 + 0.5333) + 2(0.8000 + 0.6667 + 0.5714) + 0.5]$$

$$= 0.6932.$$

Hence the value of I may be taken to be equal to 0.693, correct to three decimal places. The exact value of I is $\log_e 2$, which is equal to 0.693147…. This example demonstrates that, in general, Simpson's rule yields more accurate results than the trapezoidal rule.

Example 6.11 Use Romberg's method to compute

$$I = \int_0^1 \frac{1}{1+x}\,dx,$$

correct to three decimal places.

We take $h = 0.5$, 0.25 and 0.125 successively and use the results obtained in the previous example. We therefore have

$$I(h) = 0.7084, \quad I\left(\frac{1}{2}h\right) = 0.6970, \quad \text{and} \quad I\left(\frac{1}{4}h\right) = 0.6941$$

Hence, using Eq. (6.49), we obtain

$$I\left(h, \frac{1}{2}h\right) = 0.6970 + \frac{1}{3}(0.6970 - 0.7084) = 0.6932.$$

$$I\left(\frac{1}{2}h, \frac{1}{4}h\right) = 0.6941 + \frac{1}{3}(0.6941 - 0.6970) = 0.6931$$

Finally,

$$I\left(h, \frac{1}{2}h, \frac{1}{4}h\right) = 0.6931 + \frac{1}{3}(0.6931 - 0.6932) = 0.6931.$$

The table of values is, therefore,

0.7084		
	0.6932	
0.6970		0.6931
	0.6931	
0.6941		

An obvious advantage of this method is that the accuracy of the computed value is known at each step.

Example 6.12 Apply trapezoidal and Simpson's rules to the integral

$$I = \int_0^1 \sqrt{1 - x^2}\, dx$$

continually halving the interval h for better accuracy.

Using 10, 20, 30, 40 and 50 subintervals successively, an electronic computer, with a nine decimal precision, produced the results given in Table below. The true value of the integral is $\pi/4 = 0.785\,398\,163$.

No. of subintervals	*Trapezoidal rule*	*Simpson's's rule*
10	0.776 129 582	0.781 752 040
20	0.782 116 220	0.784 111 766
30	0.783 610 789	0.784 698 434
40	0.784 236 934	0.784 943 838
50	0.784 567 128	0.785 073 144

Example 6.13 Evaluate

$$I = \int_0^1 \sin \pi x\, dx$$

using the cubic spline method.

The exact value of I is $2/\pi$ = 0.63661978. To make the calculations easier, we take n = 2, i.e. h = 0.5. In this case, the table of values of x and y = sin px is

x	y
0	0
0.5	1.0
1.0	0.0

Using Eq. (5.32) with $M_0 = M_2 = 0$, we obtain $M_1 = -12$. Then formula (6.44) gives

$$I = \frac{1}{4}(y_0 + y_1) - \frac{1}{192}(M_0 + M_1) + \frac{1}{4}(y_1 + y_2) - \frac{1}{192}(M_1 + M_2)$$

$$= \frac{1}{4} + \frac{1}{16} + \frac{1}{4} + \frac{1}{16}$$

$$= \frac{5}{8}$$

$$= 0.62500000;$$

which shows that the absolute error in the natural spline solution is 0.01161978. It is easily verified that the Simpson's rule gives a value with an absolute error 0.03004689, which is more than the error in the spline solution.

Example 6.14 Derive Simpson's 1/3-rule using the method of undetermined coefficients.

We assume the formula

$$\int_{-h}^{h} y\,dx = a_{-1}y_{-1} + a_0 y_0 + a_1 y_1, \tag{i}$$

where the coefficients a_{-1}, a_0 and a_1 have to be determined. For this, we assume that formula (i) is exact when $y(x)$ is 1, x or x^2. Putting, therefore, $y(x) = 1$, x and x^2 successively in (i), we obtain the relations

$$a_{-1} + a_0 + a_1 = \int_{-h}^{h} dx = 2h, \tag{ii}$$

$$-a_{-1} + a_1 = \int_{-h}^{h} x\,dx = 0 \tag{iii}$$

and

$$a_{-1} + a_1 = \frac{2}{3}h. \tag{iv}$$

Solving (ii), (iii) and (iv) for a_{-1}, a_0 and a_1, we obtain

$$a_{-1} = \frac{2}{3} = a_1 \quad \text{and} \quad a_0 = \frac{4h}{3}.$$

Hence formula (i) takes the form

$$\int_{-h}^{h} y\,dx = \frac{h}{3}(y_{-1} + 4y_0 + y_1),$$

which is the Simpson's 1/3-rule given in Section 6.4.2.

6.5 EULER–MACLAURIN FORMULA

Consider the expansion of $1/(e^x - 1)$ in ascending powers of x, obtained by writing the Maclaurin expansion of e^x and simplifying

$$\frac{1}{e^x - 1} = \frac{1}{x} - \frac{1}{2} + B_1 x + B_3 x^3 + B_5 x^5 + \cdots, \tag{6.57}$$

where

$$B_{2r} = 0, \qquad B_1 = \frac{1}{12}, \qquad B_3 = -\frac{1}{720,} \qquad B_5 = \frac{1}{30,240}, \quad \text{etc.}$$

In Eq. (6.57), if we set $x = hD$ and use the relation $E \equiv e^{hD}$ (see Section 3.3.4), we obtain the identity

$$\frac{1}{E-1} \equiv \frac{1}{hD} - \frac{1}{2} + B_1 hD + B_3 h^3 D^3 + B_5 h^5 D^5 + \cdots$$

or equivalently

$$\frac{E^n - 1}{E-1} = \frac{1}{hD}(E^n - 1) - \frac{1}{2}(E^n - 1) + B_1 hD(E^n - 1) + B_3 h^3 D^3 (E^n - 1) + \cdots \tag{6.58}$$

Operating this identity on y_0, we obtain

$$\begin{aligned}\frac{E^n - 1}{E-1} y_0 &= \frac{1}{hD}(E^n - 1)y_0 - \frac{1}{2}(E^n - 1)y_0 + B_1 hD(E^n - 1)y_0 + \cdots \\ &= \frac{1}{hD}(y_n - y_0) - \frac{1}{2}(y_n - y_0) + B_1 h(y_n' - y_0') + B_3 h^3 (y_n''' - y_0''') \\ &\quad + B_5 h^5 (y_n^{\text{v}} - y_0^{\text{v}}) + \cdots \end{aligned} \tag{6.59}$$

It can be easily shown that the left-hand side denotes the sum $y_0 + y_1 + y_2 + \cdots + y_{n-1}$, whereas the term

$$\frac{1}{hD}(y_n - y_0)$$

on the right side can be written as

$$\frac{1}{h}\int_{x_0}^{x_n} y\,dx$$

since $1/D$ can be interpreted as an integration operator.

Hence, Eq. (6.59) becomes

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n) - \frac{h^2}{12}(y_n' - y_0')$$

$$+ \frac{h^4}{720}(y_n''' - y_0''') - \frac{h^6}{30,240}(y_n^{\text{v}} - y_0^{\text{v}}) + \cdots \tag{6.60}$$

which is called the *Euler–Maclaurin's formula* for integration. The first expression on the right-hand side of Eq. (6.60) denotes the approximate value of the integral obtained by using trapezoidal rule and the other expressions represent the successive *corrections* to this value. It should be noted that this formula may also be used to find the sum of a series of the form $y_0 + y_1 + y_2 + \cdots + y_n$. The use of this formula is illustrated by the following examples.

Example 6.15 Evaluate

$$I = \int_0^{\pi/2} \sin x\,dx$$

using the Euler–Maclaurin's formula.

In this case, formula (6.60) simplifies to

$$\int_0^{\pi/2} \sin x\,dx = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n) + \frac{h^2}{12} + \frac{h^4}{720} + \frac{h^6}{30,240} + \cdots \tag{i}$$

To evaluate the integral, we take $h = \pi/4$. Then we obtain

$$\int_0^{\pi/2} \sin x\,dx = \frac{\pi}{8}(0 + 2 + 0) + \frac{\pi^2}{192} + \frac{\pi^4}{1,84,320} + \cdots$$

$$= \frac{\pi}{4} + \frac{\pi^2}{192} + \frac{\pi^4}{1,84,320}, \text{ approximately}$$

$$= 0.785398 + 0.051404 + 0.000528$$

$$= 0.837330.$$

On the other hand with $h = \pi/8$, we obtain

$$\int_0^{\pi/2} \sin x\,dx = \frac{\pi}{16}[(0 + 2(0.382683) + .707117 + 0.923879 + 1.000000)]$$

$$= 0.987119 + 0.012851 + 0.000033$$

$$= 1.000003.$$

Example 6.16 Use the Euler–Maclaurin formula to prove

$$\sum_{1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}.$$

In this case, rewrite Eq. (6.60) as

$$\frac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n = \frac{1}{h}\int_{x_0}^{x_n} y\,dx + \frac{h}{12}(y_n' - y_0') - \frac{h^3}{720}(y_n''' - y_0''')$$

$$+ \frac{h^5}{30,240}(y_n^{\text{v}} - y_0^{\text{v}}) - \cdots \qquad \text{(i)}$$

Here $y(x) = x^2$, $y'(x) = 2x$ and $h = 1$.

Hence Eq. (i) gives

$$\text{Sum} = \int_{1}^{n} x^2\,dx + \frac{1}{2}(n^2+1) + \frac{1}{12}(2n-2)$$

$$= \frac{1}{3}(n^3-1) + \frac{1}{2}(n^2+1) + \frac{1}{6}(n-1)$$

$$= \frac{1}{6}(2n^3 + 3n^2 + n)$$

$$= \frac{n(n+1)(2n+1)}{6}.$$

6.6 NUMERICAL INTEGRATION WITH DIFFERENT STEP SIZES

We have so far considered integration formulae which use equally spaced abscissae. In practical problems, however, we often come across situations which require the use of different step-sizes while solving a problem. This would be so if the interval in question contains parts over which the function varies too rapidly or too slowly. For better accuracy and efficiency, it would be desirable to take a smaller size in parts of the interval over which the function variation is large. Similarly, it would be efficient to take larger step sizes over parts in which the function varies too slowly. A numerical integration procedure which *adopts* automatically a suitable step-size to solve an integration problem numerically is called *adaptive quadrature method*. We describe below an 'adaptive quadrature method' based on Simpson's (1/3)-rule.

Suppose that we wish to approximate the integral

$$I = \int_{a}^{b} y(x)\,dx \qquad (6.61)$$

to within an accuracy $\varepsilon > 0$. Using Simpson's (1/3)-rule with $h = (b - a)/2$, we obtain

$$I = \int_a^b y(x)\,dx \approx \frac{h}{3}\left[y(a) + 4y\left(\frac{a+b}{2}\right) + y(b)\right] - \frac{h^5}{90}y^{\text{iv}}(\xi_1), \quad a < \xi_1 < b$$

$$= I(a, b) - \frac{(b-a)h^4}{180}y^{\text{iv}}(\xi_1), \tag{6.62}$$

where

$$I(a, b) = \frac{h}{3}\left[y(a) + 4y\left(\frac{a+b}{2}\right) + y(b)\right]. \tag{6.63}$$

Now, we subdivide the interval and set $h = (b - 1)/4$. Simpson's (1/3)-rule then gives

$$I = \int_a^b y(x)\,dx = \frac{h}{6}\left[y(a) + 4y\left(\frac{3a+b}{4}\right) + 2y\left(\frac{a+b}{2}\right) + 4y\left(\frac{a+3b}{4}\right) + y(b)\right]$$

$$- \frac{h^4(b-a)}{180 \times 16}y^{\text{iv}}(\xi_2)$$

$$= \frac{h}{6}\left[y(a) + 4y\left(\frac{3a+b}{4}\right) + y\left(\frac{a+b}{2}\right)\right]$$

$$+ \frac{h}{6}\left[y\left(\frac{a+b}{2}\right) + 4y\left(\frac{a+3b}{4}\right) + y(b)\right] - \frac{(b-a)h^4}{180 \times 16}y^{\text{iv}}(\xi_2)$$

$$= I\left(a, \frac{a+b}{2}\right) + I\left(\frac{a+b}{2}, b\right) - \frac{(b-a)h^4}{180 \times 16}y^{\text{iv}}(\xi_2), \tag{6.64}$$

where

$$I\left(a, \frac{a+b}{2}\right) = \frac{h}{6}\left[y(a) + 4y\left(\frac{3a+b}{4}\right) + y\left(\frac{a+b}{2}\right)\right] \tag{6.65a}$$

and

$$I\left(\frac{a+b}{2}, b\right) = \frac{h}{6}\left[y\left(\frac{a+b}{2}\right) + 4y\left(\frac{a+3b}{4}\right) + y(b)\right]. \tag{6.65b}$$

Assuming

$$y^{\text{iv}}(\xi_1) = y^{\text{iv}}(\xi_2)$$

Eqs. (6.62) and (6.64) give on simplification

$$\frac{1}{15}\left[I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right]=\frac{(b-a)h^4}{180\times 16}y^{\text{iv}}(\xi_2). \qquad (6.66)$$

Substituting Eq. (6.66) in Eq. (6.64), we obtain an estimate for the error, viz.

$$\left|\int_a^b y(x)\,dx-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|$$

$$=\frac{1}{15}\left|I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|. \qquad (6.67)$$

If we suppose

$$\frac{1}{15}\left|I(a,b)-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|<\varepsilon \qquad (6.68)$$

for some $\varepsilon > 0$ in the interval $[a, b]$, then Eq. (6.67) means that

$$\left|\int_a^b y(x)\,dx-I\left(a,\frac{a+b}{2}\right)-I\left(\frac{a+b}{2},b\right)\right|<\varepsilon \qquad (6.69)$$

and that

$$\int_a^b y(x)\,dx\approx I\left(a,\frac{a+b}{2}\right)+I\left(\frac{a+b}{2},b\right) \qquad (6.70)$$

to within an accuracy of $\varepsilon > 0$.

If the inequality (6.68) is not satisfied, then the procedure is applied to each of the intervals $[a, (a + b)/2]$ and $[(a + b)/2, b]$ with the tolerance $\varepsilon/2$. If the inequality is satisfied in *both* the intervals, then the sum of the two approximations will give an approximation to the given integral. If *the test fails* in any of the intervals, then that particular interval is subdivided in to 'two subintervals' and the above procedure is applied with a tolerance which is half of the previous tolerance. The following example demonstrates the testing procedure.

Example 6.17 Test the error estimate given by Eq. (6.67) in the evaluation of the integral

$$I=\int_0^{\pi/2}\cos x\,dx.$$

Let $h = \pi/4$. Then

$$I\left(0, \frac{\pi}{2}\right) = \frac{\pi}{12}\left(1 + \frac{4}{\sqrt{2}} + 0\right) = 1.00228.$$

Also

$$I\left(0, \frac{\pi}{4}\right) = \frac{\pi}{24}\left(1 + 4\cos\frac{\pi}{8} + \frac{1}{\sqrt{2}}\right)$$

and

$$I\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi}{24}\left(\frac{1}{\sqrt{2}} + 4\cos\frac{3\pi}{8} + 0\right).$$

Hence

$$I\left(0, \frac{\pi}{4}\right) + I\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi}{24}\left(1 + \sqrt{2} + 4\cos\frac{\pi}{8} + 4\cos\frac{3\pi}{8}\right) = 1.00013.$$

It follows that

$$\frac{1}{15}\left| I\left(0, \frac{\pi}{2}\right) - I\left(0, \frac{\pi}{4}\right) - I\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right| = \frac{1}{15}(0.00215) = 0.00014.$$

It can be verified that the

$$\text{Actual error} = \left| \int_0^{\pi/2} \cos x \, dx - I\left(0, \frac{\pi}{4}\right) - I\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right| = 0.00013,$$

which is less than that obtained above.

Example 6.18 Test the error estimate given in Eq. (6.67) in the evaluation of the integral

$$I = \int_0^{\pi/2} (8 + 4\sin x)\, dx.$$

Exact value = $4(\pi + 1)$

Now,

$$I\left(0, \frac{\pi}{2}\right) = \frac{\pi}{12}\left[8 + 4\left(8 + \frac{4}{\sqrt{2}}\right) + 12\right] = \frac{\pi}{12}\left(52 + \frac{16}{\sqrt{2}}\right),$$

$$I\left(0, \frac{\pi}{4}\right) = \frac{\pi}{24}\left[8 + 4\left(8 + 4\sin\frac{\pi}{8}\right) + 8 + \frac{4}{\sqrt{2}}\right] = \frac{\pi}{24}\left(48 + 16\sin\frac{\pi}{8} + \frac{4}{\sqrt{2}}\right),$$

$$I\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi}{24}\left[8 + \frac{4}{\sqrt{2}} + 4\left(8 + 4\sin\frac{3\pi}{8}\right) + 12\right]$$

$$= \frac{\pi}{24}\left[52 + \frac{4}{\sqrt{2}} + 16\sin\frac{3\pi}{8}\right]$$

Therefore,

$$I\left(0, \frac{\pi}{4}\right) + I\left(\frac{\pi}{4}, \frac{\pi}{2}\right) = \frac{\pi}{24}\left[100 + \frac{8}{\sqrt{2}} + 16\sin\frac{\pi}{8} + 16\sin\frac{3\pi}{8}\right]$$

Hence

$$\frac{1}{15}\left[I\left(0, \frac{\pi}{2}\right) - I\left(0, \frac{\pi}{4}\right) - I\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right]$$

$$= \frac{1}{15}\frac{\pi}{24}\left[104 + \frac{32}{\sqrt{2}} - 100 - \frac{8}{\sqrt{2}} - 16\sin\frac{\pi}{8} - 16\sin\frac{3\pi}{8}\right]$$

$$= \frac{\pi}{360}\left[4 + \frac{24}{\sqrt{2}} - 16\sin\frac{\pi}{8} - 16\sin\frac{3\pi}{8}\right]$$

$$= 0.00057\,.$$

Actual Error $$= \left|\int_0^{\pi/2} (8 + 4\sin x)\,dx - I\left(0, \frac{\pi}{4}\right) - I\left(\frac{\pi}{4}, \frac{\pi}{2}\right)\right|$$

$$= \left|4(\pi + 1) - \frac{\pi}{24}\left(100 + \frac{8}{\sqrt{2}} + 16\sin\frac{\pi}{8} - 16\sin\frac{3\pi}{8}\right)\right|$$

$$= \left|16.56637 - 16.56691\right|$$

$$= 0.00054.$$

6.7 GAUSSIAN INTEGRATION

We consider the numerical evaluation of the integral

$$I = \int_a^b f(x)\,dx \tag{6.71}$$

In the preceding sections, we derived some integration formulae which require values of the function at equally-spaced points of the interval. Gauss derived a formula which uses the same number of function values but with different spacing and gives better accuracy.

Gauss' formula is expressed in the form

$$\int_{-1}^{1} F(u)du = W_1F(u_1) + W_2F(u_2) + \cdots + W_nF(u_n) = \sum_{i=1}^{n} W_iF(u_i) \tag{6.72}$$

where W_i and u_i are called the *weights* and *abscissae*, respectively.

In Eq. (6.72), there are altogether $2n$ arbitrary parameters and therefore the weights and abscissae can be determined such that the formula is *exact* when $F(u)$ is a polynomial of degree not exceeding $(2n-1)$. Hence, we start with

$$F(u) = c_0 + c_1 u + c_2 u^2 + c_3 u^3 + \cdots + c_{2n-1} u^{2n-1}. \tag{6.73}$$

We then obtain from Eq. (6.72)

$$\int_{-1}^{1} F(u)\,du = \int_{-1}^{1} (c_0 + c_1 u + c_2 u^2 + c_3 u^3 + \cdots + c_{2n-1} u^{2n-1})\,du$$

$$= 2c_0 + \frac{2}{3}c_2 + \frac{2}{5}c_4 + \cdots \tag{6.74}$$

Substituting these values on the right-hand side of Eq. (6.72), we obtain

$$\begin{aligned}\int_{-1}^{1} F(u)\,du = {} & W_1(c_0 + c_1 u_1 + c_2 u_1^2 + \cdots + c_{2n-1} u_1^{2n-1}) \\ & + W_2(c_0 + c_1 u_2 + c_2 u_2^2 + \cdots + c_{2n-1} u_2^{2n-1}) \\ & + W_3(c_0 + c_1 u_3 + c_2 u_3^2 + \cdots + c_{2n-1} u_3^{2n-1}) + \cdots \\ & + W_n(c_0 + c_1 u_n + c_2 u_n^2 + \cdots + c_{2n-1} u_n^{2n-1}),\end{aligned}$$

which can be written as

$$\begin{aligned}\int_{-1}^{1} F(u)\,du = {} & c_0(W_1 + W_2 + \cdots + W_n) \\ & + c_1(W_1 u_1 + W_2 u_2 + W_3 u_3 + \cdots + W_n u_n) \\ & + c_2(W_1 u_1^2 + W_2 u_2^2 + W_3 u_3^2 + \cdots + W_n u_n^2) + \cdots \\ & + c_{2n-1}(W_1 u_1^{2n-1} + W_2 u_2^{2n-1} + W_3 u_3^{2n-1} + \cdots + W_n u_n^{2n-1}).\end{aligned} \tag{6.75}$$

Now, Eqs. (6.74) and (6.75) are identical for all values of c_i and hence comparing the coefficients of c_i, we obtain the $2n$ equations

$$\left.\begin{aligned} W_1 + W_2 + W_3 + \cdots + W_n &= 2 \\ W_1 u_1 + W_2 u_2 + W_3 u_3 + \cdots + W_n u_n &= 0 \\ W_1 u_1^2 + W_2 u_2^2 + W_3 u_3^2 + \cdots + W_n u_n^2 &= 2/3 \\ &\vdots \\ W_1 u_1^{2n-1} + W_2 u_2^{2n-1} + W_3 u_3^{2n-1} + \cdots + W_n u_n^{2n-1} &= 0 \end{aligned}\right\} \tag{6.76}$$

in $2n$ unknowns W_i and u_i $(i = 1, 2, \ldots, n)$.

As an illustration, we consider the case $n = 2$. Then the formula is

$$\int_{-1}^{1} F(u)\, du = W_1 F(u_1) + W_2 F(u_2). \tag{6.77}$$

Since this formula is exact when $F(u)$ is a polynomial of degree not exceeding 3, we put successively $F(u) = 1$, u, u^2 and u^3. Then Eq. (6.77) gives the four equations:

$$\left.\begin{aligned} W_1 + W_2 &= 2 \\ W_1 u_1 + W_2 u_2 &= 0 \\ W_1 u_1^2 + W_2 u_2^2 &= 2/3 \\ W_1 u_1^3 + W_2 u_2^3 &= 0. \end{aligned}\right\} \tag{6.78}$$

The solution of these equations is

$$W_1 = W_2 = 1, \qquad u_2 = -u_1 = \frac{1}{\sqrt{3}}. \tag{6.79}$$

This method, when applied to the general system given in Eq. (6.76) above, will be extremely complicated and difficult, and an alternate method must be chosen to solve the nonlinear system (6.76).

It can be shown that the u_i are the zeros of the $(n + 1)$th Legendre polynomial $P_{n+1}(u)$ which can be generated using the recurrence relation

$$(n+1)P_{n+1}(u) = (2n+1)uP_n(u) - nP_{n-1}(u), \tag{6.80}$$

where $P_0(u) = 1$ and $P_1(u) = u$. The first-five Legendre polynomials are given by

$$\left.\begin{aligned} P_0(u) &= 1 \\ P_1(u) &= u \\ P_2(u) &= (1/2)\,(3u^2 - 1) \\ P_3(u) &= (1/2)\,(5u^3 - 3u) \\ P_4(u) &= (1/8)\,(35u^4 - 30u^2 + 3). \end{aligned}\right\} \tag{6.81}$$

It can also be shown that the corresponding weights W_i are given by

$$W_i = \int_{-1}^{1} \prod_{j=0,\, j\neq i}^{n} \frac{u - u_j}{u_i - u_j}\, du, \tag{6.82}$$

where the u_i are the *abscissae*.

As an example, when $n = 1$ we solve $P_2(u) = 0$, i.e.

$$\frac{1}{2}(3u^2 - 1) = 0,$$

which gives the *two abscissae*:

$$u_0 = -\frac{1}{\sqrt{3}} = -\frac{\sqrt{3}}{3} \quad \text{and} \quad u_1 = \frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{3}.$$

The corresponding weights are given by

$$W_0 = \int_{-1}^{1} \frac{u - u_1}{u_0 - u_1}\, du = \frac{1}{u_0 - u_1}\left[\frac{u^2}{2} - u_1 u\right]_{-1}^{1} = 1$$

and

$$W_1 = \int_{-1}^{1} \frac{u - u_0}{u_1 - u_0}\, du = \frac{1}{u_1 - u_0}\left[\frac{u^2}{2} - u_0 u\right]_{-1}^{1} = 1.$$

Similarly, for $n = 3$ we solve $P_4(u) = 0$. That is,

$$\frac{1}{8}(35u^4 - 30u^2 + 3) = 0,$$

which gives the four abscissae:

$$u_i = \pm\left(\frac{15 \pm 2\sqrt{30}}{35}\right)^{1/2}.$$

The weights W_i can then be obtained from Eq. (6.82). It should be noted, however, that the abscissae u_i and the weights W_i are extensively tabulated for different values of n. We list below, in Table 6.1, the abscissae and weights for values of n up to $n = 6$.

Table 6.1 Abscissae and Weights for Gaussian Integration

n	$\pm u_i$	W_i
2	0.57735 02692	1.0
3	0.0	0.88888 88889
	0.77459 66692	0.55555 55556
4	0.33998 10436	0.65214 51549
	0.86113 63116	0.34785 48451
5	0.0	0.56888 88889
	0.53846 93101	0.47862 86705
	0.90617 98459	0.23692 68851
6	0.23861 91861	0.46791 39346
	0.66120 93865	0.36076 15730
	0.93246 95142	0.17132 44924

In the general case, the limits of the integral in Eq. (6.71) have to be changed to those in Eq. (6.72) by means of the transformation

$$x = \frac{1}{2} u (b - a) + \frac{1}{2} (a + b). \tag{6.83}$$

The use of Table 6.1 is illustrated by the following example:

Example 6.19 Find $I = \int_0^1 x\, dx$, by Gauss' formula.

The first step is to change the limits by Eq. (6.83). So, we get

$$x = \frac{1}{2} (u + 1)$$

This gives

$$I = \frac{1}{4} \int_{-1}^{1} (u + 1)\, du = \frac{1}{4} \sum_{i=1}^{n} W_i F(u_i),$$

where $F(u_i) = u_i + 1$.

For simplicity, we take $n = 4$ and using the 'abscissae and weights' corresponding to $n = 4$ in Table 6.1, we obtain

$$\begin{aligned} I &\approx \frac{1}{4} [(-0.86114 + 1)(0.34785) + (-0.33998 + 1)(0.65214) \\ &\quad + (0.33998 + 1)(0.65214) + (0.86114 + 1)(0.34785)] \\ &= 0.49999\ldots, \end{aligned}$$

where the abscissae and weights have been rounded to five decimal places.

6.8 GENERALIZED QUADRATURE

In evaluating singular integrals which arise in practical applications, it will often be convenient to develop special integration formulae.

We consider, for instance, the numerical quadrature of integrals of the form

$$I(s) = \int_a^b f(t)\, \phi(t - s)\, dt, \tag{6.84}$$

where $f(t)$ is continuous but $\phi(u)$ may have an integrable singularity, e.g. $\log |s - t|$ or $|s - t|^{\alpha}$ for $\alpha > -1$. For the numerical integration, we divide the range (a, b) such that $t_j = a + jh$ ($j = 0, 1, 2, \ldots, n$), with $b = a + nh$. Then Eq. (6.84) can be written as

$$I(s) = \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} f(t)\, \phi(t - s)\, dt. \tag{6.85}$$

The method to be followed here is to approximate $f(t)$ in (6.85) by the linear interpolating function $f_n(t)$, where

$$f_n(t) = \frac{1}{h}[(t_{j+i} - t) f(t_j) + (t - t_j) f(t_{j+1})]. \tag{6.86}$$

Substituting $f_n(t)$ for $f(t)$ in (6.85), we obtain

$$I(s) = \frac{1}{h} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} [(t_{j+1} - t)\ f(t_j) + (t - t_j)\ f(t_{j+1})] \phi(t - s)\, dt.$$

Setting $t = t_j + ph$ this becomes

$$I(s) = h \sum_{j=0}^{n-1} \int_0^1 [(1 - p) f(t_j) + pf(t_{j+1})]\, \phi\,(t_j + ph - s)\, dp,$$

which can be written as

$$I(s) = h \sum_{j=0}^{n-1} [\alpha_j f(t_j) + \beta_j f(t_{j+1})], \tag{6.87}$$

where

$$\alpha_j = h \int_0^1 (1 - p)\, \phi\,(t_j + ph - s)\, dp \tag{6.88a}$$

and

$$\beta_j = h \int_0^1 p\phi(t_j + ph - s)\, dp. \tag{6.88b}$$

It is clear from Eqs. (6.88a) and (6.88b) that if $\phi(u) \equiv 1$, then $\alpha_j = \beta_j = h/2$, and hence Eq. (6.87) gives

$$I(s) = \frac{h}{2}[f(t_0) + 2f(t_1) + 2f(t_2) + \cdots + 2f(t_{n-1}) + f(t_n)],$$

which is the trapezoidal rule deduced in Section 6.4.1. Hence the rule defined by Eqs. (6.87), (6.88a) and (6.88b) is called the *generalized trapezoidal* rule and is due to Atkinson [1967]. When $\phi(u) = \log|u|$, this rule finds important applications in the numerical solution of certain singular integral equations.* In practice, the computation of the weights α_j and β_j may be difficult, but they can be evaluated once and for all, for a given $\phi(u)$.

In a similar way, one can deduce the *generalized Simpson's* rule—analogous to the ordinary Simpson's rule—by approximating $f(t)$ by means of a quadratic in the interval (t_j, t_{j+1}).**

---

* See, for example, Sastry [1973; 1976].

** See, Noble [1964], p. 241.

The error in generalized quadrature can also be estimated by the method outlined in Section 6.4.1. For example, it can be shown that the error in the generalized trapezoidal rule is of order h^2, assuming that f'' is continuous in $[a, b]$.

6.9 NUMERICAL CALCULATION OF FOURIER INTEGRALS

We consider, in this section, the problem of computing integrals which involve oscillatory functions, i.e., integrals of the form

$$I_c = \int_a^b f(x) \cos \omega x \, dx \tag{6.89}$$

and

$$I_s = \int_a^b f(x) \sin \omega x \, dx \tag{6.90}$$

Such integrals, called the *Fourier integrals*, occur in practical applications, e.g. *spectral analysis*. Einarsson [1972] described three methods for the numerical integration of these integrals, namely, the trapezoidal method, Filon's formula [1928] and cubic spline method. Since the derivations of Filon's formula and the cubic spline solution are quite involved, we omit them here but refer the reader to Einarsson's paper. In this section, we consider the evaluation of

$$I = \int_0^\infty e^{-x} \cos \omega x \, dx = \frac{1}{1+\omega^2} \tag{6.91}$$

by the trapezoidal rule. Using this rule, we obtain

$$I = \frac{h}{2}[1 + 2(e^{-h} \cos \omega h + e^{-2h} \cos 2\omega h + e^{-3h} \cos 3\omega h + \cdots)]$$

$$= \frac{h}{2} + h \sum_{n=1}^{\infty} e^{-nh} \cos \omega nh$$

$$= \frac{h}{2} + h \sum_{n=1}^{\infty} e^{-nh} \operatorname{Re}(e^{i\omega nh})$$

$$= h\left[\frac{1}{2} + \operatorname{Re} \sum_{n=1}^{\infty} e^{(i\omega - 1)nh}\right]$$

$$= h\left[\frac{1}{2} + \text{Re}\left\{\frac{e^{(-1+i\omega)h}}{1-e^{(-1+i\omega)h}}\right\}\right]$$

$$= h\left[\frac{1}{2} + \text{Re}\left\{\frac{e^{i\omega h}}{e^{h}-e^{i\omega h}}\right\}\right]$$

$$= \frac{h}{2}\cdot\frac{e^{2h}-1}{1+e^{2h}-2e^{h}\cos\omega h}, \text{ on simplification} \tag{6.92}$$

The right side of Eq. (6.92) can be further simplified by using hyperbolic functions.

With $h = 0.1$ and $\omega = 1$, formula (6.92) gives

$$I = \int_0^\infty e^{-x}\cos x\,dx$$

$$= \frac{0.1}{2}\cdot\frac{e^{0.2}-1}{1+e^{0.2}-2e^{0.1}\cos(0.1)} \qquad \left[\frac{1}{1+\omega^2} = \frac{1}{10} = 0.1, \omega = 3\right]$$

$$= \frac{0.1}{2}\cdot\frac{0.221402758}{2.221402758-2.199299334}$$

$$= 0.500833622,$$

so that the absolute error in the above result is 0.000833622. Table 6.2 gives the values of the integral given in Eq. (6.91) for different values of ω and $h = 0.1$. For comparison, exact values of the integral are also tabulated.

Table 6.2 Values of the Integral (6.91)

ω	*Exact value*	*Value obtained by using* (6.92)
1.0	0.500 000 000	0.500 833 622
3.0	0.100 000 000	0.100 836 955
6.0	0.027 027 027	0.027 875 426
9.0	0.012 195 121	0.013 063 154
12.0	0.006 896 552	0.007 793 302
20.0	0.002 493 766	0.003 524 142

6.10 NUMERICAL DOUBLE INTEGRATION

Formulae for the evaluation of a double integral can be obtained by repeatedly applying the trapezoidal and Simpson's rules derived in Sections 6.4.1 and 6.4.2. We consider, as an example, the double integral defined by

$$I = \int_{y_j}^{y_{j+1}} \int_{x_i}^{x_{i+1}} f(x, y)\, dx\, dy, \tag{6.93}$$

where

$$x_{i+1} = x_i + h \quad \text{and} \quad y_{j+1} = y_j + k.$$

By the repeated 'application of trapezoidal rule' to Eq. (6.93), we get

$$\begin{aligned} I &= \frac{h}{2} \int_{y_j}^{y_{j+1}} [f(x_i, y) + f(x_{i+1}, y)]\, dy \\ &= \frac{hk}{4} [f(x_i, y_j) + f(x_{i+1}, y_j) + f(x_i, y_{j+1}) + f(x_{i+1}, y_{j+1})] \\ &= \frac{hk}{4} [f_{i,j} + f_{i+1,j} + f_{i,j+1} + f_{i+1,j+1}], \end{aligned} \tag{6.94}$$

where $f_{i,j} = f(x_i, y_j)$, etc.

Similarly, applying Simpson's rule to the integral

$$I = \int_{y_{j-1}}^{y_{j+1}} \int_{x_{i-1}}^{x_{i+1}} f(x, y)\, dx\, dy, \tag{6.95}$$

we obtain

$$\begin{aligned} I &= \frac{h}{3} \int_{y_{j-1}}^{y_{j+1}} [f(x_{i-1}, y) + 4 f(x_i, y) + f(x_{i+1}, y)]\, dy \\ &= \frac{hk}{9} [f(x_{i-1}, y_{j-1}) + 4 f(x_{i-1}, y_j) + f(x_{i-1}, y_{j+1}) \\ &\qquad + 4\{f(x_i, y_{j-1}) + 4 f(x_i, y_j) + f(x_i, y_{j+1})\} \\ &\qquad + f(x_{i+1}, y_{j-1}) + 4 f(x_{i+1}, y_j) + f(x_{i+1}, y_{j+1})] \\ &= \frac{hk}{9} [f_{i-1,j-1} + f_{i-1,j+1} + f_{i+1,j-1} + f_{i+1,j+1} \\ &\qquad + 4(f_{i-1,j} + f_{i,j-1} + f_{i,j+1} + f_{i+1,j}) + 16 f_{i,j}]. \end{aligned} \tag{6.96}$$

A numerical example is given here.

Example 6.20 Evaluate

$$I = \int_0^1 \int_0^1 e^{x+y}\, dx\, dy,$$

using the trapezoidal and Simpson's rules. With $h = k = 0.5$, we have the following table of values of e^{x+y}.

	x		
y	0	0.5	1.0
0	1	1.6487	2.7183
0.5	1.6487	2.7183	4.4817
1.0	2.7183	4.4817	7.3891

Using the 'trapezoidal rule' from Eq. (6.94) repeatedly, we obtain

$$I = \frac{0.25}{4}[1.0 + 4(1.6487) + 6(2.7183) + 4(4.4817) + 7.3891]$$
$$= \frac{12.3050}{4}$$
$$= 3.0762.$$

Using 'Simpson's rule' given in Eq. (6.96) repeatedly, we obtain

$$I = \frac{0.25}{9}[1.0 + 2.7183 + 7.3891 + 2.7183$$
$$+ 4(1.6487 + 4.4817 + 4.4817 + 1.6487) + 16(2.7183)]$$
$$= \frac{26.59042}{9}$$
$$= 2.9545.$$

The 'exact value of the double integral is 2.9525' and therefore it can be verified that the result given by *Simpson's rule* is about sixty times more accurate than that given by the *trapezoidal rule*.

EXERCISES

6.1 Find $\frac{d}{dx} J_0(x)$ at $x = 0.1$ from the following table:

(0, 1.0), (0.1, 0.9975), (0.2, 0.9900), (0.3, 0.9776), (0.4, 0.9604).

6.2 The following table gives angular displacements θ (in radians) at different times t (seconds):
(0, 0.052), (0.02, 0.105), (0.04, 0.168), (0.06, 0.242), (0.08, 0.327), (0.10, 0.408), (0.12, 0.489).
Calculate the angular velocity at $t = 0.06$.

6.3 From the following values of x and y, find $\frac{dy}{dx}$ at $x = 0.6$.

(0.4, 1.5836), (0.5, 1.7974), (0.6, 2.0442), (0.7, 2.3275), (0.8, 2.6511).

6.4 The distances (x cm) traversed by a particle at different times (t seconds) are given below.

t	0.0	0.1	0.2	0.3	0.4	0.5	0.6
x	3.01	3.16	3.29	3.36	3.40	3.38	3.32

Find the velocity of the particle at $t = 0.3$ seconds.

6.5 From the following values of x and y, find $\frac{dy}{dx}$ when (a) $x = 1$, (b) $x = 3$, (c) $x = 6$ and (d) $\frac{d^2y}{dx^2}$ at $x = 3$.

x	0	1	2	3	4	5	6
y	6.9897	7.4036	7.7815	8.1291	8.4510	8.7506	9.0309

6.6 A rod is rotating in a plane about one of its ends. The angle θ (in radians) at different times t (seconds) are given below.

t	0	0.2	0.4	0.6	0.8	1.0
θ	0.0	0.15	0.50	1.15	2.0	3.20

Find its angular velocity and angular acceleration when $t = 0.6$ seconds.

6.7 Tabulate the function $y = f(x) = x^3 - 10x + 6$ at $x_0 = -0.5$, $x_1 = 1.00$ and $x_2 = 2.0$. Compute its first and second derivatives at $x = 1.00$ using Lagrange's interpolation formula. Compare your results with true values.

6.8 Given the following values of x and y, find $\frac{dy}{dx}$ at $x = 2$:

(0, 2), (2, –2), (3, –1).

6.9 A cubic function $y = f(x)$ satisfies the following data:

x	0	1	3	4
$f(x)$	1	4	40	85

Determine $f(x)$ and hence find $f'(2)$ and $f''(2)$.

6.10 The function $y = 3xe^{-x}$ is tabulated below.

(3, 0.4481), (4, 0.2198), (5, 0.1011).

Find $y'(x)$ at $x = 3$, 4 and 5 and compare your results with the exact values.

6.11 From the following values of x and y, find $\frac{dy}{dx}$ at $x = 2$ using the cubic spline method.

(2, 11), (3, 49), (4, 123)

6.12 From the following values of x and y, determine the value of $\frac{dy}{dx}$ at each of the points by fitting a natural cubic spline through them.

(1, 3), (2, 11), (4, 69), (5, 131).

6.13 Given the values of x and y:

(1.2, 0.9320), (1.3, 0.9636), (1.4, 0.9855), (1.5, 0.9975), (1.6, 0.9996),

find x, correct to two decimal places, for which y is maximum and find this value of y.

6.14 If $y = A + Bx + Cx^2$ and y_0, y_1, y_2 are the values of y corresponding to $x = 0$, h and $2h$, respectively, prove that

$$\int_0^{2h} y\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2)\,.$$

6.15 Evaluate

(a) $\int_0^{\pi} x \sin x\,dx$ and (b) $\int_{-2}^{2} \frac{x}{5+2x}\,dx$

using the trapezoidal rule with five ordinates.

6.16 State the trapezoidal rule for finding an approximate area under a given curve. A curve is given by the points (x, y) given below.

(0, 23), (0.5, 19), (1.0, 14), (1.5, 11), (2.0, 12.5), (2.5, 16), (3.0, 19), (3.5, 20), (4.0, 20).

Estimate the area bounded by the curve, the x-axis and the extreme ordinates.

6.17 Write an algorithm to evaluate the integral

$$I = \int_{x_0}^{x_n} y\,dx$$

by the trapezoidal rule with step size h. Given the values of x and $y(x)$;

(0, 0.399), (0.5, 0.352), (1.0, 0.242) (1.5, 0.129), (2.0, 0.054)

find an approximate value of

$$\int_0^{2} y(x)\,dx.$$

6.18 Estimate the value of the integral

$$\int_1^3 \frac{1}{x}\, dx$$

by Simpson's rule with 4 strips and 8 strips respectively. Determine the error in each case.

6.19 Evaluate

$$I = \int_0^{\pi/2} \sqrt{\sin x}\, dx$$

using Simpson's $\frac{1}{3}$-rule with $h = \pi/12$.

6.20 Using Simpson's $\frac{1}{3}$-rule with $h = 1$, evaluate the integral

$$I = \int_3^7 x^2 \log x\, dx.$$

6.21 Write an algorithm to evaluate $\int_{x_0}^{x_{2n}} y\, dx$ using Simpson's $\frac{1}{3}$-rule when $y(x)$ is given at x_0, $x_0 + h$, …, $x_0 + 2nh$. Evaluate

$$\int_0^1 e^{-x^2} \sin x\, dx$$

using Simpson's $\frac{1}{3}$-rule with $h = 0.1$.

6.22 Compute the values of

$$I = \int_0^1 \frac{dx}{1 + x^2}$$

using the trapezoidal rule with $h = 0.5$, 0.25 and 0.125. Then obtain a better estimate using Romberg's method. Compare your results with the true value.

6.23 Determine the maximum error in evaluating the integral

$$I = \int_0^{\pi/2} \sin x\, dx$$

by both the trapezoidal and Simpson's $\frac{1}{3}$-rules using four subintervals.

6.24 Estimate the value of the integral

$$I = \int_0^{1/2} \frac{dx}{\sqrt{x}\,\sqrt{1-x}}$$

using the trapezoidal rule. What is its exact value?

6.25 Derive Simpson's $\frac{3}{8}$-rule,

$$\int_{x_0}^{x_3} y\,dx = \frac{3}{8}h\,(y_0 + 3y_1 + 3y_2 + y_3)$$

Using this rule, evaluate

$$\int_0^1 \frac{1}{1+x}\,dx$$

with $h = \frac{1}{6}$. Evaluate the integral by Simpson's $\frac{1}{3}$-rule and compare the results.

6.26 Using the method of undetermined coefficients, derive the formula

$$\int_0^h y(x)\,dx = \frac{h}{2}(y_0 + y_1) + \frac{h^2}{12}(y_0' - y_1')$$

6.27 The function $y = e^x$ satisfies the following data which is unequally spaced:

(1.00, 2.7183), (1.05, 2.8577), (1.10, 3.0042),
(1.15, 3.1582), (1.25, 3.4903), (1.30, 3.6693).

Evaluate the integral

$$\int_{1.0}^{1.3} y\,dx,$$

as accurately as possible.

6.28 Evaluate

$$\int_0^2 \frac{dx}{x^3 + x + 1}$$

by Simpson's $\frac{1}{3}$-rule with $h = 0.25$.

6.29 Use Euler–Maclaurin formula to evaluate the integral

$$I = \int_1^2 (\cos x + \ln x - e^x)\, dx$$

6.30 Use Euler–Maclaurin formula to sum the series

$$S = 1^3 + 2^3 + 3^3 + \cdots + n^3.$$

6.31 Derive the Gauss integration formula when $n = 2$ and apply it to evaluate the integral

$$\int_{-1}^{1} \frac{1}{1+x^2}\, dx.$$

6.32 Use the three-point Gauss formula to evaluate the integral

$$I = \int_0^1 \frac{1}{1+x}\, dx.$$

Compare this result with that obtained by Simpson's $\frac{1}{3}$-rule with $h = 0.125$.

6.33 Let

$$\int_0^2 x^2 f(x)\, dx = k_0\ f(0) + k_1 f(1) + k_2 f(2).$$

If $f(x)$ is approximated by a quadratic through the points $x = 0$, 1, and 2, then find the values of k_0, k_1 and k_2.

6.34 Using the trapezoidal rule, show that

$$\int_0^\infty e^{-x} \cos \omega x\, dx \approx \frac{h}{2} \frac{\sinh h}{\cosh h - \cos \omega h} \qquad \text{(Einarsson)}$$

6.35 Einarsson [1972] derived the formula

$$\int_0^\infty e^{-x} \cos \omega x\, dx \approx \frac{h \sinh h}{\cosh 2h - \cos 2\omega h}\ (c_1 \cosh h + c_2 \cos \omega h),$$

where

$$c_1 = 2\left[\frac{1+\cos^2 \omega h}{\omega^2 h^2} - \frac{\sin 2\omega h}{\omega^3 h^3}\right]$$

and $$c_2 = 4\left[\frac{\sin \omega h}{\omega^3 h^3} - \frac{\cos \omega h}{\omega^2 h^2}\right]$$

use this formula to show that

$$\int_0^\infty e^{-x} \cos x \, dx = 0.500001391$$

6.36 If

$$I = \int_0^{\pi/4} \cos^2 x \, dx,$$

compute

$$I\left(0, \frac{\pi}{4}\right), \; I\left(0, \frac{\pi}{8}\right) \text{ and } I\left(\frac{\pi}{8}, \frac{\pi}{4}\right)$$

6.37 Verify the error estimate given in Eq. (6.67) for Problem 6.36.

6.38 Test the error estimate given in Eq. (6.67) in the evaluation of the integral

$$I = \int_0^{\pi/2} \cos x \, dx$$

6.39 Use the trapezoidal rule to evaluate the double integral

$$\int_{-2}^{2} \int_0^4 (x^2 - xy + y^2) \, dx \, dy.$$

6.40 Use Simpson's $\frac{1}{3}$-rule to evaluate the double integral in Problem 6.39 and compare the results with the exact value.

Answers to Exercises

6.1 –0.0505

6.2 3.975

6.3 2.6444

6.4 0.55

6.5 (a) 0.3931 (b) 0.3341 (c) 0.2706 (d) –0.0256.

6.6 3.73 rad/s; 4.48 rad/s^2

6.7 –5.5; 5.0.

6.8 0

6.9 17; 6

6.10 (a) –0.2831 (b) –0.1735 (c) –0.0639.

6.11 29

6.12 $s_1(x) = 2(x - 1)^3 + 3(2 - x) + 9(x - 1),\ 1 \le x \le 2.$

6.13 1.0

6.15 (a) π, (b) –1.0794

6.16 66.5

6.17 0.475

6.18 (a) 1.1000, (b) 1.0987

6.19 1.1873

6.20 177.4816

6.21 0.2947

6.22 0.7855

6.23 $I_T = 0.9871,\quad I_s = 1.0002$

6.24 1.570858

6.25 0.69319; 0.69317

6.27 0.9513

6.28 0.815

6.29 –4.21667

6.31 1.5

6.32 0.693122

6.33 $k_0 = -\dfrac{2}{15},\ k_1 = \dfrac{8}{5},\ \ k_2 = \dfrac{6}{5}.$

6.36 0.269583

6.37 See Example 6.16.

6.38 Error = 0.00014

6.39 112

6.40 106.6667

Chapter 7

Numerical Linear Algebra

7.1 INTRODUCTION

Most problems arising from engineering and applied sciences require the solution of systems of linear algebraic equations and computation of eigenvalues and eigenvectors of a matrix. We assume that the readers are familiar with the theory of determinants and elements of matrix algebra since these provide a convenient way to represent linear algebraic equations. For example, the system of equations.

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3$$

may be represented as the matrix equation, where

$$AX = b$$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

In Section 7.5, we discuss the Gauss-elimination method and also the method of LU decomposition which is particularly useful in the cases where a system has to be evaluated for several righthand side vectors. Iterative methods of Jacobi and Gauss–Seidel are discussed in Section 7.6.

The eigenvalues of a matrix are of great importance in many engineering problems. For example, problems concerning the stability of an aircraft and those on vibrations of a beam require the computation of eigenvalues of a matrix. The matrix eigenvalue problem is discussed in Section 7.7 and finally Section 7.8 is devoted to a brief discussion of singular value decomposition of a matrix.

7.2 TRIANGULAR MATRICES

A square matrix is said to be *triangular* if the elements above (or below) of the main diagonal are zero. For example, the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \text{ and } B = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

are triangular matrices where A is called an upper triangular matrix and B is a lower triangular matrix.

It is clear that in A, $a_{ij} = 0$ for $i > j$, and $b_{ij} = 0$ for $j > i$ in B. It is also easily seen that a triangular matrix is nonsingular only when all its diagonal elements are nonzero. The following properties hold for triangular matrices:

(i) If A_1 and A_2 are two upper triangular matrices of the same order, then $A_1 + A_2$ and A_1A_2 are also upper triangular matrices of the same order. Similar results hold good for lower triangular matrices also.

(ii) The inverse of a nonsingular lower triangular matrix is also a lower triangular matrix. Similar result holds good for an upper triangular matrix also. This property enables us to invert a triangular matrix easily.

Example 7.1 Find the inverse of the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

Let

$$A^{-1} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

Since $AA^{-1} = I$, we write

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Multiplying the matrices on the left side and equating corresponding elements on both sides, we obtain

$$\begin{aligned} a_{11} &= 1, & a_{22} &= 1, \\ 2a_{11} + a_{12} &= 0, & 2a_{22} + a_{23} &= 0, \\ \Rightarrow a_{12} &= -2, & \Rightarrow a_{23} &= -2, \\ 3a_{11} + 2a_{12} + a_{13} &= 0 & a_{33} &= 1. \\ \Rightarrow a_{13} &= 1. \end{aligned}$$

Hence

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

Since the inverse of a triangular matrix is easily computed, it follows that the inverse of a nonsingular matrix A can be easily obtained if A is expressed as a product of two triangular matrices.

In particular, if $A = LU$, where L and U are lower and upper triangular matrices, then it follows that

$$\begin{aligned} A^{-1} &= (LU)^{-1} \\ &= U^{-1}L^{-1} \end{aligned} \tag{7.1}$$

The next section will be devoted to the LU decomposition of a nonsingular square matrix and this will be used in the solution of a system of linear algebraic equations.

7.3 *LU* DECOMPOSITION OF A MATRIX

Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

be a nonsingular square matrix. Then A can be factorized into the form LU, where

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & \cdots & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix},$$

if

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0, \quad \text{and so on.}$$

It is a standard result of linear algebra that such a factorization, when it exists, is *unique*. Similarly, the factorization LU where

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n_1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & u_{2n} \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

is also a unique factorization. We outline below the procedure for finding L and U with a square matrix of order 3.

Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \tag{7.2}$$

Multiplying the matrices on the right side of Eq. (7.2) and equating the corresponding elements of both sides, we get

$$\left.\begin{array}{lll} u_{11} = a_{11}, & u_{12} = a_{12}, & u_{13} = a_{13}, \\ l_{21}u_{11} = a_{21}, & l_{21}u_{12} + u_{22} = a_{22}, & l_{21}u_{13} + u_{23} = a_{23}, \\ l_{31}u_{11} = a_{31}, & l_{31}u_{12} + l_{32}u_{22} = a_{32}, & l_{31}u_{13} + l_{32}u_{23} + u_{33} = a_{33} \end{array}\right\} \tag{7.3}$$

From the above equations, we obtain

$$\left.\begin{array}{ll} l_{21} = \dfrac{a_{21}}{a_{11}}, & l_{31} = \dfrac{a_{31}}{a_{11}}, \\ u_{22} = a_{22} - \dfrac{a_{21}}{a_{11}} a_{12}, & u_{23} = a_{23} - \dfrac{a_{21}}{a_{11}} a_{13} \\ l_{32} = \dfrac{a_{32} - \dfrac{a_{31}}{a_{11}} a_{12}}{u_{22}}, & \end{array}\right\} \tag{7.4}$$

from which u_{33} can be computed.

The given procedure is a systematic one to evaluate the elements of L and U (where L is *unit* lower triangular and U upper triangular). First, we determine the first row of U and the first column of L, then we determine the second row of U and the second column of L, and finally, we compute the third row of U. It is obvious that this procedure can be generalized. When the factorization is complete, the inverse of A can be computed from formula (7.1).

Example 7.2 Factorize the matrix

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$$

into the LU form.

Let

$$\begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

From Eq. (7.4), we obtain

$$u_{11} = 2, \qquad u_{12} = 3, \qquad u_{13} = 1,$$

$$l_{21} = \frac{1}{2}, \qquad l_{31} = \frac{3}{2},$$

$$u_{22} = \frac{1}{2}, \qquad u_{23} = \frac{5}{2},$$

$$l_{32} = -7, \text{ and} \qquad u_{33} = 18.$$

It follows that

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{3}{2} & -7 & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 2 & 3 & 1 \\ 0 & \frac{1}{2} & \frac{5}{2} \\ 0 & 0 & 18 \end{bmatrix}$$

7.4 VECTOR AND MATRIX NORMS

The distance between a vector and the null vector is a measure of the *size* or *length* of the vector. This is called a *norm* of the vector. The norm of the vector x, written as $\| x \|$, is a real number which satisfies the following conditions or axioms:

$$\| x \| \geq 0 \quad \text{and} \quad \| x \| = 0 \quad \text{if and only if } x = 0 \tag{7.5}$$

$$\| \alpha x \| = | \alpha | \| x \| \quad \text{for any real } \alpha \tag{7.6}$$

$$\| x + y \| \leq \| x \| + \| y \| \quad \text{(triangle inequality).} \tag{7.7}$$

For the vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{7.8}$$

some useful norms are

$$\| x \|_1 = |x_1| + |x_2| + \cdots + |x_n| = \sum_{i=1}^{m} |x_i| \tag{7.9}$$

$$\| x \|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2} = \left[\sum_{i=1}^{n} |x_i|^2 \right]^{1/2} = \| x \|_e \tag{7.10}$$

$$\| x \|_\infty = \max_i |x_i| . \tag{7.11}$$

The norm $\|\bullet\|_2$ is called the *Euclidean* norm since it is just the formula for distance in the three-dimensional Euclidean space. The norm $\|\bullet\|_\infty$ is called the *maximum* norm or the *uniform* norm.

It is easy to show that the three norms $\| x \|_1, \| x \|_2$ and $\| x \|_\infty$ satisfy the conditions (7.5) to (7.7), given above. Conditions (7.5) and (7.6) are trivially satisfied. Only condition (7.7), the triangle inequality, needs to be shown to be *true*. For the norm $\| x \|_1$ we observe that

$$\begin{aligned} \| x + y \| &= \sum_{i=1}^{n} |x_i + y_i| \\ &\leq \sum_{i=1}^{n} (|x_i| + |y_i|) \\ &= \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| \\ &= \| x \|_1 + \| y \|_1 \end{aligned} \tag{7.12}$$

Similarly, for $\| x \|_\infty$, we have

$$\begin{aligned} \| x + y \|_\infty &= \max_i \, | x_i + y_i | \\ &\le \max_i \, (| x_i | + | y_i |) \\ &= \| x \|_\infty + \| y \|_\infty . \end{aligned} \tag{7.13}$$

The proof for the Euclidean norm is left as an exercise to the reader.

To define matrix norms, we consider two matrices A and B for which the operations $A + B$ and AB are defined. Then,

$$| A + B | \le | A | + | B | \tag{7.14}$$

$$| AB | \le | A | | B | \tag{7.15}$$

$$| \alpha A | = | \alpha | | A | \quad (\alpha \text{ a scalar}). \tag{7.16}$$

From Eq. (7.15) it follows that

$$| A^p | \le | A |^p, \tag{7.17}$$

where p is a natural number. In the above equations, $|A|$ denotes the matrix A with absolute values of the elements.

By the norm of a matrix $A = | a_{ij} |$, we mean a nonnegative number, denoted by $\| A \|$, which satisfies the following conditions

$$\| A \| \ge 0 \quad \text{and} \quad \| A \| = 0 \quad \text{if and only if} \quad A = 0 \tag{7.18}$$

$$\| \alpha A \| = | \alpha | \| A \| \quad (\alpha \text{ a scalar}) \tag{7.19}$$

$$\| A + B \| \le \| A \| + \| B \| \tag{7.20}$$

$$\| AB \| \le \| A \| \| B \| . \tag{7.21}$$

From Eq. (7.21), it easily follows that

$$\| A^p \| \le \| A \|^p, \tag{7.22}$$

where p is a natural number.

Corresponding to the vector norms given in Eqs. (7.9)–(7.11), we have the three matrix norms

$$\| A \|_1 = \max_j \sum_i | a_{ij} | \quad \text{(the column norm)} \tag{7.23}$$

$$\| A \|_e = \left[\sum_{i, j} | a_{ij} |^2 \right]^{1/2} \quad \text{(the Euclidean norm)} \tag{7.24}$$

$$\| A \|_\infty = \max_i \sum_j | a_{ij} | \qquad \text{(the row norm).} \tag{7.25}$$

In addition to the above, we have $\| A \|_2$ defined by

$$\| A \|_2 = (\text{Maximum eigenvalue of } A^{\mathrm{T}}A)^{1/2}. \tag{7.26}$$

The eigenvalues of a matrix will be discussed in Section 7.7.

The choice of a particular norm is dependent mostly on practical considerations. The row-norm is, however, most widely used because it is easy to compute and, at the same time, provides a fairly adequate measure of the size of the matrix.

The following example demonstrates the computation of some of these norms.

Example 7.3 Given the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

find $\| A \|_1, \| A \|_e$ and $\| A \|_\infty$.

We have

$$\| A \|_1 = \max\,[1+4+7,\ 2+5+8,\ 3+6+9] = \max\,[12, 15, 18] = 18$$

$$\begin{aligned} \| A \|_e &= (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2)^{1/2} \\ &= (1+4+9+16+25+36+49+64+81)^{1/2} \\ &= (285)^{1/2} \\ &= 16.88. \end{aligned}$$

$$\begin{aligned} \| A \|_\infty &= \max\,[1+2+3,\ 4+5+6,\ 7+8+9] \\ &= \max\,[6,\ 15,\ 24] \\ &= 24. \end{aligned}$$

The concept of the norm of a matrix will be useful in the study of the convergence of iterative methods of solving linear systems. It is also used in defining the 'stability' of a system of equations.

7.5 SOLUTION OF LINEAR SYSTEMS—DIRECT METHODS

The solution of a linear system of equations can be accomplished by a numerical method which falls in one of two categories: *direct* or *iterative*

methods. Amongst the direct methods, we will describe the elimination method by Gauss as also its modification and the *LU* decomposition method. About the iterative types, we will describe only the Jacobi and Gauss–Seidel methods.

7.5.1 Gauss Elimination

This is the elementary elimination method and it reduces the system of equations to an equivalent upper-triangular system, which can be solved by *back substitution*.

Let the system of n linear equations in n unknowns be given by

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n. \end{aligned}\right\} \tag{7.27}$$

There are two steps in the solution of the system given in Eq. (7.27), viz., the elimination of unknowns and back substitution.

Step 1: The unknowns are eliminated to obtain an upper-triangular system.

To eliminate x_1 from the second equation, we multiply the first equation by $(-a_{21}/a_{11})$ and obtain

$$-a_{21}x_1 - a_{12}\frac{a_{21}}{a_{11}}x_2 - a_{13}\frac{a_{21}}{a_{11}}x_3 - \cdots - a_{1n}\frac{a_{21}}{a_{11}}x_n = -b_1\frac{a_{21}}{a_{11}}.$$

Adding the above equation to the second equation of Eq. (7.27), we obtain

$$\left(a_{22} - a_{12}\frac{a_{21}}{a_{11}}\right)x_2 + \left(a_{23} - a_{13}\frac{a_{21}}{a_{11}}\right)x_3 + \cdots + \left(a_{2n} - a_{1n}\frac{a_{21}}{a_{11}}\right)x_n = b_2 - b_1\frac{a_{21}}{a_{11}}, \tag{7.28}$$

which can be written as

$$a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n = b'_2,$$

where $a'_{22} = a_{22} - a_{12}(a_{21}/a_{11})$, etc. Thus the primes indicate that the original element has changed its value. Similarly, we can multiply the first equation by $-a_{31}/a_{11}$ and add it to the third equation of the system (7.27). This eliminates the unknown x_1 from the third equation of Eq. (7.27) and we obtain

$$a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n = b'_3. \tag{7.29}$$

In a similar fashion, we can eliminate x_1 from the remaining equations and after eliminating x_1 from the last equation of Eq. (7.27), we obtain the system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n &= b'_3 \\ &\vdots \\ a'_{n2}x_2 + a'_{n3}x_3 + \cdots + a'_{nn}x_n &= b'_n. \end{aligned}\right\} \tag{7.30}$$

We next eliminate x_2 from the last $(n-2)$ equations of Eq. (7.30). Before this, it is important to notice that in the process of obtaining the above system, we have multiplied the first row by $(-a_{21}/a_{11})$, i.e. we have divided it by a_{11} which is therefore assumed to be nonzero. For this reason, the first equation in the system (7.30) is called the *pivot equation*, and a_{11} is called the *pivot* or *pivotal element*. The method obviously fails if $a_{11} = 0$. We shall discuss this important point after completing the description of the elimination method. Now, to eliminate x_2 from the third equation of Eq. (7.30), we multiply the second equation by $(-a'_{32}/a'_{22})$ and add it to the third equation. Repeating this process with the remaining equations, we obtain the system

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a''_{33}x_3 + \cdots + a''_{3n}x_n &= b''_3 \\ &\vdots \\ a''_{n3}x_3 + \cdots + a''_{nn}x_n &= b''_n. \end{aligned}\right\} \tag{7.31}$$

In Eq. (7.31), the 'double primes' indicate that the *elements have changed twice*. It is easily seen that this procedure can be continued to eliminate x_3 from the fourth equation onwards, x_4 from the fifth equation onwards, etc., till we finally obtain the upper-triangular form:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\ a''_{33}x_3 + \cdots + a''_{3n}x_n &= b''_3 \\ &\vdots \\ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)}, \end{aligned}\right\} \tag{7.32}$$

where $a_{nn}^{(n-1)}$ indicates that the element a_{nn} has changed $(n-1)$ times. We thus have completed the first step of elimination of unknowns and reduction to the upper-triangular form.

Step 2: We now have to obtain the required solution from the system (7.32). From the last equation of this system, we obtain

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}. \tag{7.33}$$

This is then substituted in the $(n-1)$th equation to obtain x_{n-1} and the process is repeated to compute the other unknowns. We have therefore first computed x_n, then $x_{n-1}, x_{n-2}, \ldots, x_2, x_1$, in that order. Due to this reason, the process is called *back substitution*.

7.5.2 Necessity for Pivoting

We now come to the important case of the pivot being zero or very close to zero. If the *pivot is zero*, the *entire process fails* and if it is close to zero, round-off errors may occur. These problems can be avoided by adopting a procedure called *pivoting*. If a_{11} is either zero or very small compared to the other coefficients of the equation, then we find the largest available coefficient in the columns below the pivot equation and then *interchange* the two rows. In this way, we obtain a new pivot equation with a nonzero pivot. Such a process is called *partial pivoting*, since in this case we search only the columns below for the largest element. If, on the other hand, we search both columns and rows for the largest element, the procedure is called *complete pivoting*. It is obvious that complete pivoting involves more complexity in computations since interchange of columns means change of 'order' of unknowns which invariably requires more programming effort. In comparison, partial pivoting, i.e. row interchanges, is easily adopted in programming. Due to this reason, complete pivoting is rarely used.

Example 7.4 Use Gauss elimination to solve the system

$$2x + y + z = 10$$

$$3x + 2y + 3z = 18$$

$$x + 4y + 9z = 16.$$

We first eliminate x from the second and third equations. For this we multiply the first equation by (–3/2) and add to the second to get

$$y + 3z = 6. \tag{i}$$

Similarly, we multiply the first equation by (–1/2) and add it to the third to get

$$7y + 17z = 22. \tag{ii}$$

We thus have eliminated x from the second and third equations. Next, we have to eliminate y from (i) and (ii). For this we multiply (i) by –7 and add to (ii). This gives

$$-4z = -20 \quad \text{or} \quad z = 5.$$

The upper-triangular form is therefore given by

$$2x + y + z = 10$$
$$y + 3z = 6$$
$$z = 5.$$

It follows that the required solution is $x = 7$, $y = -9$ and $z = 5$.

The next example demonstrates the necessity of pivoting in the elimination method.

Example 7.5 Solve the system

$$0.0003120x_1 + 0.006032x_2 = 0.003328$$
$$0.5000x_1 + 0.8942x_2 = 0.9471$$

The exact solution is $x_1 = 1$ and $x_2 = 0.5$.

We first solve the system with pivoting. We write the given system as

$$0.5000x_1 + 0.8942x_2 = 0.9471$$
$$0.000312x_1 + 0.006032x_2 = 0.003328$$

using Gaussian elimination, the above system reduces to

$$0.5000x_1 + 0.8942x_2 = 0.9471$$
$$0.005474x_2 = 0.002737$$

Back substitution gives: $x_2 = 0.5$ and $x_1 = 1.0$.

Without pivoting, Gaussion elimination gives the system

$$0.000312x_1 + 0.006032x_2 = 0.003328$$
$$-8.7725x_2 = -5.3300$$

The back substitution process gives

$$x_2 = 0.6076 \text{ and } x_1 = -1.0803$$

The effect of pivoting is clearly seen.

7.5.3 Gauss–Jordan Method

This is a modification of the Gauss elimination method, the essential difference being that when an unknown is eliminated, it is eliminated from all equations. The method does not require back substitution to obtain the solution and is best illustrated by the following example.

Example 7.6 Solve the system (*see* Example 7.4)

$$2x + y + z = 10$$
$$3x + 2y + 3z = 18$$
$$x + 4y + 9z = 16.$$

by the Gauss–Jordan method.

Elimination of x from the second and third equations is done as in 'Gauss elimination' and we obtain the system

$$2x + y + z = 10$$

$$(1/2)y + (3/2)z = 3$$

$$(7/2)y + (17/2)z = 11.$$

Next, the unknown y is eliminated from *both* the first and third equations. This gives us

$$x - z = 2 \quad \text{and} \quad z = 5.$$

Hence the system becomes:

$$x - z = 2$$

$$y + 3z = 6$$

$$z = 5.$$

Evaluation of y and z is trivial and the result is the same as before.

7.5.4 Modification of the Gauss Method to Compute the Inverse

We know that X will be the inverse of A if

$$AX = I, \tag{7.34}$$

where I is the unit matrix of the same order as A. It is required to determine the elements of X such that Eq. (7.34) is satisfied. For example, for third-order matrices, Eq. (7.34) may be written as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The reader can easily see that this equation is equivalent to the three equations

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

We can therefore apply the Gaussian elimination method to each of these systems and the result in each case will be the *corresponding* column of A^{-1}. Since the matrix of coefficients is the same in each case, we can solve all the three systems simultaneously. Starting with the 'augmented system'

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \vdots & 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & \vdots & 0 & 1 & 0 \\ a_{31} & a_{32} & a_{33} & \vdots & 0 & 0 & 1 \end{bmatrix}$$

we obtain at the end of the first and second stage, respectively

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \vdots & 1 & 0 & 0 \\ 0 & a'_{22} & a'_{23} & \vdots & -a_{21}/a_{11} & 1 & 0 \\ 0 & a'_{32} & a'_{33} & \vdots & -a_{31}/a_{11} & 0 & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \vdots & 1 & 0 & 0 \\ 0 & a'_{22} & a'_{23} & \vdots & \alpha_{21} & 1 & 0 \\ 0 & 0 & a''_{33} & \vdots & \alpha_{31} & \alpha_{32} & 1 \end{bmatrix},$$

where

$$\alpha_{21} = -\frac{a_{21}}{a_{11}}, \quad \alpha_{31} = \frac{a_{21}}{a_{11}}\frac{a'_{32}}{a'_{22}} - \frac{a_{31}}{a_{11}}, \quad \alpha_{32} = -\frac{a'_{32}}{a'_{22}}.$$

The inverse can now be obtained easily, since the back-substitution process with each column of the matrix I will yield the corresponding column of A^{-1}, where I is given by

$$I = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_{21} & 1 & 0 \\ \alpha_{31} & \alpha_{32} & 1 \end{bmatrix}.$$

Example 7.7 We shall consider again the system given in Example 7.6. We have here

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 3 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix}.$$

The augmented system is

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 1 & 0 & 0 \\ 3 & 2 & 3 & \vdots & 0 & 1 & 0 \\ 1 & 4 & 9 & \vdots & 0 & 0 & 1 \end{bmatrix}.$$

After the first stage, this becomes

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 1 & 0 & 0 \\ 0 & 1/2 & 3/2 & \vdots & -3/2 & 1 & 0 \\ 0 & 7/2 & 17/2 & \vdots & -1/2 & 0 & 1 \end{bmatrix}.$$

Finally, at the end of the second stage, the system becomes:

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 1 & 0 & 0 \\ 0 & 1/2 & 3/2 & \vdots & -3/2 & 1 & 0 \\ 0 & 0 & -2 & \vdots & 10 & -7 & 1 \end{bmatrix}.$$

This is equivalent to the three systems:

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 1 \\ 0 & 1/2 & 3/2 & \vdots & -3/2 \\ 0 & 0 & -2 & \vdots & 10 \end{bmatrix},$$

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 0 \\ 0 & 1/2 & 3/2 & \vdots & 1 \\ 0 & 0 & -2 & \vdots & -7 \end{bmatrix}$$

and

$$\begin{bmatrix} 2 & 1 & 1 & \vdots & 0 \\ 0 & 1/2 & 3/2 & \vdots & 0 \\ 0 & 0 & -2 & \vdots & 1 \end{bmatrix},$$

whose solution by back substitution yields the three columns of the matrix:

$$\begin{bmatrix} -3 & 5/2 & -1/2 \\ 12 & -17/2 & 3/2 \\ -5 & 7/2 & -1/2 \end{bmatrix},$$

which is the required inverse A^{-1}.

We can also find

$$|A| = 2\left(\frac{1}{2}\right)(-2) = -2$$

by looking at the triangulated coefficient matrix. If this value is zero, then we cannot back substitute and the matrix has no inverse.

7.5.5 Number of Arithmetic Operations

Since the total execution time depends on the number of multiplications and divisions in Gaussian elimination, we give below a count of the total number of floating-point multiplications or divisions in this method.

For eliminating x_1, i.e. in Eq. (7.28), the factor a_{21}/a_{11} is computed once. There are $(n - 1)$ multiplications in the $(n - 1)$ terms on the left side and 1 multiplication on the right side. Hence the number of 'floating-point' multiplications/divisions required for eliminating x_1 is $1 + n - 1 + 1 = n + 1$. But x_1 is eliminated from $(n - 1)$ equations. Therefore, the total number of multiplications/divisions required to eliminate x_1 from $(n - 1)$ equations is

$$(n-1)\,(n+1) = (n-1)\,(n+2-1).$$

Similarly, the total number of multiplications/divisions required to eliminate x_2 from $(n - 2)$ equations is

$$(n-2)\,n = (n-2)\,(n+2-2).$$

The total number of multiplications/divisions required to eliminate x_3 from $(n - 3)$ equations is

$$(n-3)\,(n-1) = (n-3)\,(n+2-3).$$

Similarly, the total number of multiplications/divisions required to eliminate x_p from $(n - p)$ equations is

$$(n-p)\,(n+2-p),$$

and finally, x_{n-1} is eliminated in

$$[n-(n-1)]\,[n+2-(n-1)] = 1.3.$$

Summing up all the above, we can write the total number of arithmetic operations (i.e. multiplications/divisions) as

$$\begin{aligned}
\sum_{p=1}^{n-1} (n-p)\,(n+2-p) &= \sum [(n-p)^2 + 2(n-p)] \\
&= \sum_{p=1}^{n-1} (n^2 + p^2 - 2np + 2n - 2p) \\
&= n^2(n-1) + \frac{(n-1)\,(n)\,(2n-2+1)}{6} - 2n\,\frac{(n-1)\,n}{2} \\
&\quad + 2n(n-1) - 2\,\frac{(n-1)\,n}{2} \\
&\approx \frac{n^3}{3},
\end{aligned}$$

where we have used the formulae:

$$1+2+3+\cdots+n = \frac{n(n+1)}{2} \quad \text{and} \quad 1^2+2^2+3^2+\cdots+n^2 = \frac{n(n+1)\,(2n+1)}{6}.$$

It follows that the total number of 'floating-point' multiplications or divisions in Gaussian elimination is $n^3/3$. In a similar way, it can be shown that the Gauss–Jordan method requires $n^3/2$ arithmetic operations. Hence, Gauss elimination is preferred to Gauss–Jordan method while solving large systems of equations.

7.5.6 *LU* Decomposition Method

In Section 7.3, we described a scheme for computing the matrices L and U such that

$$A = LU \tag{7.35}$$

where L is unit lower triangular and U an upper triangular matrix, and these are given in Eq. (7.2). Let the system of equations be given by

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}\right\} \tag{7.36}$$

which can be written in the matrix form

$$AX = B \tag{7.37}$$

or

$$LUX = B \tag{7.38}$$

If we set

$$UX = Y, \tag{7.39}$$

then Eq. (7.38) becomes

$$LY = B \tag{7.40}$$

which is equivalent to the system

$$\left.\begin{aligned} y_1 &= b_1 \\ l_{21}y_1 + y_2 &= b_2 \\ l_{31}y_1 + l_{32}y_2 + y_3 &= b_3 \end{aligned}\right\} \tag{7.41}$$

where

$$(y_1, y_2, y_3)^T = Y \text{ and } (b_1, b_2, b_3)^T = B.$$

The system (7.41) can be solved for y_1, y_2 and y_3 by forward substitution. When Y is known, the system (7.39) becomes

$$\left.\begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 &= y_1 \\ u_{22}x_2 + u_{23}x_3 &= y_2 \\ u_{33}x_3 &= y_3 \end{aligned}\right\} \tag{7.42}$$

which can be solved for x_1, x_2, x_3 by backsubstitution process. As noted earlier, this method has the advantage of being applicable to solve systems with the same coefficient matrix but different right-hand side vectors.

7.5.7 Computational Procedure for *LU* Decomposition Method

Given any nonsingular square matrix A, the LU decomposition, where L is unit lower triangular and U an upper triangular matrix, can be achieved by the following computational steps:

```
Do i = 1(1)N - 1
Do j = i + 1(1)N
A(j, i) = A(j, i)/A(i, i)
Do M = i + 1(1) N
A(j, M) = A(j, M) - A(i, M) * A (j, i)
Next M
Next j
Next i
End
```

To save storage space, the elements of L and U are stored in the space occupied by A, the elements l_{ii} being omitted. Thus, after the factorization is effected, the layout of the store for a (4×4) matrix is given by

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ l_{21} & u_{22} & u_{23} & u_{24} \\ l_{31} & l_{32} & u_{33} & u_{34} \\ l_{41} & l_{42} & l_{43} & u_{44} \end{bmatrix}$$

If the right-hand side vector is $B = (b_1, b_2, b_3)$, then the forward substitution can be accomplished by the statements:

```
Do j = 1(1)N - 1
Do i = j + 1(1) N
b(i) = b(i) - l(i, j) * b(j)
Next i
Next j
End.
```

Similarly, the backsubstitution can be effected by the steps:

```
Do j = N(-1)1
b(j) = b(j)/u(j, j)
Do i =1(1) j - 1
b(i) =b(i) - u(i, j) * b(j)
Next i
Next j
End
```

Example 7.8 Solve the equations

$$2x + 3y + z = 9$$
$$x + 2y + 3z = 6$$
$$3x + y + 2z = 8$$

by the method of LU decomposition.
We have

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} \text{ and } B = \begin{bmatrix} 9 \\ 6 \\ 8 \end{bmatrix}$$

In Example 7.2, we obtained the LU decomposition of A. This is

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 1.5 & -7 & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 0.5 & 2.5 \\ 0 & 0 & 18 \end{bmatrix}$$

If $Y = [y_1, y_2, y_3]^T$, then the equation $LY = B$ gives the solution:

$$y_1 = 9, \quad y_2 = \frac{3}{2} \quad \text{and} \quad y_3 = 5.$$

Finally, the matrix equation

$$UX = Y \quad \text{where } X = [x, y, z]^T,$$

gives the required solution

$$x = \frac{35}{18}, \; y = \frac{29}{18} \text{ and } z = \frac{5}{18}.$$

7.5.8 *LU* Decomposition from Gauss Elimination

We have seen that Gaussian elimination consists in reducing the coefficient matrix to an upper-triangular form. We show that the LU decomposition of the coefficient matrix can also be obtained from Gauss elimination. The upper-triangular form to which the coefficient matrix is reduced is actually the upper-triangular matrix U of the decomposition LU. Then, what is the lower-triangular matrix L? For this, we consider the system defined by

$$AX = b, \tag{7.43}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

To eliminate x_1 from the second equation, we multiply the first equation by a_{21}/a_{11} and subtract it from the second equation. We then obtain

$$\left(a_{22} - a_{12}\frac{a_{21}}{a_{11}}\right)x_2 + \left(a_{23} - a_{13}\frac{a_{21}}{a_{11}}\right)x_3 = \left(b_2 - b_1\frac{a_{21}}{a_{11}}\right)$$

or

$$a'_{22}x_2 + a'_{23}x_3 = b'_2. \tag{7.44}$$

The factor $l_{21} = a_{21}/a_{11}$ is called the *multiplier* for eliminating x_1 from the second equation. Similarly, the multiplier for eliminating x_1 from the third equation is given by $l_{31} = a_{31}/a_{11}$. After this elimination, the system is of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 &= b'_2 \\ a'_{32}x_2 + a'_{33}x_3 &= b'_3. \end{aligned}\right\} \tag{7.45}$$

In the final step, we have to eliminate x_2 from the third equation. For this we multiply the second equation by a'_{32}/a'_{22} and subtract it from the third equation. We then obtain

$$a''_{33}x_3 = b''_3, \tag{7.46}$$

where the double primes indicate that the concerned elements have changed their values twice. In this step, the multiplier is given by $l_{32} = a'_{32}/a'_{22}$. The final form of the matrix of coefficients is the upper-triangular matrix given by

$$U = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & 0 & a''_{33} \end{bmatrix}. \tag{7.47}$$

Equation (7.47) suggests that in any computer program, the places occupied by the zero elements may be used to store the values of the multipliers. Thus, after elimination, the matrix A may be written as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ l_{21} & a'_{22} & a'_{23} \\ l_{31} & l_{32} & a''_{33} \end{bmatrix}, \tag{7.48}$$

which represents the storage of the LU decomposition of A with

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & 0 & a''_{33} \end{bmatrix}.$$

It is easily verified that $A = LU$.

7.5.9 Solution of Tridiagonal Systems

Consider the system of equations defined by

$$\left.\begin{aligned} b_1u_1 + c_1u_2 &= d_1 \\ a_2u_1 + b_2u_2 + c_2u_3 &= d_2 \\ a_3u_2 + b_3u_3 + c_3u_4 &= d_3 \\ &\vdots \\ a_nu_{n-1} + b_nu_n &= d_n. \end{aligned}\right\} \tag{7.49}$$

The matrix of coefficients is

$$A = \begin{bmatrix} b_1 & c_1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & a_3 & b_3 & c_3 & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & a_n & b_n \end{bmatrix} \tag{7.50}$$

Matrices of the type, given in Eq. (7.50), called the *tridiagonal matrices,* occur frequently in the solution of ordinary and partial differential equations by finite difference methods. The method of factorization described earlier can be conveniently applied to solve the system (7.49). For example, for a (3×3) matrix we have

$$\begin{bmatrix} b_1 & c_1 & 0 \\ a_2 & b_2 & c_2 \\ 0 & a_3 & b_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ 0 & l_{32} & 1 \end{bmatrix} \begin{bmatrix} b_1 & c_1 & 0 \\ 0 & u_{22} & c_2 \\ 0 & 0 & u_{33} \end{bmatrix}$$

This matrix equation gives

$$\left.\begin{aligned} l_{21}b_1 &= a_2, \quad l_{21}c_1 + u_{22} = b_2, \\ l_{32}u_{22} &= a_3, \quad l_{32}c_2 + u_{33} = b_3 \end{aligned}\right\} \tag{7.51}$$

From these four equations, we can compute l_{21}, u_{22}, l_{32} and u_{33} and these values are stored in the locations occupied by a_2, b_2, a_3 and b_3, respectively. These computations can be achieved by the following statements:

```
Do i = 2(1)N
a(i) = a(i)/b(i – 1)
b(i) = b(i) – a(i) c(i –1)
Next i
```

When the decomposition is complete, forward and back substitutions give the required solution. This algorithm is due to Thomas and possesses all the advantages of the LU decomposition.

7.5.10 Ill-conditioned Linear Systems

In practical applications, one usually encounters systems of equations in which small changes in the coefficients of the system produce large changes in the solution. Such systems are said to be *ill-conditioned.* On the other hand, if the corresponding changes in the solution are also small, then the system is *well-conditioned.*

Ill-conditioning can usually be expected when $|A|$, in the system $AX = b$, is small. The quantity $c(A)$ defined by

$$c(A) = \|A\| \; \|A^{-1}\|, \tag{7.52}$$

where $\|A\|$ is any matrix norm, gives a *measure of the condition of the matrix.* It is, therefore, called the *condition number* of the matrix-Large condition numbers indicate that the matrix is ill-conditioned. Again, let $A = [a_{ij}]$ and

$$s_i = \left[a_{i1}^2 + a_{i2}^2 + \cdots + a_{in}^2\right]^{1/2} \tag{7.53}$$

If we define

$$k = \frac{|A|}{s_1 s_2 \cdots s_n}, \tag{7.54}$$

then the system is ill-conditioned if k is very small compared to unity. Otherwise, it is well-conditioned.

Example 7.9 The system

$$\left.\begin{aligned} 2x + y &= 2 \\ 2x + 1.01y &= 2.01 \end{aligned}\right\} \tag{i}$$

has the solution

$$x = 0.5 \quad \text{and} \quad y = 1.$$

But the system

$$\left.\begin{aligned} 2x + y &= 2 \\ 2.01x + y &= 2.05 \end{aligned}\right\} \tag{ii}$$

has the solution $x = 5$ and $y = -8$.

Also,

$$\|A\|_e = 3.165 \quad \text{and} \quad \|A^{-1}\|_e = 158.273$$

Therefore, condition number $c(A) = \|A\| \, \|A^{-1}\| = 500.974$.
Hence the system is ill-conditioned.
Also

$$|A| = 0.02$$

$$s_1 = \sqrt{5} \quad \text{and} \quad s_2 = 2.24$$

So,

$$k = 4.468 \times 10^{-3}$$

Hence the system is ill-conditioned.

Example 7.10 Let

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{8} & \frac{1}{9} & \frac{1}{10} \end{bmatrix}$$

which is called *Hilbert's matrix.*

Now,

$$|A| = 0.0000297, \text{ which is small compared to } 1.$$

Hence A is ill-conditioned.

Example 7.11 Let

$$A = \begin{bmatrix} 25 & 24 & 10 \\ 66 & 78 & 37 \\ 92 & -73 & -80 \end{bmatrix}$$

Now,

$$|A| = 1.0.$$

Also,

$$s_1 = 36.0694, \quad s_2 = 108.6692 \quad \text{and} \quad s_3 = 142.1021.$$

Therefore,

$$k = 1.7954 \times 10^{-6}.$$

which shows that A is ill-conditioned.

7.5.11 Method for Ill-conditioned Systems

In general, the accuracy of an approximate solution can be improved upon by an iterative procedure. This is described below. Let the system be

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}\right\} \tag{7.55}$$

Let $x_1^{(1)}$, $x_2^{(1)}$ and $x_3^{(1)}$ be an approximate solution. Substituting these values in the left side of Eq. (7.55), we get new values of b_1, b_2 and b_3. Let these new values be $b_1^{(1)}$, $b_2^{(1)}$ and $b_3^{(1)}$. The new system of equations is given by

$$\left.\begin{aligned} a_{11}x_1^{(1)} + a_{12}x_2^{(1)} + a_{13}x_3^{(1)} &= b_1^{(1)} \\ a_{21}x_1^{(1)} + a_{22}x_2^{(1)} + a_{23}x_3^{(1)} &= b_2^{(1)} \\ a_{31}x_1^{(1)} + a_{32}x_2^{(1)} + a_{33}x_3^{(1)} &= b_3^{(1)} \end{aligned}\right\} \tag{7.56}$$

Subtracting each equation given in Eq. (7.56) from the corresponding equation given in Eq. (7.55), we obtain

$$\left.\begin{aligned} a_{11}e_1 + a_{12}e_2 + a_{13}e_3 &= d_1 \\ a_{21}e_1 + a_{22}e_2 + a_{23}e_3 &= d_2 \\ a_{31}e_1 + a_{32}e_2 + a_{33}e_3 &= d_3 \end{aligned}\right\} \tag{7.57}$$

where $e_i = x_i - x_i^{(1)}$ and $d_i = b_i - b_i^{(1)}$. We now solve the system (7.57) for e_1, e_2 and e_3. Since $e_i = x_i - x_i^{(1)}$, we obtain

$$x_i = x_i^{(1)} + e_i, \tag{7.58}$$

which is a letter approximation for x_i. The procedure can be repeated to improve upon the accuracy.

Example 7.12 Solve the system

$$2x + y = 2$$
$$2x + 1.01y = 2.01$$

Let an approximate solution of the given system be given by

$$x^{(1)} = 1 \quad \text{and} \quad y^{(1)} = 1.$$

Substituting these values in the given system, we obtain

$$\left.\begin{aligned} 2x^{(1)} + y^{(1)} &= 3 \\ \text{and}\ \ 2x^{(1)} + 1.01y^{(1)} &= 3.01 \end{aligned}\right\} \tag{i}$$

Subtracting each equation of (i) from the corresponding equation of the given system, we get

$$2(x - x^{(1)}) + (y - y^{(1)}) = -1$$

and

$$2(x - x^{(1)}) + 1.01(y - y^{(1)}) = -1.$$

Solving the above system of equations, we obtain

$$x - x^{(1)} = -\frac{1}{2} \quad \text{and} \quad y - y^{(1)} = 0.$$

Hence

$$x = \frac{1}{2} \quad \text{and} \quad y = 1,$$

which is the exact solution of the given system.

7.6 SOLUTION OF LINEAR SYSTEMS—ITERATIVE METHODS

We have so far discussed some direct methods for the solution of simultaneous linear equations and we have seen that these methods yield the solution after an amount of computation that is known in advance. We shall now describe the *iterative* or *indirect* methods, which start from an *approximation* to the true solution and, if convergent, derive a sequence of closer approximations—*the cycle of computations being repeated till the required accuracy is obtained*. This means that in a direct method the amount of computation is fixed, while in an iterative method the amount of computation depends on the accuracy required.

In general, one should prefer a direct method for the solution of a linear system, but in the case of matrices with a large number of zero elements, it will be advantageous to use iterative methods which preserve these elements.

Let the system be given by

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n &= b_3 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{7.59}$$

in which the diagonal elements a_{ii} do not vanish. If this is not the case, then the equations should be rearranged so that this condition is satisfied. Now, we rewrite the system (7.59) as

$$\left.\begin{aligned} x_1 &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \cdots - \frac{a_{1n}}{a_{11}}x_n \\ x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_2 - \cdots - \frac{a_{2n}}{a_{22}}x_n \\ x_3 &= \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}}x_1 - \frac{a_{32}}{a_{33}}x_2 - \cdots - \frac{a_{3n}}{a_{33}}x_n \\ &\vdots \\ x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \cdots - \frac{a_{n,n-1}}{a_{nn}}x_{n-1}. \end{aligned}\right\} \tag{7.60}$$

Suppose $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$ are any first approximations to the unknowns $x_1, x_2, \ldots, x_n$. Substituting in the right side of Eq. (7.60), we find a system of second approximations

$$\left.\begin{aligned}
x_1^{(2)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} x_2^{(1)} - \cdots - \frac{a_{1n}}{a_{11}} x_n^{(1)}, \\
x_2^{(2)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_1^{(1)} - \cdots - \frac{a_{2n}}{a_{22}} x_n^{(1)}, \\
x_3^{(2)} &= \frac{b_3}{a_{33}} - \frac{a_{31}}{a_{33}} x_1^{(1)} - \cdots - \frac{a_{3n}}{a_{33}} x_n^{(1)}, \\
&\vdots \\
x_n^{(2)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}} x_1^{(1)} - \cdots - \frac{a_{n,n-1}}{a_{nn}} x_{n-1}^{(1)}.
\end{aligned}\right\} \tag{7.61}$$

Similarly, if $x_1^{(n)}, x_2^{(n)}, \ldots, x_n^{(n)}$ are a system of nth approximations, then the next approximation is given by the formula

$$\left.\begin{aligned}
x_1^{(n+1)} &= \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}} x_2^{(n)} - \cdots - \frac{a_{1n}}{a_{11}} x_n^{(n)}, \\
x_2^{(n+1)} &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_1^{(n)} - \cdots - \frac{a_{2n}}{a_{22}} x_n^{(n)}, \\
&\vdots \\
x_n^{(n+1)} &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}} x_1^{(n)} - \cdots - \frac{a_{n,\,n-1}}{a_{nn}} x_{n-1}^{(n)}.
\end{aligned}\right\} \tag{7.62}$$

If we write Eq. (7.60) in the matrix form

$$X = BX + C \tag{7.63}$$

then the iteration formula (7.62) may be written as

$$X^{(n+1)} = BX^{(n)} + C. \tag{7.64}$$

This method is due to Jacobi and is called the *method of simultaneous displacements*. It can be shown that a sufficient condition for the convergence of this method is that

$$\| B \| < 1. \tag{7.65}$$

A simple modification in this method sometimes yields faster convergence and is described below:

In the first equation of Eq. (7.60), we substitute the first approximation $(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_n^{(1)})$ into the right-hand side and denote the result as $x_1^{(2)}$. In the second equation, we substitute $(x_1^{(2)}, x_2^{(1)}, x_3^{(1)}, \ldots, x_n^{(1)})$ and denote

the result as $x_2^{(2)}$. In the third, we substitute $(x_1^{(2)}, x_2^{(2)}, x_3^{(1)}, \ldots, x_n^{(1)})$ and call the result as $x_3^{(2)}$. In this manner, we complete the first stage of iteration and the entire process is repeated till the values of $x_1, x_2, \ldots, x_n$ are obtained to the accuracy required. It is clear, therefore, that this method uses an improved component as soon as it is available and it is called the *method of successive displacements*, or the *Gauss–Seidel* method.

The Jacobi and Gauss–Seidel methods converge, for any choice of the first approximation $x_j^{(1)}$ $(j = 1, 2, \ldots, n)$, if every equation of the system (7.60) satisfies the condition that the sum of the absolute values of the coefficients a_{ij}/a_{ii} is almost equal to, or in at least one equation less than unity, i.e. provided that

$$\sum_{j=1,\, j \neq i}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \leq 1, \qquad (i = 1, 2, \ldots, n), \tag{7.66}$$

where the '<' sign should be valid in the case of 'at least' one equation. It can be shown that the *Gauss–Seidel method converges twice as fast as the Jacobi method*. The working of the methods is illustrated in the following examples:

Example 7.13 We consider the equations:

$$\begin{aligned} 10x_1 - 2x_2 - x_3 - x_4 &= 3 \\ -2x_1 + 10x_2 - x_3 - x_4 &= 15 \\ -x_1 - x_2 + 10x_3 - 2x_4 &= 27 \\ -x_1 - x_2 - 2x_3 + 10x_4 &= -9. \end{aligned}$$

To solve these equations by the iterative methods, we re-write them as follows:

$$\begin{aligned} x_1 &= 0.3 + 0.2x_2 + 0.1x_3 + 0.1x_4 \\ x_2 &= 1.5 + 0.2x_1 + 0.1x_3 + 0.1x_4 \\ x_3 &= 2.7 + 0.1x_1 + 0.1x_2 + 0.2x_4 \\ x_4 &= -0.9 + 0.1x_1 + 0.1x_2 + 0.2x_3. \end{aligned}$$

It can be verified that these equations satisfy the condition given in Eq. (7.66). The results are given in Tables 7.1 and 7.2:

Table 7.1 Gauss–Seidel Method

n	x_1	x_2	x_3	x_4
1	0.3	1.56	2.886	−0.1368
2	0.8869	1.9523	2.9566	−0.0248
3	0.9836	1.9899	2.9924	−0.0042
4	0.9968	1.9982	2.9987	−0.0008
5	0.9994	1.9997	2.9998	−0.0001
6	0.9999	1.9999	3.0	0.0
7	1.0	2.0	3.0	0.0

Table 7.2 Jacobi's Method

n	x_1	x_2	x_3	x_4
1	0.3	1.5	2.7	–0.9
2	0.78	1.74	2.7	–0.18
3	0.9	1.908	2.916	–0.108
4	0.9624	1.9608	2.9592	–0.036
5	0.9845	1.9848	2.9851	–0.0158
6	0.9939	1.9938	2.9938	–0.006
7	0.9975	1.9975	2.9976	–0.0025
8	0.9990	1.9990	2.9990	–0.0010
9	0.9996	1.9996	2.9996	–0.0004
10	0.9998	1.9998	2.9998	–0.0002
11	0.9999	1.9999	2.9999	–0.0001
12	1.0	2.0	3.0	0.0

From Tables 7.1 and 7.2, it is clear that twelve iterations are required by Jacobi's method to achieve the same accuracy as seven Gauss–Seidel iterations.

Example 7.14 Solve the system

$$\begin{aligned} 6x + y + z &= 20 \\ x + 4y - z &= 6 \\ x - y + 5z &= 7 \end{aligned}$$

using both Jacobi and Gauss–Seidel methods.

(a) *Jacobi's method*

We rewrite the given system as

$$x = \frac{20}{6} - \frac{1}{6}y - \frac{1}{6}z = 3.3333 - 0.1667y - 0.1667z$$

$$y = 1.5 - 0.25x + 0.25z$$

$$z = 1.4 - 0.2x + 0.2y$$

In matrix form, the above system may be written as

$$X = C + BX$$

where

$$C = \begin{bmatrix} 3.3333 \\ 1.5 \\ 1.4 \end{bmatrix}, \; B = \begin{bmatrix} 0 & -0.1667 & -0.1667 \\ -0.25 & 0 & 0.25 \\ -0.2 & 0.2 & 0 \end{bmatrix} \text{ and } X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Assuming

$$X^0 = \begin{bmatrix} 3.3333 \\ 1.5 \\ 1.4 \end{bmatrix}, \text{ we obtain}$$

$$X^{(1)} = \begin{bmatrix} 3.3333 \\ 1.5 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 0 & -0.1667 & -0.1667 \\ -0.25 & 0 & 0.25 \\ -0.2 & 0.2 & 0 \end{bmatrix} \begin{bmatrix} 3.3333 \\ 1.5 \\ 1.4 \end{bmatrix} = \begin{bmatrix} 2.8499 \\ 1.0167 \\ 1.0333 \end{bmatrix}$$

$$X^{(2)} = \begin{bmatrix} 3.3333 \\ 1.5 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 0 & -0.1667 & -0.1667 \\ -0.25 & 0 & 0.25 \\ -0.2 & 0.2 & 0 \end{bmatrix} \begin{bmatrix} 2.8499 \\ 1.0167 \\ 1.0333 \end{bmatrix} = \begin{bmatrix} 2.9647 \\ 1.0458 \\ 1.0656 \end{bmatrix}$$

Proceeding in this way, we obtain

$$X^{(8)} = \begin{bmatrix} 2.9991 \\ 1.0012 \\ 1.0010 \end{bmatrix} \quad \text{and} \quad X^{(9)} = \begin{bmatrix} 2.9995 \\ 1.0005 \\ 1.0004 \end{bmatrix}.$$

We, therefore, conclude that

$$X = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \text{ i.e., } \quad x = 3, \ y = 1 \quad \text{and} \quad z = 1.$$

(b) *Gauss–Seidel method*

As before, we obtain the first approximation as

$$X^{(1)} = \begin{bmatrix} 2.8499 \\ 1.0167 \\ 1.0333 \end{bmatrix}$$

Then

$$x^{(2)} = 3.3333 - 0.1667 \times 1.0167 - 0.1667 \times 1.0333 = 2.9916$$
$$y^{(2)} = 1.5 - 0.25 \times 2.9916 + 0.25 \times 1.0333 = 1.0104$$
$$z^{(2)} = 1.4 - 0.2 \times 2.9916 + 0.2 \times 1.0104 = 1.0038$$

Similarly, we find

$$x^{(3)} = 2.9975, \quad y^{(3)} = 1.0016, \quad z^{(3)} = 1.0008,$$
$$x^{(4)} = 2.9995, \quad y^{(4)} = 1.0003, \quad z^{(4)} = 1.0002,$$
$$x^{(5)} = 2.9998, \quad y^{(5)} = 1.0001, \quad z^{(5)} = 1.0001.$$

At this stage, we can conclude that

$$x = 3, \quad y = 1, \quad z = 1.$$

7.7 MATRIX EIGENVALUE PROBLEM

Let A be a square matrix of order n with elements a_{ij}. We wish to find a column vector X and a constant λ such that

$$AX = \lambda X \tag{7.67}$$

In Eq. (7.67), λ is called the *eigenvalue* and X is called the corresponding *eigenvector*.

The matrix Eq. (7.67), when written out in full, represents a set of homogeneous linear equations:

$$\left.\begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22}-\lambda)x_2 + \ldots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn}-\lambda)x_n &= 0. \end{aligned}\right\} \tag{7.68}$$

A nontrivial solution exists only when the coefficient determinant in (7.68) vanishes. Hence, we have

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn}-\lambda \end{vmatrix} = 0. \tag{7.69}$$

This equation, called the *characteristic equation* of the matrix A, is a polynomial equation of degree n in λ, the polynomial being called the *characteristic-polynomial* of A. If the roots of Eq. (7.69) be given by $\lambda_i (i = 1, 2, \ldots, n)$, then for each value of λ_i, there exists a corresponding X_i such that

$$AX_i = \lambda_i X_i. \tag{7.70}$$

The eigenvalues λ_i may be either *distinct* (i.e. all different) or *repeated*. The evaluation of eigenvectors in the case of the repeated roots is a much involved process and will not be attempted here. The set of all eigenvalues, λ_i, of a matrix A is called the *spectrum* of A and the largest of $|\lambda_i|$ is called the *spectral radius* of A.

The eigenvalues are obtained by solving the algebraic Eq. (7.69). This method, which is demonstrated in Example 7.15, is unsuitable for matrices of higher order and better methods must be applied. For symmetric matrices, in particular, several methods are available and a recent method, known as Householder's method, is described in a subsequent section.

In some practical applications, only the numerically largest eigenvalue and the corresponding eigenvector are required, and in Example 7.16, we will describe an iterative method to compute the largest eigenvalue. This method is easy of application and also well-suited for machine computations.

Example 7.15 Find the eigenvalues and eigenvectors of the matrix:

$$A = \begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix}.$$

The characteristic equation of this matrix is given by

$$\begin{vmatrix} 5-\lambda & 0 & 1 \\ 0 & -2-\lambda & 0 \\ 1 & 0 & 5-\lambda \end{vmatrix} = 0.$$

which gives $\lambda_1 = -2$, $\lambda_2 = 4$ and $\lambda_3 = 6$. The corresponding eigenvectors are obtained thus

(i) $\lambda_1 = -2$. Let the eigenvector be

$$X_1 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Then we have:

$$A\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = -2\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

which gives the equations

$$7x_1 + x_3 = 0 \quad \text{and} \quad x_1 + 7x_3 = 0$$

The solution is $x_1 = x_3 = 0$ with x_2 arbitrary. In particular, we take $x_2 = 1$ and the eigenvector is

$$X_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

(ii) $\lambda_2 = 4$. With

$$X_2 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

as the eigenvector, the equations are

$$x_1 + x_3 = 0 \quad \text{and} \quad -6x_2 = 0,$$

from which we obtain

$$x_1 = -x_3 \quad \text{and} \quad x_2 = 0.$$

We choose, in particular, $x_1 = 1/\sqrt{2}$ and $x_3 = -1/\sqrt{2}$ so that $x_1^2 + x_2^2 + x_3^2 = 1$. The eigenvector chosen in this way is said be *normalized*. We, therefore, have

$$X_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix}.$$

(iii) $\lambda_3 = 6$. If

$$X_3 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

is the required eigenvector, then the equations are

$$-x_1 + x_3 = 0$$

$$-8x_2 = 0$$

$$x_1 - x_3 = 0,$$

which give $x_1 = x_3$ and $x_2 = 0$.

Choosing $x_1 = x_3 = 1/\sqrt{2}$, the normalized eigenvector is given by

$$X_3 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}.$$

Example 7.16 Determine the largest eigenvalue and the corresponding eigenvector of the matrix

$$A = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Let the initial eigenvector be

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = X^{(0)}, \text{ say.}$$

Then we have

$$AX^{(0)} = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = X^{(1)}, \quad \text{say}$$

Hence an approximate eigenvalue is 1 and an approximate eigenvector is $X^{(1)}$. Hence we have

$$AX^{(1)} = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 7 \\ 3 \\ 0 \end{bmatrix} = 3 \begin{bmatrix} 2.3 \\ 1 \\ 0 \end{bmatrix}$$

from which we see that

$$X^{(2)} = \begin{bmatrix} 2.3 \\ 1 \\ 0 \end{bmatrix}$$

and that an approximate eigenvalue is 3.
Repeating the above procedure, we successively obtain

$$4 \begin{bmatrix} 2.1 \\ 1.1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2.2 \\ 1.1 \\ 0 \end{bmatrix}; \quad 4.4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}; \quad 4 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}.$$

It follows that the largest eigenvalue is 4 and the corresponding eigenvector is

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}.$$

7.7.1 Eigenvalues of a Symmetric Tridiagonal Matrix

Since symmetric matrices can be reduced to symmetric tridiagonal matrices, the determination of eigenvalues of a symmetric tridiagonal matrix is of particular interest. Let

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{12} & a_{22} & a_{23} \\ 0 & a_{23} & a_{33} \end{bmatrix}. \tag{7.71}$$

To obtain the eigenvalues of A_1, we form the determinant equation

$$|A_1| = \begin{vmatrix} a_{11} - \lambda & a_{12} & 0 \\ a_{12} & a_{22} - \lambda & a_{23} \\ 0 & a_{23} & a_{33} - \lambda \end{vmatrix} = 0.$$

Suppose that the above equation is written in the form

$$\phi_3(\lambda) = 0. \tag{7.72}$$

Expanding the determinant in terms of the third row, we obtain

$$\begin{aligned}\phi_3(\lambda) &= (a_{33}-\lambda)\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} - a_{23}\begin{vmatrix} a_{11}-\lambda & 0 \\ a_{12} & a_{23} \end{vmatrix} \\ &= (a_{33}-\lambda)\,\phi_2(\lambda) - a_{23}(a_{11}-\lambda)\,a_{23} \\ &= (a_{33}-\lambda)\,\phi_2(\lambda) - a_{23}^2\phi_1(\lambda) \\ &= 0. \end{aligned} \tag{7.73}$$

We, thus, obtain the recursion formula

$$\phi_0(\lambda) = 1 \tag{7.74}$$

$$\begin{aligned}\phi_1(\lambda) &= a_{11}-\lambda \\ &= (a_{11}-\lambda)\,\phi_0(\lambda)\end{aligned} \tag{7.75}$$

$$\begin{aligned}\phi_2(\lambda) &= \begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{12} & a_{22}-\lambda \end{vmatrix} \\ &= (a_{11}-\lambda)\,(a_{22}-\lambda) - a_{12}^2 \\ &= \phi_1(\lambda)\,(a_{22}-\lambda) - a_{12}^2\,\phi_0(\lambda)\end{aligned} \tag{7.76}$$

$$\phi_3(\lambda) = \phi_2(\lambda)\,(a_{33}-\lambda) - a_{23}^2\,\phi_1(\lambda). \tag{7.77}$$

In general, if

$$\phi_k(\lambda) = \begin{vmatrix} a_{11}-\lambda & a_{12} & 0 & \dots & 0 \\ a_{12} & a_{22}-\lambda & a_{23} & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & a_{k-1,k} & a_{kk}-\lambda \end{vmatrix}, \qquad (2 \le k \le n), \tag{7.78}$$

then the recursion formula is

$$\phi_k(\lambda) = (a_{kk}-\lambda)\,\phi_{k-1}(\lambda) - a_{k-1,k}^2\,\phi_{k-2}(\lambda), \qquad (2 \le k \le n). \tag{7.79}$$

The equation $\phi_k(\lambda) = 0$ is the characteristic equation and can be solved by one of the methods described in Chapter 2. We might therefore consider the problem as solved, but we would like to remark that the sequence $\{\phi_k(\lambda), 0 \le k \le n\}$ has special properties which make it a *Sturm sequence* and from these properties one can isolate the eigenvalues of A_1. Once the eigenvalues have been isolated, one of the methods of Chapter 2 can be used to calculate the roots rapidly. The theory of Sturm sequences will not be discussed here, but the interested reader may refer to Henrici [1974] for details. When the eigenvalues of the tridiagonal matrix are known its eigenvectors can be calculated by the general method of solving a homogeneous system.

We next describe Householder's method for reducing a real symmetric matrix to a tridiagonal form.

7.7.2 Householder's Method

To describe this method, we consider a third order real symmetric matrix A given by

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \tag{7.80}$$

The method consists in finding a *real symmetric orthogonal* matrix P such that PAP is a symmetric tridiagonal matrix, i.e.

$$PAP = \begin{bmatrix} a_{11} & a'_{12} & 0 \\ a'_{12} & a'_{22} & a'_{23} \\ 0 & a'_{23} & a'_{33} \end{bmatrix} \tag{7.81}$$

where the primes denote that the elements have changed. Householder suggests that P could be of the form

$$P = I - 2VV^{\mathrm{T}}, \tag{7.82}$$

where

$$V = [0,\ v_2,\ v_3]^{\mathrm{T}} \quad \text{and} \quad V^{\mathrm{T}}V = I \tag{7.83}$$

The matrix equation in (7.83) means that

$$v_2^2 + v_3^2 = 1. \tag{7.84}$$

It follows that

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - 2v_2^2 & -2v_2v_3 \\ 0 & -2v_2v_3 & 1 - 2v_3^2 \end{bmatrix} \tag{7.85}$$

It can be verified that P is a symmetric orthogonal matrix as required. By direct multiplications, we find PAP and equating it to Eq. (7.81) and after some manipulations, we obtain

$$v_2^2 = \frac{1}{2}\left[1 \mp \frac{a_{12}}{\sqrt{a_{12}^2 + a_{13}^2}}\right] \tag{7.86}$$

and

$$v_3 = \mp \frac{a_{13}}{2v_2\sqrt{a_{12}^2 + a_{13}^2}} \tag{7.87}$$

In Eq. (7.86), if the sign chosen is the same as that of a_{12}, then v_2 would have a larger value and v_3 can be computed from Eq. (7.87). See Wilkinson [1960].

Example 7.17 Reduce the matrix

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 3 & 2 & -1 \\ 4 & -1 & 1 \end{bmatrix}$$

to the tridiagonal form.

Here

$$S = \sqrt{a_{12}^2 + a_{13}^2} = 5.$$

From Eq. (7.86), we obtain

$$v_2^2 = \frac{1}{2}\left(1 + \frac{3}{5}\right), \qquad \text{since } a_{12} \text{ is positive.}$$

and so

$$v_2 = \frac{2}{\sqrt{5}}.$$

Equation (7.87) now gives

$$v_3 = \frac{4}{2(2/\sqrt{5})\,(5)} = \frac{1}{\sqrt{5}}.$$

Thus,

$$V = \begin{bmatrix} 0 & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}^{\mathrm{T}}$$

and

$$P_1 = I - 2VV^{\mathrm{T}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3/5 & -4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix}.$$

Hence we have

$$A_1 = P_1 A P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3/5 & -4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix} \begin{bmatrix} 1 & 3 & 4 \\ 3 & 2 & -1 \\ 4 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -3/5 & -4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -5 & 0 \\ -5 & 2/5 & 1/5 \\ 0 & 1/5 & 13/5 \end{bmatrix}, \quad \text{on simplification.}$$

7.7.3 *QR* Method

This is the most efficient and widely used general method for the computation of all the eigenvalues of a general nonsymmetric matrix. Originally, due to J.G.F. Francis, the method is quite complicated and, therefore, only a brief presentation is given below.

Let $A_1 = A$ be the given matrix. Suppose that A_1 is factorized into the form

$$A_1 = Q_1 R_1 \tag{7.88}$$

where Q_1 is an orthogonal matrix and R_1 is an upper triangular matrix.

Therefore,

$$Q_1^{-1} A_1 = R_1 \tag{7.89}$$

The essential feature of this method is to find orthogonal matrices $P_1 P_2 \ldots P_{n-1}$ such that

$$P_{n-1} P_{n-2} \ldots P_2 P_1 A_1 = R_1 \tag{7.90}$$

The matrices P are of the form $I - 2VV^{\mathrm{T}}$ such that $P_1 A_1$ will contain zeros below the diagonal in its first column, $P_2\, P_1 A_1$ will contain zeros in its second column below the diagonal, and so on. If we carry out this procedure with each column of A_1, then the final result will be R_1, which is an upper triangular matrix. The sequence $\{A_i\}$ converges either to a triangular matrix or to a near triangular matrix. In either case, the eigenvalues can be computed easily.

7.8 SINGULAR VALUE DECOMPOSITION

We have so far considered square matrices only and in Section 7.5.6 we obtained the LU decomposition of a square matrix. For rectangular matrices, a similar decomposition is possible and this is called the *singular value decomposition* (SVD). This decomposition is of great importance in matrix theory since it is useful in finding the inverse of a singular matrix, called the *generalized inverse.*

Let A be an $(m \times n)$ matrix with $m \geq n$. Then we know that the matrices $A^{\mathrm{T}}A$ and AA^{T} are both non-negative and symmetric. Their eigenvalues are also identical. Let the eigenvalues of $A^{\mathrm{T}}A$ be $\lambda_1, \lambda_2, \ldots \lambda_n$ with corresponding orthonormalized eigenvectors $X_1, X_2, \ldots, X_n$. Let these eigenvectors be the columns of the matrix V.

Therefore,

$$A^{\mathrm{T}} A X_n = \lambda_n X_n \tag{7.91}$$

Similarly, let Y_n the orthonormalized eigenvectors of AA^{T}, so that we have

$$AA^{\mathrm{T}} Y_n = \lambda_n Y_n \tag{7.92}$$

Then X_n and Y_n are related through the equation

$$Y_n = \frac{1}{\sqrt{\lambda_n}} \cdot AX_n \tag{7.93}$$

A can be decomposed into the form

$$A = UDV^{\mathrm{T}}, \tag{7.94}$$

where U is the matrix whose columns are the eigenvectors Y_n,

$$D = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right), \tag{7.95}$$

and

$$U^{\mathrm{T}}U = V^{\mathrm{T}}V = VV^{\mathrm{T}} = I_n \tag{7.96}$$

The decomposition defined by Eq. (7.94) is called the *singular value decomposition* (SVD) of the matrix A.

Example 7.18 Obtain the singular-value decomposition of

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 3 \end{bmatrix}.$$

We have

$$A^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} \quad \text{and} \quad A^{\mathrm{T}}A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

The eigenvalues of $A^{\mathrm{T}}A$ are given by $\lambda_1 = 16.64$ and $\lambda_2 = 0.36$. For the corresponding eigenvectors, we have

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 16.64\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

which gives the system:

$$3x_1 + 6x_2 = 16.64x_1$$

$$6x_1 + 14x_2 = 16.64x_2.$$

The solution is given by

$$x_1 = \begin{bmatrix} 0.4033 \\ 0.9166 \end{bmatrix}.$$

Again, we have

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.36\begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

which gives the system

$$3x_1 + 6x_2 = 0.36x_1$$
$$6x_1 + 14x_2 = 0.36x_2.$$

The solution is

$$x_2 = \begin{bmatrix} 0.9166 \\ -0.4033 \end{bmatrix}$$

we also have $\sqrt{\lambda_1} = 4.080$ and $\sqrt{\lambda_2} = 0.60$.

The eigenvectors of AA^T can then be obtained from Eq. (7.93). These are given by

$$\begin{bmatrix} 0.5480 \\ 0.3235 \\ 0.7727 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.1833 \\ 0.8555 \\ -0.4889 \end{bmatrix}.$$

The singular-value decomposition of A is then given by

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 0.5480 & 0.1833 \\ 0.3235 & 0.8555 \\ 0.7727 & -0.4889 \end{bmatrix} \begin{bmatrix} 4.080 & 0 \\ 0 & 0.60 \end{bmatrix} \begin{bmatrix} 0.4033 & 0.9166 \\ 0.9166 & -0.4033 \end{bmatrix}.$$

EXERCISES

7.1. Express the following systems of equations in the matrix form:

(a) $3x + 2y + 4z = 7$
$2x + y + z = 7$
$x + 3y + 5z = 2$

(b) $2x - z - 2u = -8$
$y + 2z - u = -1$
$x - y - u = -6$
$-x + 3y - 2u = 7$

7.2 Write an algorithm to compute the product $A = BC$, where B and C are matrices of sizes $(p \times q)$ and $(q \times r)$, respectively.

If $B = \begin{bmatrix} 2 & 5 & -2 \\ -1 & 0 & 0 \\ 2 & 3 & 4 \end{bmatrix}$ and $C = \begin{bmatrix} 3 & 5 \\ 1 & 0 \\ 2 & 0 \end{bmatrix}$, find $A = BC$.

7.3 Show that the product of two upper triangular matrices is also an upper triangular matrix.

If $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} -1 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{bmatrix}$, find AB.

7.4 Explain the back substitution process for the solution of the system

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ & \vdots & & \\ 0 & 0 & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

and give an algorithm to implement it. Solve the following system

$$\begin{aligned} 2x_1 - 3x_2 + x_3 &= -1 \\ -3x_2 - x_3 &= -9 \\ 5x_3 &= 15 \end{aligned}$$

7.5 Find the inverse of the matrix

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 1.5 & 2.5 \\ 0 & 0 & 18 \end{bmatrix}$$

7.6 Find the inverse of the matrix

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 2 \end{bmatrix}$$

7.7 Factorize the matrix

$$A = \begin{bmatrix} 4 & 3 & -1 \\ 1 & 1 & 1 \\ 3 & 5 & 3 \end{bmatrix}$$

into the product LU where L is unit lower triangular and U upper triangular.

7.8 *Crout's Decomposition* If a matrix A is decomposed into the product of a lower triangular matrix and a unit upper triangular matrix, it is called *Crout's decomposition.*
Decompose the matrix.

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 3 & -1 \\ 3 & 5 & 3 \end{bmatrix}$$

by Crout's method.

7.9 Define norm of a matrix.
If

$$A = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 4 & 3 \\ 1 & 3 & 2 \end{bmatrix}$$

find $\| A \|_1$, $\| A \|_e$ and $\| A \|_\infty$.

7.10 Use Gauss elimination with partial pivoting to solve the system

$$2x_1 + x_2 - x_3 = -1$$
$$x_1 - 2x_2 + 3x_3 = 9$$
$$3x_1 - x_2 + 5x_3 = 14$$

Check your answer by substitution into the original equations.

7.11 Explain Gauss elimination method with partial pivoting to solve a system of linear algebraic equations and apply it to solve the system

$$1.2x_1 + 2.1x_2 - 1.1x_3 = 1.8776$$
$$-1.1x_1 + 2.0x_2 + 3.1x_3 = -0.1159$$
$$-2.1x_1 - 2.2x_2 + 3.7x_3 = -4.2882.$$

7.12 Use Gauss–Jordan method to solve the system

$$4x_1 + 3x_2 - x_3 = 6$$
$$3x_1 + 5x_2 + 3x_3 = 4$$
$$x_1 + x_2 + x_3 = 1.$$

7.13 Use Gauss elimination to find the inverse of the matrix

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -2 & 4 \\ 1 & 2 & 2 \end{bmatrix}$$

7.14 Solve the system of equations

$$10x + y + z = 12$$
$$2x + 10y + z = 13$$
$$x + y + 3z = 5$$

by both Gauss elimination and Gauss-Jordan methods.

7.15 Decompose the matrix

$$A = \begin{bmatrix} 5 & -2 & 1 \\ 7 & 1 & -5 \\ 3 & 7 & 4 \end{bmatrix}$$

into the form LU where L is unit lower triangular and U an upper triangular matrix. Hence solve the system $AX = b$ where $b = [4\ 8\ 10]^T$.

7.16 For the matrix in Problem 7.15, find L^{-1}, U^{-1} and A^{-1}.

7.17 Design an algorithm to reduce a given system of equations to upper triangular form. Test your algorithm on the system:

$$4x + 3y + 2z = 16$$
$$2x + 3y + 4z = 20$$
$$x + 2y + z = 8.$$

7.18 Decompose the matrix

$$A = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 3 & 4 \\ 1 & 2 & 1 \end{bmatrix}$$

into the form LU, where L is a lower triangular matrix and U is unit upper triangular.

7.19 Solve the following system by Gauss elimination:

$$2x_1 + 3x_2 - x_3 + 2x_4 = 7$$
$$x_1 + x_2 + x_3 + x_4 = 2$$
$$x_1 + x_2 + 3x_3 - 2x_4 = -6$$
$$x_1 + 2x_2 + x_3 - x_4 = -2$$

7.20 An approximate solution of the system

$$10x_1 + x_2 + x_3 = 12$$
$$x_1 + 10x_2 + x_3 = 12$$
$$x_1 + x_2 + 10x_3 = 12$$

is given as

$$x_1^{(0)} = 0.4,\ x_2^{(0)} = 0.6 \text{ and } x_3^{(0)} = 0.8.$$

Use the iterative method of Section 7.5.11 to improve this solution.

7.21 Solve the system

$$10x + 2y + z = 9$$
$$2x + 20y - 2z = -44$$
$$-2x + 3y + 10z = 22$$

by Jacobi's method.

7.22 Solve the system given in Problem 7.21 by Gauss–Seidel method.

7.23 State the condition of convergence of Gauss-Seidel Iterative method. Apply this method, upto six iterations, to solve the system defined by

$$28x + 4y - z = 32$$
$$2x + 17y + 4z = 35$$
$$x + 3y + 10z = 24$$

7.24 *Cholesky's method* A matrix A is said to be a *symmetric* matrix if $a_{ij} = a_{ji}$. Symmetric systems occur frequently in engineering and science. For symmetric systems, the LU decomposition is more conveniently obtained, since $U = L^{\mathrm{T}}$. This method, called Cholesky's method, requires less storage space and less computer time. Solve the system given by

$$\begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -4 \\ 16 \end{bmatrix}$$

by Cholesky's method.

7.25 Explain what is meant by ill-conditioning of a matrix. Give two examples of ill-conditioned matrices.

Is the matrix

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$$

ill-conditioned?

7.26 Define norm of a matrix. List the different types of norms of a matrix. What is condition number of a matrix.? Explain how the condition number is useful in determining whether a matrix is ill-conditioned.

7.27 *Centro-symmetric systems* Equations of the type

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = b_1 \quad \text{(i)}$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = b_2 \quad \text{(ii)}$$
$$a_{24}x_1 + a_{23}x_2 + a_{22}x_3 + a_{21}x_4 = b_3 \quad \text{(iii)}$$
$$a_{14}x_1 + a_{13}x_2 + a_{12}x_3 + a_{11}x_4 = b_4 \quad \text{(iv)}$$

are called centro-symmetric systems. Such systems can be solved easily by the following method. Adding (i) and (iv), we obtain an

equation for $(x_1 + x_4)$ and $(x_2 + x_3)$. Similarly, adding (ii) and (iii), we obtain another equation for $(x_1 + x_4)$ and $(x_2 + x_3)$. These two equations give the values of $(x_1 + x_4)$ and $(x_2 + x_3)$. Again by subtractions, we get two equations in $(x_1 - x_4)$ and $(x_2 - x_3)$. Computation of x_1, x_4 and x_2, x_3 is now fairly easy.

Solve the system

$$\begin{aligned} x_1 + x_2 + 3x_3 - 2x_4 &= -6 \\ 2x_1 + 3x_2 - x_3 + 2x_4 &= 7 \\ 2x_1 - x_2 + 3x_3 + 2x_4 &= 3 \\ -2x_1 + 3x_2 + x_3 + x_4 &= -1 \end{aligned}$$

7.28 Determine the eigenvalues and the corresponding eigenvectors for the following matrices:

(a) $\begin{bmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 5 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 5 \end{bmatrix}$

7.29 Use the iterative method to find the smallest eigenvalue of the matrix

$$A = \begin{bmatrix} 1 & 6 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

7.30 Determine the largest eigenvalue and the corresponding eigenvector of the matrix

$$B = \begin{bmatrix} 1 & 3 & -1 \\ 3 & 2 & 4 \\ -1 & 4 & 10 \end{bmatrix}$$

7.31 Using Householder's method, obtain the tridiagonal form of the matrix

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 3 & 1 & 2 \\ 4 & 2 & 1 \end{bmatrix}$$

7.32 Given the matrix

$$A = \begin{bmatrix} 0 & 1 & 4 \\ 1 & 3 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

find a symmetric orthogonal matrix P such that the matrix PA will contain zeros below the diagonal in its first column.

Answers to Exercises

7.1 (a) $\begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 1 \\ 1 & 3 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \\ 2 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & 0 & ?\,1 & ?\,2 \\ 0 & 1 & 2 & ?\,1 \\ 1 & ?\,1 & 0 & ?\,1 \\ ?\,1 & 3 & 0 & ?\,2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ u \end{bmatrix} = \begin{bmatrix} ?\,8 \\ ?\,1 \\ ?\,6 \\ 7 \end{bmatrix}$

7.2 $\begin{bmatrix} 7 & 10 \\ -3 & -5 \\ 17 & 10 \end{bmatrix}$

7.4 $x_1 = 1,\ x_2 = 2,\ x_3 = 3$

7.5 $A^{-1} = \begin{bmatrix} \frac{1}{2} & -3 & \frac{7}{18} \\ 0 & +2 & -\frac{5}{18} \\ 0 & 0 & \frac{1}{18} \end{bmatrix}$

7.6 $L^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ -1 & 1 & 0 \\ \frac{1}{4} & -1 & \frac{1}{2} \end{bmatrix}$

7.7 $L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{3}{4} & 11 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & 3 & -1 \\ 0 & \frac{1}{4} & \frac{5}{4} \\ 0 & 0 & -10 \end{bmatrix}$

7.8 $\begin{bmatrix} 1 & 0 & 0 \\ 4 & -1 & 0 \\ 3 & 2 & -10 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix}$

7.9 $\| A \|_1 = 10, \quad \| A \|_e = 8.6602, \quad \| A \|_\infty = 9.$

7.10 $x_1 = 1, \quad x_2 = -1, \quad x_3 = 2.$

7.11 $x_1 = -2.1557, \quad x_2 = 1.2746, \quad x_3 = -1.6246.$

7.12 $x_1 = 1.0, \quad x_2 = 0.5, \quad x_3 = -0.5.$

7.13 $A^{-1} = \begin{bmatrix} 1.2 & -0.4 & 0.2 \\ -0.2 & -0.1 & 0.3 \\ -0.4 & 0.3 & 0.1 \end{bmatrix}$

7.14 $x = 1, \quad y = 1, \quad z = 1.$

7.15 $x_1 = 1.1193, \quad x_2 = 0.8685, \quad x_3 = 0.1407.$

7.16 $L^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1.4 & 1 & 0 \\ 2.4211 & -2.1579 & 1 \end{bmatrix}, \quad U^{-1} = \begin{bmatrix} 0.2 & -0.1053 & -0.0507 \\ 0 & 0.2632 & 0.0978 \\ 0 & 0 & 0.0581 \end{bmatrix}$

7.17 $x = 1, \quad y = 2, \quad z = 3.$

7.18 $L = \begin{bmatrix} 4 & 0 & 0 \\ 2 & \frac{3}{2} & 0 \\ 1 & \frac{5}{4} & -2 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & \frac{3}{4} & \frac{1}{2} \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$

7.19 $x_4 = 2, \quad x_3 = -1, \quad x_2 = 0, \quad x_1 = 1.$

7.20 $x_1 = x_2 = x_3 = 1.$

7.21 5th iteration values are 0.9989, –1.9993, 2.9984.

7.22 4th iteration values are 0.9991, –1.9998, 2.9998.

7.23 6th iteration values are 0.9936, 1.5070, 1.8485.

7.24 $x_3 = 3, \quad x_2 = 2, \quad x_1 = 1.$

7.25 ill-conditioned

7.27 $x_1 = 1, \quad x_2 = 0, \quad x_3 = -1, \quad x_4 = 2.$

7.28 (a) $\lambda = 0, \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix}; \quad \lambda = 3, \begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix}$

(b) $\lambda = 0, \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}; \quad \lambda = 1, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad \lambda = 2, \begin{bmatrix} 0 \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

(c) $\lambda = -2, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \ \lambda = 4, \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix}; \ \lambda = 6, \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$

7.29 $\lambda = -1, \begin{bmatrix} 1 \\ -0.33 \\ 0 \end{bmatrix}$

7.30 $\lambda = 10.6, \begin{bmatrix} 0.02 \\ 0.46 \\ 1.0 \end{bmatrix}$

7.31 $A_1 = \begin{bmatrix} 1 & -5 & 0 \\ -5 & 2.92 & 0.56 \\ 0 & 0.56 & -0.92 \end{bmatrix}$

7.32 $\begin{bmatrix} 0 & 1 & 4 \\ 0 & 2 & 0.8 \\ 0 & -1 & -0.4 \end{bmatrix}$

Chapter 8

Numerical Solution of Ordinary Differential Equations

8.1 INTRODUCTION

Many problems in science and engineering can be reduced to the problem of solving differential equations satisfying certain given conditions. The analytical methods of solution, with which the reader is assumed to be familiar, can be applied to solve only a selected class of differential equations. Those equations which govern physical systems do not possess, in general closed-form solutions, and hence recourse must be made to numerical methods for solving such differential equations.

To describe various numerical methods for the solution of ordinary differential equations, we consider the general first order differential equation

$$\frac{dy}{dx} = f(x, y) \tag{8.1a}$$

with the initial condition,

$$y(x_0) = y_0 \tag{8.1b}$$

and illustrate the theory with respect to this equation. The methods so developed can, in general, be applied to the solution of systems of first-order equations, and will yield the solution in one of the two forms:

(i) A series for y in terms of powers of x, from which the value of y can be obtained by direct substitution.

(ii) A set of tabulated values of x and y.

The methods of Taylor and Picard belong to class (i), whereas those of Euler, Runge–Kutta, Adams–Bashforth, etc., belong to class (ii). These latter

methods are called *step-by-step* methods or *marching* methods because the values of y are computed by short steps ahead for equal intervals h of the independent variable. In the methods of Euler and Runge–Kutta, the interval length h should be kept small and hence these methods can be applied for tabulating y over a limited range only. If, however, the function values are desired over a wider range, the methods due to Adams–Bashforth, Adams–Moulton, Milne, etc., may be used. These methods use finite-differences and require 'starting values' which are usually obtained by Taylor's series or Runge–Kutta methods.

It is well-known that a differential equation of the nth order will have n arbitrary constants in its general solution. In order to compute the numerical solution of such an equation, we therefore need n conditions. Problems in which all the initial conditions are specified at the *initial* point only are called *initial value problems*. For example, the problem defined by Eqs. (8.1) is an *initial value problem*. On the other hand, in problems involving second-and higher-order differential equations, we may prescribe the conditions at two or more points. Such problems are called *boundary value problems*.

We shall first describe methods for solving initial value problems of the type (8.1), and at the end of the chapter we will outline methods for solving boundary value problems for second-order differential equations.

8.2 SOLUTION BY TAYLOR'S SERIES

We consider the differential equation

$$y' = f(x, y) \tag{8.1a}$$

with the initial condition

$$y(x_0) = y_0. \tag{8.1b}$$

If $y(x)$ is the exact solution of Eq. (8.1), then the Taylor's series for $y(x)$ around $x = x_0$ is given by

$$y(x) = y_0 + (x - x_0)y_0' + \frac{(x - x_0)^2}{2!}y_0'' + \cdots \tag{8.2}$$

If the values of $y_0', y_0'', \ldots$ are known, then Eq. (8.2) gives a power series for y. Using the formula for total derivatives, we can write

$$y'' = f' = f_x + y'f_y = f_x + ff_y,$$

where the suffixes denote partial derivatives with respect to the variable concerned. Similarly, we obtain

$$\begin{aligned} y''' = f'' &= f_{xx} + f_{xy}f + f(f_{yx} + f_{yy}f) + f_y(f_x + f_y f) \\ &= f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_x f_y + ff_y^2 \end{aligned}$$

and other higher derivatives of y. The method can easily be extended to simultaneous and higher-order differential equations.

Example 8.1 From the Taylor series for $y(x)$, find $y(0.1)$ correct to four decimal places if $y(x)$ satisfies

$$y' = x - y^2 \quad \text{and} \quad y(0) = 1.$$

The Taylor series for $y(x)$ is given by

$$y(x) = 1 + xy_0' + \frac{x^2}{2}y_0'' + \frac{x^3}{6}y_0''' + \frac{x^4}{24}y_0^{\text{iv}} + \frac{x^5}{120}y_0^{\text{v}} + \cdots$$

The derivatives $y_0', y_0'', \ldots$ etc. are obtained thus:

$$y'(x) = x - y^2 \qquad y_0' = -1$$

$$y''(x) = 1 - 2yy' \qquad y_0'' = 3$$

$$y'''(x) = -2yy'' - 2y'^2 \qquad y_0''' = -8$$

$$y^{\text{iv}}(x) = -2yy''' - 6y'y'' \qquad y_0^{\text{iv}} = 34$$

$$y^{\text{v}}(x) = -2yy^{\text{iv}} - 8y'y''' - 6y''^2 \qquad y_0^{\text{v}} = -186$$

Using these values, the Taylor series becomes

$$y(x) = 1 - x + \frac{3}{2}x^2 - \frac{4}{3}x^3 + \frac{17}{12}x^4 - \frac{31}{20}x^5 + \cdots$$

To obtain the value of $y(0.1)$ correct to four decimal places, it is found that the terms up to x^4 should be considered, and we have $y(0.1) = 0.9138$.

Suppose that we wish to find the range of values of x for which the above series, truncated after the term containing x^4, can be used to compute the values of y correct to four decimal places. We need only to write

$$\frac{31}{20}x^5 \le 0.00005 \quad \text{or} \quad x \le 0.126.$$

Example 8.2 Given the differential equation

$$y'' - xy' - y = 0$$

with the conditions $y(0) = 1$ and $y'(0) = 0$, use Taylor's series method to determine the value of $y(0.1)$.

We have $y(x) = 1$ and $y'(x) = 0$ when $x = 0$. The given differential equation is

$$y''(x) = xy'(x) + y(x) \tag{i}$$

Hence $y''(0) = y(0) = 1$. Successive differentiation of (i) gives

$$y'''(x) = xy''(x) + y'(x) + y'(x) = xy''(x) + 2y'(x), \tag{ii}$$

$$y^{\text{iv}}(x) = xy'''(x) + y''(x) + 2y''(x) = xy'''(x) + 3y''(x), \tag{iii}$$

$$y^{\text{v}}(x) = xy^{\text{iv}}(x) + y'''(x) + 3y'''(x) = xy^{\text{iv}}(x) + 4y'''(x), \tag{iv}$$

$$y^{\text{vi}}(x) = xy^{\text{v}}(x) + y^{\text{iv}}(x) + 4y^{\text{iv}}(x) = xy^{\text{v}}(x) + 5y^{\text{iv}}(x), \tag{v}$$

and similarly for higher derivatives. Putting $x = 0$ in (ii) to (v), we obtain

$$y'''(0) = 2y'(0) = 0, \quad y^{\text{iv}}(0) = 3y''(0) = 3, \quad y^{\text{v}}(0) = 0, \quad y^{\text{vi}}(0) = 5.$$

By Taylor's series, we have

$$y(x) = y(0) + xy'(0) + \frac{x^2}{2}y''(0) + \frac{x^3}{6}y'''(0) + \frac{x^4}{24}y^{\text{iv}}(0)$$
$$+ \frac{x^5}{120}y^{\text{v}}(0) + \frac{x^6}{720}y^{\text{vi}}(0) + \cdots$$

Hence

$$y(0.1) = 1 + \frac{(0.1)^2}{2} + \frac{(0.1)^4}{24}(3) + \frac{(0.1)^6}{720}(5) + \cdots$$
$$= 1 + 0.005 + 0.0000125, \text{ neglecting the last term}$$
$$= 1.0050125, \text{ correct to seven decimal places.}$$

8.3 PICARD'S METHOD OF SUCCESSIVE APPROXIMATIONS

Integrating the differential equation given in Eq. (8.1), we obtain

$$y = y_0 + \int_{x_0}^{x} f(x, y)\, dx. \tag{8.3}$$

Equation (8.3), in which the unknown function y appears under the integral sign, is called an *integral equation*. Such an equation can be solved by the method of successive approximations in which the first approximation to y is obtained by putting y_0 for y on right side of Eq. (8.3), and we write

$$y^{(1)} = y_0 + \int_{x_0}^{x} f(x, y_0)\, dx$$

The integral on the right can now be solved and the resulting $y^{(1)}$ is substituted for y in the integrand of Eq. (8.3) to obtain the second approximation $y^{(2)}$:

$$y^{(2)} = y_0 + \int_{x_0}^{x} f(x, y^{(1)})\, dx$$

Proceeding in this way, we obtain $y^{(3)}, y^{(4)}, \ldots, y^{(n-1)}$ and $y^{(n)}$, where

$$y^{(n)} = y_0 + \int_{x_0}^{x} f(x, y^{(n-1)})\, dx \qquad \text{with } y^{(0)} = y_0 \tag{8.4}$$

Hence this method yields a sequence of approximations $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$ and it can be proved (*see*, for example, the book by Levy and Baggot) that if the function $f(x, y)$ is bounded in some region about the point (x_0, y_0) and if $f(x, y)$ satisfies the *Lipschitz condition*, viz.,

$$|f(x, y) - f(x, \bar{y})| \le K\,|y - \bar{y}| \qquad K \text{ being a constant} \tag{8.5}$$

then the sequence $y^{(1)}, y^{(2)}, \ldots$ converges to the solution of Eq. (8.1).

Example 8.3 Solve the equation $y' = x + y^2$, subject to the condition $y = 1$ when $x = 0$.

We start with $y^{(0)} = 1$ and obtain

$$y^{(1)} = 1 + \int_0^x (x+1)\,dx = 1 + x + \frac{1}{2}x^2.$$

Then the second approximation is

$$y^{(2)} = 1 + \int_0^x \left[x + \left(1 + x + \frac{1}{2}x^2\right)^2 \right] dx$$

$$= 1 + x + \frac{3}{2}x^2 + \frac{2}{3}x^3 + \frac{1}{4}x^4 + \frac{1}{20}x^5.$$

It is obvious that the integrations might become more and more difficult as we proceed to higher approximations.

Example 8.4 Given the differential equation

$$\frac{dy}{dx} = \frac{x^2}{y^2 + 1}$$

with the initial condition $y = 0$ when $x = 0$, use Picard's method to obtain y for $x = 0.25$, 0.5 and 1.0 correct to three decimal places.

We have

$$y = \int_0^x \frac{x^2}{y^2 + 1}\,dx.$$

Setting $y^{(0)} = 0$, we obtain

$$y^{(1)} = \int_0^x x^2 dx = \frac{1}{3}x^3$$

and

$$y^{(2)} = \int_0^x \frac{x^2}{(1/9)\,x^6 + 1}\,dx = \tan^{-1}\left(\frac{1}{3}x^3\right) = \frac{1}{3}x^3 - \frac{1}{81}x^9 + \cdots$$

so that $y^{(1)}$ and $y^{(2)}$ agree to the first term, viz., $(1/3)x^3$. To find the range of values of x so that the series with the term $(1/3)x^3$ alone will give the result correct to three decimal places, we put

$$\frac{1}{81}x^9 \le 0.0005$$

which yields

$$x \le 0.7$$

Hence

$$y(0.25) = \frac{1}{3}(0.25)^3 = 0.005$$

$$y(0.5) = \frac{1}{3}(0.5)^3 = 0.042$$

$$y(1.0) = \frac{1}{3} - \frac{1}{81} = 0.321$$

8.4 EULER'S METHOD

We have so far discussed the methods which yield the solution of a differential equation in the form of a power series. We will now describe the methods which give the solution in the form of a set of tabulated values.

Suppose that we wish to solve the Eqs. (8.1) for values of y at $x = x_r = x_0 + rh (r = 1, 2, \ldots)$. Integrating Eq. (8.1), we obtain

$$y_1 = y_0 + \int_{x_0}^{x_1} f(x, y)\,dx. \tag{8.6}$$

Assuming that $f(x, y) = f(x_0, y_0)$ in $x_0 \le x \le x_1$, this gives Euler's formula

$$y_1 \approx y_0 + hf(x_0, y_0). \tag{8.7a}$$

Similarly for the range $x_1 \le x \le x_2$, we have

$$y_2 = y_1 + \int_{x_1}^{x_2} f(x, y)\,dx.$$

Substituting $f(x_1, y_1)$ for $f(x, y)$ in $x_1 \le x \le x_2$ we obtain

$$y_2 \approx y_1 + hf(x_1, y_1). \tag{8.7b}$$

Proceeding in this way, we obtain the general formula

$$y_{n+1} = y_n + hf(x_n, y_n), \quad n = 0, 1, 2, \ldots \tag{8.8}$$

The process is very slow and to obtain reasonable accuracy with Euler's method, we need to take a smaller value for h. Because of this restriction

on h, the method is unsuitable for practical use and a modification of it, known as the *modified Euler method*, which gives more accurate results, will be described in Section 8.4.2.

Example 8.5 To illustrate Euler's method, we consider the differential equation $y' = -y$ with the condition $y(0) = 1$.

Successive application of Eq. (8.8) with $h = 0.01$ gives

$$y(0.01) = 1 + 0.1(-1) = 0.99$$

$$y(0.02) = 0.99 + 0.01(-0.99) = 0.9801$$

$$y(0.03) = 0.9801 + 0.01(-0.9801) = 0.9703$$

$$y(0.04) = 0.9703 + 0.01(-0.9703) = 0.9606.$$

The exact solution is $y = e^{-x}$ and from this the value at $x = 0.04$ is 0.9608.

8.4.1 Error Estimates for the Euler Method

Let the true solution of the differential equation at $x = x_n$ be $y(x_n)$ and also let the approximate solution be y_n. Now, expanding $y(x_{n+1})$ by Taylor's series, we get

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(x_n) + \cdots \\ &= y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(\tau_n), \qquad \text{where } x_n \le \tau_n \le x_{n+1}. \end{aligned} \tag{8.9}$$

We usually encounter two types of errors in the solution of differential equations. These are (i) local errors, and (ii) rounding errors. The local error is the result of replacing the given differential equation by means of the equation

$$y_{n+1} = y_n + hy'_n.$$

This error is given by

$$L_{n+1} = -\frac{1}{2} h^2 y''(\tau_n) \tag{8.10}$$

The total error is then defined by

$$e_n = y_n - y(x_n) \tag{8.11}$$

Since y_0 is exact, it follows that $e_0 = 0$.

Neglecting the rounding error, we write the total solution error as

$$\begin{aligned} e_{n+1} &= y_{n+1} - y(x_{n+1}) \\ &= y_n + hy'_n - \left[y(x_n) + hy'(x_n) - L_{n+1}\right] \\ &= e_n + h\left[f(x_n, y_n) - y'(x_n)\right] + L_{n+1}. \end{aligned}$$

$$\Rightarrow \qquad e_{n+1} = e_n + h\left[f(x_n, y_n) - f(x_n, y(x_n))\right] + L_{n+1}.$$

By mean value theorem, we write

$$f(x_n, y_n) - f(x_n, y(x_n)) = \left[y_n - y(x_n)\right]\frac{\partial f}{\partial y}(x_n, \xi_n), \quad y(x_n) \le \xi_n \le y_n.$$

Hence, we have

$$e_{n+1} = e_n\left[1 + hf_y(x_n, \xi_n)\right] + L_{n+1} \tag{8.12}$$

Since $e_0 = 0$, we obtain successively:

$$e_1 = L_1; \quad e_2 = \left[1 + hf_y(x_1, \xi_1)\right]L_1 + L_2;$$

$$e_3 = \left[1 + hf_y(x_2, \xi_2)\right]\left[1 + hf_y(x_1, \xi_1)\right](L_1 + L_2) + L_3; \text{ etc.}$$

See the book by Isaacson and Keller [1966] for more details.

Example 8.6 We consider, again, the differential equation $y' = -y$ with the condition $y(0) = 1$, which we have solved by Euler's method in Example 8.5.

Choosing $h = 0.01$, we have

$$1 + hf_y(x_n, \xi_n) = 1 + 0.01(-1) = 0.99.$$

and

$$L_{n+1} = -\frac{1}{2}h^2 y''(\rho_n) = -0.00005\, y(\rho_n).$$

In this problem, $y(\rho_n) \le y(x_n)$, since y' is negative. Hence we successively obtain

$$|L_1| \le 0.00005 = 5\times10^{-5},$$

$$|L_2| \le (0.00005)\,(0.99) < 5\times10^{-5},$$

$$|L_3| \le (0.00005)\,(0.9801) < 5\times10^{-5},$$

and so on. For computing the total solution error, we need an estimate of the rounding error. If we neglect the rounding error, i.e., if we set

$$R_{n+1} = 0,$$

then using the above bounds, we obtain from Eq. (8.12) the estimates

$$e_0 = 0,$$

$$|e_1| \le 5\times10^{-5}$$

$$|e_2| \le 0.99e_1 + 5\times10^{-5} < 10^{-4}$$

$$|e_3| \le 0.99e_2 + 5\times10^{-5} < 10^{-4} + 5\times10^{-5}$$

$$|e_4| \le 0.99e_3 + 5\times10^{-5} < 10^{-4} + 10^{-4} = 2\times10^{-4} = 0.0002$$

$$\vdots$$

It can be verified that the estimate for e_4 agrees with the actual error in the value of $y(0.04)$ obtained in Example 8.5.

8.4.2 Modified Euler's Method

Instead of approximating $f(x, y)$ by $f(x_0, y_0)$ in Eq. (8.6), we now approximate the integral given in Eq. (8.6) by means of trapezoidal rule to obtain

$$y_1 = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_1, y_1)] \tag{8.13}$$

We thus obtain the iteration formula

$$y_1^{(n+1)} = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_1, y_1^{(n)})], \quad n = 0, 1, 2, \ldots \tag{8.14}$$

where $y_1^{(n)}$ is the nth approximation to y_1. The iteration formula (8.14) can be started by choosing $y_1^{(0)}$ from Euler's formula:

$$y_1^{(0)} = y_0 + hf(x_0, y_0).$$

Example 8.7 Determine the value of y when $x = 0.1$ given that

$$y(0) = 1 \quad \text{and} \quad y' = x^2 + y$$

We take $h = 0.05$. With $x_0 = 0$ and $y_0 = 1.0$, we have $f(x_0, y_0) = 1.0$. Hence Euler's formula gives

$$y_1^{(0)} = 1 + 0.05(1) = 1.05$$

Further, $x_1 = 0.05$ and $f(x_1, y_1^{(0)}) = 1.0525$. The average of $f(x_0, y_0)$ and $f(x_1, y_1^{(0)})$ is 1.0262. The value of $y_1^{(1)}$ can therefore be computed by using Eq. (8.14) and we obtain

$$y_1^{(1)} = 1.0513.$$

Repeating the procedure, we obtain $y_1^{(2)} = 1.0513$. Hence we take $y_1 = 1.0513$, which is correct to four decimal places.

Next, with $x_1 = 0.05$, $y_1 = 1.0513$ and $h = 0.05$, we continue the procedure to obtain y_2, i.e., the value of y when $x = 0.1$. The results are

$$y_2^{(0)} = 1.1040, \quad y_2^{(1)} = 1.1055, \quad y_2^{(2)} = 1.1055.$$

Hence we conclude that the value of y when $x = 0.1$ is 1.1055.

8.5 RUNGE–KUTTA METHODS

As already mentioned, Euler's method is less efficient in practical problems since it requires h to be small for obtaining reasonable accuracy. The

Runge–Kutta methods are designed to give greater accuracy and they possess the advantage of requiring only the function values at some selected points on the subinterval.

If we substitute $y_1 = y_0 + hf(x_0, y_0)$ on the right side of Eq. (8.13), we obtain

$$y_1 = y_0 + \frac{h}{2}[f_0 + f(x_0 + h, y_0 + hf_0)],$$

where $f_0 = f(x_0, y_0)$. If we now set

$$k_1 = hf_0 \quad \text{and} \quad k_2 = hf(x_0 + h,\ y_0 + k_1)$$

then the above equation becomes

$$y_1 = y_0 + \frac{1}{2}(k_1 + k_2), \tag{8.15}$$

which is the *second-order Runge–Kutta* formula. The error in this formula can be shown to be of order h^3 by expanding both sides by Taylor's series. Thus, the left side gives

$$y_0 + hy_0' + \frac{h^2}{2}y_0'' + \frac{h^3}{6}y_0''' + \cdots$$

and on the right side

$$k_2 = hf\ (x_0 + h, y_0 + hf_0) = h\left[f_0 + h\frac{\partial f}{\partial x_0} + hf_0\frac{\partial f}{\partial y_0} + \mathrm{O}(h^2)\right].$$

Since

$$\frac{df(x, y)}{dx} = \frac{\partial f}{\partial x} + f\frac{\partial f}{\partial y},$$

we obtain

$$k_2 = h[f_0 + hf_0' + \mathrm{O}(h^2)] = hf_0 + h^2 f_0' + \mathrm{O}(h^3),$$

so that the right side of Eq. (8.15) gives

$$y_0 + \frac{1}{2}[hf_0 + hf_0 + h^2 f_0' + \mathrm{O}(h^3)] = y_0 + hf_0 + \frac{1}{2}h^2 f_0' + \mathrm{O}(h^3)$$

$$= y_0 + hy_0' + \frac{h^2}{2}y_0'' + \mathrm{O}(h^3).$$

It therefore follows that the Taylor series expansions of both sides of Eq. (8.15) agree up to terms of order h^2, which means that the error in this formula is of order h^3.

More generally, if we set

$$y_1 = y_0 + W_1k_1 + W_2k_2 \tag{8.16a}$$

where

$$\left.\begin{aligned} k_1 &= hf_0 \\ k_2 &= hf\ (x_0 + \alpha_0 h,\ y_0 + \beta_0 k_1) \end{aligned}\right\} \tag{8.16b}$$

then the Taylor series expansions of both sides of the last equation in (8.16a) gives the identity

$$y_0 + hf_0 + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + f_0\frac{\partial f}{\partial y}\right) + \mathrm{O}(h^3) = y_0 + (W_1 + W_2)hf_0$$

$$+ W_2 h^2\left(\alpha_0\frac{\partial f}{\partial x} + \beta_0 f_0\frac{\partial f}{\partial y}\right) + \mathrm{O}(h^3).$$

Equating the coefficients of $f(x, y)$ and its derivatives on both sides, we obtain the relations

$$W_1 + W_2 = 1, \quad W_2\alpha_0 = \frac{1}{2}, \quad W_2\beta_0 = \frac{1}{2}. \tag{8.17}$$

Clearly, $\alpha_0 = \beta_0$ and if α_0 is assigned any value arbitrarily, then the remaining parameters can be determined uniquely. If we set, for example, $\alpha_0 = \beta_0 = 1$, then we immediately obtain $W_1 = W_2 = 1/2$, which gives formula (8.15).

It follows, therefore, that there are several second-order Runge–Kutta formulae and that formulae (8.16) and (8.17) constitute just one of several such formulae.

Higher-order Runge–Kutta formulae exist, of which we mention only the *fourth-order formula* defined by

$$y_1 = y_0 + W_1k_1 + W_2k_2 + W_3k_3 + W_4k_4 \tag{8.18a}$$

where

$$\left.\begin{aligned} k_1 &= hf(x_0, y_0) \\ k_2 &= hf(x_0 + \alpha_0 h, y_0 + \beta_0 k_1) \\ k_3 &= hf(x_0 + \alpha_1 h, y_0 + \beta_1 k_1 + v_1 k_2) \\ k_4 &= hf(x_0 + \alpha_2 h, y_0 + \beta_2 k_1 + v_2 k_2 + \delta_1 k_3), \end{aligned}\right\} \tag{8.18b}$$

where the parameters have to be determined by expanding both sides of the first equation of (8.18a) by Taylor's series and securing agreement of terms up to and including those containing h^4. The choice of the parameters is, again, arbitrary and we have therefore several fourth-order Runge–Kutta formulae. If, for example, we set

$$\left.\begin{aligned} &\alpha_0 = \beta_0 = \frac{1}{2}, && \alpha_1 = \frac{1}{2}, && \alpha_2 = 1, \\ &\beta_1 = \frac{1}{2}(\sqrt{2} - 1), && \beta_2 = 0 && \\ &v_1 = 1 - \frac{1}{\sqrt{2}}, && v_2 = -\frac{1}{\sqrt{2}}, && \delta_1 = 1 + \frac{1}{\sqrt{2}}, \\ &W_1 = W_4 = \frac{1}{6}, && W_2 = \frac{1}{3}\left(1 - \frac{1}{\sqrt{2}}\right), && W_3 = \frac{1}{3}\left(1 + \frac{1}{\sqrt{2}}\right), \end{aligned}\right\} \tag{8.19}$$

we obtain the method of Gill, whereas the choice

$$\left.\begin{aligned} &\alpha_0 = \alpha_1 = \frac{1}{2}, && \beta_0 = \nu_1 = \frac{1}{2} \\ &\beta_1 = \beta_2 = \nu_2 = 0, && \alpha_2 = \delta_1 = 1 \\ &W_1 = W_4 = \frac{1}{6}, && W_2 = W_3 = \frac{2}{6} \end{aligned}\right\} \tag{8.20}$$

leads to the fourth-order Runge–Kutta formula, the most commonly used one in practice:

$$y_1 = y_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{8.21a}$$

where

$$\left.\begin{aligned} k_1 &= hf(x_0, y_0) \\ k_2 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right) \\ k_3 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2\right) \\ k_4 &= hf(x_0 + h, y_0 + k_3) \end{aligned}\right\} \tag{8.21b}$$

in which the error is of order h^5. Complete derivation of the formula is exceedingly complicated, and the interested reader is referred to the book by Levy and Baggot. We illustrate here the use of the fourth-order formula by means of examples.

Example 8.8 Given $dy/dx = y - x$ where $y(0) = 2$, find $y(0.1)$ and $y(0.2)$ correct to four decimal places.

(i) *Runge–Kutta second-order formula:* With $h = 0.1$, we find $k_1 = 0.2$ and $k_2 = 0.21$. Hence

$$y_1 = y(0.1) = 2 + \frac{1}{2}(0.41) = 2.2050.$$

To determine $y_2 = y(0.2)$, we note that $x_0 = 0.1$ and $y_0 = 2.2050$. Hence, $k_1 = 0.1(2.105) = 0.2105$ and $k_2 = 0.1(2.4155 - 0.2) = 0.22155$.

It follows that

$$y_2 = 2.2050 + \frac{1}{2}(0.2105 + 0.22155) = 2.4210.$$

Proceeding in a similar way, we obtain

$$y_3 = y(0.3) = 2.6492 \quad \text{and} \quad y_4 = y(0.4) = 2.8909$$

We next choose $h = 0.2$ and compute $y(0.2)$ and $y(0.4)$ directly. With $h = 0.2$. $x_0 = 0$ and $y_0 = 2$, we obtain $k_1 = 0.4$ and $k_2 = 0.44$ and hence y(0.2) = 2.4200. Similarly, we obtain $y(0.4) = 2.8880$.

From the analytical solution $y = x + 1 + e^x$, the exact values of $y(0.2)$ and $y(0.4)$ are respectively 2.4214 and 2.8918. To study the order of convergence of this method, we tabulate the values as follows:

x	*Computed y*	*Exact y*	*Difference*	*Ratio*
0.2	$h = 0.1$: 2.4210	2.4214	0.0004	3.5
	$h = 0.2$: 2.4200		0.0014	
0.4	$h = 0.1$: 2.8909	2.8918	0.0009	4.2
	$h = 0.2$: 2.8880		0.0038	

It follows that the method has an h^2-order of convergence.

(ii) *Runge–Kutta fourth-order formula:* To determine $y(0.1)$, we have $x_0 = 0$, $y_0 = 2$ and $h = 0.1$. We then obtain

$$k_1 = 0.2,$$
$$k_2 = 0.205$$
$$k_3 = 0.20525$$
$$k_4 = 0.21053.$$

Hence

$$y(0.1) = 2 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 2.2052.$$

Proceeding similarly, we obtain $y(0.2) = 2.4214$.

Example 8.9 Given $dy/dx = 1 + y^2$, where $y = 0$ when $x = 0$, find $y(0.2)$, $y(0.4)$ and $y(0.6)$.

We take $h = 0.2$. With $x_0 = y_0 = 0$, we obtain from (8.21a) and (8.21b),

$$k_1 = 0.2,$$
$$k_2 = 0.2\,(1.01) = 0.202,$$
$$k_3 = 0.2\,(1 + 0.010201) = 0.20204,$$
$$k_4 = 0.2\,(1 + 0.040820) = 0.20816,$$

and

$$y(0.2) = 0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 0.2027,$$

which is correct to four decimal places.

To compute $y(0.4)$, we take $x_0 = 0.2$, $y_0 = 0.2027$ and $h = 0.2$. With these values, Eqs. (8.21a) and (8.21b) give

$$k_1 = 0.2\,[1 + (0.2027)^2] = 0.2082,$$

$$k_2 = 0.2\,[1 + (0.3068)^2] = 0.2188,$$

$$k_3 = 0.2\,[1 + (0.3121)^2] = 0.2195,$$

$$k_4 = 0.2\,[1 + (0.4222)^2] = 0.2356,$$

and

$$y\,(0.4) = 0.2027 + 0.2201 = 0.4228,$$

correct to four decimal places.

Finally, taking $x_0 = 0.4$, $y_0 = 0.4228$ and $h = 0.2$, and proceeding as above, we obtain $y(0.6) = 0.6841$.

Example 8.10 We consider the initial value problem $y' = 3x + y/2$ with the condition $y(0) = 1$.

The following table gives the values of $y(0.2)$ by different methods, the exact value being 1.16722193. It is seen that the *fourth-order* Runge–Kutta method gives the accurate value for $h = 0.05$.

Method	*h*	*Computed value*
Euler	0.2	1.100 000 00
	0.1	1.132 500 00
	0.05	1.149 567 58
Modified Euler	0.2	1.100 000 00
	0.1	1.150 000 00
	0.05	1.162 862 42
Fourth-order Runge–Kutta	0.2	1.167 220 83
	0.1	1.167 221 86
	0.05	1.167 221 93

8.6 PREDICTOR–CORRECTOR METHODS

In the methods described so far, to solve a differential equation over a single interval, say from $x = x_n$ to $x = x_{n+1}$, we required information only at the beginning of the interval, i.e. at $x = x_n$. *Predictor–corrector* methods are the ones which require function values at x_n, x_{n-1}, x_{n-2}, … for the computation of the function value at x_{n+1} A *predictor* formula is used to predict the value of y at x_{n+1} and then a *corrector* formula is used to improve the value of y_{n+1}.

In Section 8.6.1 we derive Predictor–corrector formulae which use backward differences and in Section 8.6.2 we describe Milne's method which uses forward differences.

8.6.1 Adams–Moulton Method

Newton's backward difference interpolation formula can be written as

$$f(x, y) = f_0 + n\nabla f_0 + \frac{n(n+1)}{2}\nabla^2 f_0 + \frac{n(n+1)(n+2)}{6}\nabla^3 f_0 + \cdots \tag{8.22}$$

where

$$n = \frac{x - x_0}{h} \quad \text{and} \quad f_0 = f(x_0, y_0).$$

If this formula is substituted in

$$y_1 = y_0 + \int_{x_0}^{x_1} f(x, y)\,dx, \tag{8.23}$$

we get

$$y_1 = y_0 + \int_{x_0}^{x_1} \left[f_0 + n\nabla f_0 + \frac{n(n+1)}{2}\nabla^2 f_0 + \cdots \right] dx$$

$$= y_0 + h\int_0^1 \left[f_0 + n\nabla f_0 + \frac{n(n+1)}{2}\nabla^2 f_0 + \cdots \right] dn$$

$$= y_0 + h\left(1 + \frac{1}{2}\nabla + \frac{5}{12}\nabla^2 + \frac{3}{8}\nabla^3 + \frac{251}{720}\nabla^4 + \cdots\right) f_0.$$

It can be seen that the right side of the above relation depends only on y_0, y_{-1}, y_{-2}, …, all of which are known. Hence this formula can be used to compute y_1. We therefore write it as

$$y_1^{\mathrm{p}} = y_0 + h\left(1 + \frac{1}{2}\nabla + \frac{5}{12}\nabla^2 + \frac{3}{8}\nabla^3 + \frac{251}{720}\nabla^4 + \cdots\right) f_0 \tag{8.24}$$

This is called *Adams–Bashforth* formula and is used as a *predictor* formula (the superscript p indicating that it is a predicted value).

A corrector formula can be derived in a similar manner by using Newton's backward difference formula at f_1:

$$f(x, y) = f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \frac{n(n+1)(n+2)}{6}\nabla^3 f_1 + \cdots \tag{8.25}$$

Substituting Eq. (8.25) in Eq. (8.23), we obtain

$$y_1 = y_0 + \int_{x_0}^{x_1} \left[f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \cdots \right] dx$$

$$= y_0 + h\int_{1}^{0} \left[f_1 + n\nabla f_1 + \frac{n(n+1)}{2}\nabla^2 f_1 + \cdots \right] dn$$

$$= y_0 + h\left(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \frac{19}{720}\nabla^4 - \cdots\right) f_1 \qquad (8.26)$$

The right side of Eq. (8.26) depends on y_1, y_0, y_{-1}, … where for y_1 we use y_1^{p}, the predicted value obtained from (8.24). The new value of y_1 thus obtained from Eq. (8.26) is called the *corrected* value, and hence we rewrite the formula as

$$y_1^{\mathrm{c}} = y_0 + h\left(1 - \frac{1}{2}\nabla - \frac{1}{12}\nabla^2 - \frac{1}{24}\nabla^3 - \frac{19}{720}\nabla^4 - \cdots\right) f_1^{\mathrm{p}} \qquad (8.27)$$

This is called *Adams–Moulton corrector* formula the superscript c indicates that the value obtained is the corrected value and the superscript p on the right indicates that the predicted value of y_1 should be used for computing the value of $f(x_1, y_1)$.

In practice, however, it will be convenient to use formulae (8.24) and (8.27) by ignoring the higher-order differences and expressing the lower-order differences in terms of function values. Thus, by neglecting the fourth and higher-order differences, formulae (8.24) and (8.27) can be written as

$$y_1^{\mathrm{p}} = y_0 + \frac{h}{24}(55 f_0 - 59 f_{-1} + 37 f_{-2} - 9 f_{-3}) \qquad (8.28)$$

and

$$y_1^{\mathrm{c}} = y_0 + \frac{h}{24}(9 f_1^{\mathrm{p}} + 19 f_0 - 5 f_{-1} + f_{-2}) \qquad (8.29)$$

in which the errors are approximately

$$\frac{251}{720}h^5 f_0^{(4)} \quad \text{and} \quad -\frac{19}{720}h^5 f_0^{(4)} \quad \text{respectively.}$$

The general forms of formulae (8.28) and (8.29) are given by

$$y_{n+1}^{\mathrm{p}} = y_n + \frac{h}{24}[55 f_n - 59 f_{n-1} + 37 f_{n-2} - 9 f_{n-3}] \qquad (8.28a)$$

and

$$y_{n+1}^{\mathrm{c}} = y_n + \frac{h}{24}[9 f_{n+1}^{\mathrm{p}} + 19 f_n - 5 f_{n-1} + f_{n-2}] \qquad (8.29a)$$

Such formulae, expressed in ordinate form, are often called *explicit predictor–corrector* formulae.

The values y_{-1}, y_{-2} and y_{-3}, which are required on the right side of Eq. (8.28) are obtained by means of the Taylor's series, or Euler's method, or Runge–Kutta method. Due to this reason, these methods are called *starter* methods. For practical problems, Runge–Kutta fourth-order formula together with formulae (8.28) and (8.29) have been found to be the most successful combination. The following example will illustrate the application of this method.

Example 8.11 We consider once again the differential equation given in Example 8.9 with the same condition, and we wish to compute $y(0.8)$.

For this example, the starter values are $y(0.6)$, $y(0.4)$ and $y(0.2)$, which are already computed in Example 8.9 by the fourth-order Runge–Kutta method. Using now Eq. (8.28) with $y_0 = 0.6841$, $y_{-1} = 0.4228$, $y_{-2} = 0.2027$ and $y_{-3} = 0$, we obtain

$$\begin{aligned} y^{\mathrm{p}}(0.8) &= 0.6841 + \frac{0.2}{24}\{55[1+(0.6841)^2] - 59[1+(0.4228)^2] \\ &\quad + 37[1+(0.2027)^2] - 9\} \\ &= 1.0233, \text{ on simplification.} \end{aligned}$$

Using this predicted value on the right side of Eq. (8.29), we obtain

$$\begin{aligned} y^{\mathrm{c}}(0.8) &= 0.6841 + \frac{0.2}{24}\{9[1+(0.0233)^2] + 19[1+(0.6841)^2] \\ &\quad - 5[1+(0.4228)^2] + [1+(0.2027)^2]\} \\ &= 1.0296, \text{ which is correct to four decimal places} \end{aligned}$$

The importance of the method lies in the fact that when once y_1^{p} is computed from formula (8.28), formula (8.29) can be used iteratively to obtain the value of y_1 to the accuracy required.

8.6.2 Milne's Method

This method uses Newton's forward difference formula in the form

$$f(x, y) = f_0 + n\Delta f_0 + \frac{n(n-1)}{2}\Delta^2 f_0 + \frac{n(n-1)(n-2)}{6}\Delta^3 f_0 + \cdots \tag{8.30}$$

Substituting Eq. (8.30) in the relation

$$y_4 = y_0 + \int_{x_0}^{x_4} f(x, y)\, dx \tag{8.31}$$

we obtain

$$y_4 = y_0 + \int_{x_0}^{x_4} \left[f_0 + n\Delta f_0 + \frac{n(n-1)}{2}\Delta^2 f_0 + \cdots \right] dx$$

$$= y_0 + h\int_0^4 \left[f_0 + n\Delta f_0 + \frac{n(n-1)}{2}\Delta^2 f_0 + \cdots \right] dn$$

$$= y_0 + h\left(4f_0 + 8\Delta f_0 + \frac{20}{3}\Delta^2 f_0 + \frac{8}{3}\Delta^3 f_0 + \cdots \right)$$

$$= y_0 + \frac{4h}{3}(2f_1 - f_2 + 2f_3) \tag{8.32}$$

after neglecting fourth- and higher-order differences and expressing differences $\Delta f_0, \Delta^2 f_0$ and $\Delta^3 f_0$ in terms of the function values.

This formula can be used to 'predict' the value of y_4 when those of y_0, y_1, y_2 and y_3 are known. To obtain a 'corrector' formula, we substitute Newton's formula from (8.30) in the relation

$$y_2 = y_0 + \int_{x_0}^{x_2} f(x, y)\, dx \tag{8.33}$$

and get

$$y_2 = y_0 + h\int_0^2 \left[f_0 + n\Delta f_0 + \frac{n(n-1)}{2}\Delta^2 f_0 + \cdots \right] dn$$

$$= y_0 + h\left(2f_0 + 2\Delta f_0 + \frac{1}{3}\Delta^2 f_0 + \cdots \right)$$

$$= y_0 + \frac{h}{3}(f_0 + 4f_1 + f_2) \tag{8.34}$$

The value of y_4 obtained from Eq. (8.32) can therefore be checked by using Eq. (8.34).

The general form of Eqs. (8.32) and (8.34) are:

$$y_{n+1}^{\mathrm{p}} = y_{n-3} + \frac{4h}{3}(2f_{n-2} - f_{n-1} + 2f_n) \tag{8.32a}$$

and

$$y_{n+1}^{\mathrm{c}} = y_{n-1} + \frac{h}{3}(f_{n-1} + 4f_n + f_{n+1}) \tag{8.34a}$$

The application of this method is illustrated by the following example.

Example 8.12 We consider again the differential equation discussed in Examples 8.9 and 8.10, viz., to solve $y' = 1 + y^2$ with $y(0) = 0$ and we wish to compute $y(0.8)$ and $y(1.0)$.

With $h = 0.2$, the values of $y(0.2)$, $y(0.4)$ and $y(0.6)$ are computed in Example 8.9 and these values are given in the table below:

x	y	$y' = 1 + y^2$
0	0	1.0
0.2	0.2027	1.0411
0.4	0.4228	1.1787
0.6	0.6841	1.4681

To obtain $y(0.8)$, we use Eq. (8.32) and obtain

$$y(0.8) = 0 + \frac{0.8}{3}[2(1.0411) - 1.1787 + 2(1.4681)] = 1.0239$$

This gives

$$y'(0.8) = 2.0480.$$

To correct this value of $y(0.8)$, we use formula (8.34) and obtain

$$y(0.8) = 0.4228 + \frac{0.2}{3}[1.1787 + 4(1.4681) + 2.0480] = 1.0294.$$

Proceeding similarly, we obtain $y(1.0) = 1.5549$. The accuracy in the values of $y(0.8)$ and $y(1.0)$ can, of course, be improved by repeatedly using formula (8.34).

Example 8.13 The differential equation $y' = x^2 + y^2 - 2$ satisfies the following data:

x	y
–0.1	1.0900
0	1.0000
0.1	0.8900
0.2	0.7605

Use Milne's method to obtain the value of $y(0.3)$.

We first form the following table:

x	y	$y' = x^2 + y^2 - 2$
–0.1	1.0900	–0.80190
0	1.0	–1.0
0.1	0.8900	–1.19790
0.2	0.7605	–1.38164

Using Eq. (8.32), we obtain

$$y(0.3) = 1.09 + \frac{4(0.1)}{3}[2(-1) - (-1.19790) + 2(-1.38164)] = 0.614616.$$

In order to apply Eq. (8.34), we need to compute $y'(0.3)$. We have

$$y'(0.3) = (0.3)^2 + (0.614616)^2 - 2 = -1.532247.$$

Now, Eq. (8.34) gives the corrected value of $y(0.3)$:

$$y(0.3) = 0.89 + \frac{0.1}{3}[-1.197900 + 4(-1.38164) + (-1.532247)] = 0.614776.$$

8.7 CUBIC SPLINE METHOD

The governing equations of a cubic spline have been discussed in detail in Section 5.2, where the cubic spline function has been obtained in terms of its second derivatives, M_i. In certain applications, e.g. the solution of initial-value problems, it would be convenient to use the governing equations in terms of its first derivatives, i.e., m_i. Using Hermite's interpolation formula (see Section 3.9.3), it would not be difficult to derive the following formula for the cubic spline $s(x)$ in $x_{i-1} \le x \le x_i$ in terms of its first derivatives $s'(x_i) = m_i$:

$$s(x) = m_{i-1}\frac{(x_i - x)^2(x - x_{i-1})}{h^2} - m_i\frac{(x - x_{i-1})^2(x_i - x)}{h^2}$$
$$+ y_{i-1}\frac{(x_i - x)^2[2(x - x_{i-1}) + h]}{h^3} + y_i\frac{(x - x_{i-1})^2[2(x_i - x) + h]}{h^3}, \quad (8.35)$$

where $h = x_i - x_{i-1}$. Differentiating Eq. (8.35) with respect to x and simplifying, we obtain

$$s'(x) = \frac{m_{i-1}}{h^2}(x_i - x)(2x_{i-1} + x_i - 3x) - \frac{m_i}{h^2}(x - x_{i-1})(x_{i-1} + 2x_i - 3x)$$
$$+ \frac{6(y_i - y_{i-1})}{h^3}(x - x_{i-1})(x_i - x). \quad (8.36)$$

Again,

$$s''(x) = -\frac{2m_{i-1}}{h^2}(x_{i-1} + 2x_i - 3x) - \frac{2m_i}{h^2}(2x_{i-1} + x_i - 3x)$$
$$+ \frac{6(y_i - y_{i-1})}{h^3}(x_{i-1} + x_i - 2x), \quad (8.37)$$

which gives

$$s''(x_i) = \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(y_i - y_{i-1})$$

$$= \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}). \tag{8.38}$$

If we now consider the initial-value problem

$$\frac{dy}{dx} = f(x, y) \tag{8.39a}$$

and

$$y(x_0) = y_0 \tag{8.39b}$$

then from Eq. (8.39a), we obtain

$$\frac{d^2y}{dx^2} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx},$$

or

$$y''(x_i) = f_x(x_i, y_i) + f_y(x_i, y_i)\, f(x_i, y_i)$$

$$= f_x(x_i, s_i) + f_y(x_i, s_i)\, f(x_i, s_i). \tag{8.40}$$

Equating Eqs. (8.38) and (8.40), we obtain

$$\frac{2m_{i-1}}{h} + \frac{4m_i}{h} - \frac{6}{h^2}(s_i - s_{i-1}) = f_x(x_i, s_i) + f_y(x_i, s_i) f(x_i, s_i) \tag{8.41}$$

from which s_i can be computed. Substitution in Eq. (8.35) gives the required solution.

The following example demonstrates the usefulness of the spline method.

Example 8.14 We consider again the initial-value problem defined by

$$y' = 3x + \frac{1}{2}y, \qquad y(0) = 1, \tag{i}$$

whose exact solution is given by

$$y = 13e^{x/2} - 6x - 12 \tag{ii}$$

We take, for simplicity, $n = 2$, i.e. $h = 0.5$ and compute the value of $y(0.5)$. Here $f(x, y) = 3x + y/2$ and therefore we have $f_x = 3$ and $f_y = 1/2$. Also,

$$f(x_i, s_i) = 3x_i + \frac{1}{2}s_i.$$

Hence, Eq. (8.41) gives

$$4m_0 + 8m_1 - 24(s_1 - s_0) = 3 + \frac{1}{2}\left(\frac{3}{2} + \frac{1}{2}s_1\right)$$

and

$$4m_1 + 8m_2 - 24(s_2 - s_1) = 3 + \frac{1}{2}\left(3 + \frac{1}{2}s_2\right)$$

Since $m_0 = 1/2, m_1 = 3/2 + s_1/2$ and $m_2 = 3 + s_2/2$, the above equations give on simplification

$$s_1 = 1.691358 \quad \text{and} \quad s_2 = 3.430879.$$

The errors in these solutions are given by 0.000972 and 0.002497, respectively. It can be shown that, under certain conditions, the spline method gives $O(h^4)$ convergence and compares well with the multi-step Milne's method. For details, the reader is referred to Patricio [1978].

8.8 SIMULTANEOUS AND HIGHER-ORDER EQUATIONS

We consider the two equations

$$\frac{dx}{dt} = f(t, x, y) \quad \text{and} \quad \frac{dy}{dt} = \phi(t, x, y) \tag{8.42}$$

with the initial conditions $x = x_0$ and $y = y_0$, when $t = t_0$. Assuming that $\Delta t = h$, $\Delta x = k$, and $\Delta y = l$, the fourth-order Runge–Kutta method gives

$$\left.\begin{aligned}
k_1 &= hf(t_0, x_0, y_0);\\
l_1 &= h\phi(t_0, x_0, y_0);\\
k_2 &= hf\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_1, y_0 + \frac{1}{2}l_1\right);\\
l_2 &= h\phi\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_1, y_0 + \frac{1}{2}l_1\right);\\
k_3 &= hf\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_2, y_0 + \frac{1}{2}l_2\right);\\
l_3 &= h\phi\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_2, y_0 + \frac{1}{2}l_2\right);\\
k_4 &= hf(t_0 + h, x_0 + k_3, y_0 + l_3);\\
l_4 &= h\phi(t_0 + h, x_0 + k_3, y_0 + l_3);\\
x_1 &= x_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)\\
y_1 &= y_0 + \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4).
\end{aligned}\right\} \tag{8.43}$$

In a similar manner, one can extend the Taylor series method or Picard's method to the system (8.42). The extension of the Runge–Kutta method to a system of n equations is quite straightforward.

We now consider the second-order differential equation

$$y'' = F(x, y, y') \tag{8.44a}$$

with the initial conditions

$$y(x_0) = y_0 \quad \text{and} \quad y'(x_0) = y'_0. \tag{8.45a}$$

By setting $z = y'$, the problem given in Eqs. (8.44a) and (8.45a) can be reduced to the problem of solving the system

$$y' = z \quad \text{and} \quad z' = F(x, y, z) \tag{8.44b}$$

with the conditions

$$y(x_0) = y_0 \quad \text{and} \quad z(x_0) = y'_0 \tag{8.45b}$$

which can be solved by the method described above. Similarly, any higher-order differential equation, in which we can solve for the highest derivative, can be reduced to a system of first-order differential equations.

8.9 SOME GENERAL REMARKS

In the preceding sections, we have given a brief discussion of some well-known methods for the numerical solution of an ordinary differential equation satisfying certain given initial conditions. If the solution is required over a wider range, it is important to get the starting values as accurately as possible by one of the methods described.

It is outside the scope of this book to present a comprehensive review of the different methods described in this text for the numerical solution of differential equations, but the following points are relevant to the methods discussed.

The Taylor's series method suffers from the serious disadvantage that all the higher derivatives of $f(x, y)$ [see Eqs. (8.1)] must exist and that h should be small such that successive terms of the series diminish quite rapidly. Likewise, in the modified Euler method, the value of h should be so small that one or two applications of the iteration formula (8.14) will give the final result for that value of h. The Picard method has probably little practical value because of the difficulty in performing the successive integrations.

Although laborious, the Runge–Kutta method is the most widely used one since it gives reliable starting values and is particularly suitable when the computation of higher derivatives is complicated. When the starting values have been found, the computations for the rest of the interval can be continued by means of the predictor–corrector methods.

The cubic spline method is a one-step method and at the same time a global one. The step-size can be changed during computations and, under certain conditions, gives $O(h^4)$ convergence. The method can also be extended to systems of ordinary differential equations.

8.10 BOUNDARY-VALUE PROBLEMS

Some simple examples of two-point linear boundary-value problems are:

(a) $$y''(x) + f(x)y'(x) + g(x)y(x) = r(x) \tag{8.46}$$

with the boundary conditions

$$y(x_0) = a \quad \text{and} \quad y(x_n) = b \tag{8.47}$$

(b) $$y^{\text{iv}}(x) = p(x)\, y(x) + q(x) \tag{8.48}$$

with

$$y(x_0) = y'(x_0) = A \quad \text{and} \quad y(x_n) = y'(x_n) = B. \tag{8.49}$$

Problems of the type (b), which involve the fourth-order differential equation, are much involved and will not be discussed here. There exist many methods of solving second-order boundary-value problems of type (a). Of these, the finite difference method is a popular one and will be described in Section 8.10.1. Finally, in Sections 8.10.2 and 8.10.3 we discuss methods based on the application of cubic splines and weighted residuals.

8.10.1 Finite-difference Method

The finite-difference method for the solution of a two-point boundary value problem consists in replacing the derivatives occurring in the differential equation (and in the boundary conditions as well) by means of their finite-difference approximations and then solving the resulting linear system of equations by a standard procedure.

To obtain the appropriate finite-difference approximations to the derivatives, we proceed as follows.

Expanding $y(x+h)$ in Taylor's series, we have

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x) + \cdots \tag{8.50}$$

from which we obtain

$$y'(x) = \frac{y(x+h) - y(x)}{h} - \frac{h}{2}y''(x) - \cdots$$

Thus we have

$$y'(x) = \frac{y(x+h) - y(x)}{h} + O(h) \tag{8.51}$$

which is the forward difference approximation for $y'(x)$. Similarly, expansion of $y(x-h)$ in Taylor's series gives

$$y(x-h) = y(x) - hy'(x) + \frac{h^2}{2}y''(x) - \frac{h^3}{6}y'''(x) + \cdots \tag{8.52}$$

from which we obtain

$$y'(x) = \frac{y(x) - y(x-h)}{h} + \mathrm{O}(h) \tag{8.53}$$

which is the backward difference approximation for $y'(x)$.

A central difference approximation for $y'(x)$ can be obtained by subtracting Eq. (8.52) from Eq. (8.50). We thus have

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + \mathrm{O}(h^2). \tag{8.54}$$

It is clear that Eq. (8.54) is a better approximation to $y'(x)$ than either Eq. (8.51) or Eq. (8.53). Again, adding Eqs. (8.50) and (8.52), we get an approximation for $y''(x)$

$$y''(x) = \frac{y(x-h) - 2y(x) + y(x+h)}{h^2} + \mathrm{O}(h^2). \tag{8.55}$$

In a similar manner, it is possible to derive finite-difference approximations to higher derivatives.

To solve the boundary-value problem defined by Eqs. (8.46) and (8.47), we divide the range $[x_0, x_n]$ into n equal subintervals of width h so that

$$x_i = x_0 + ih, \quad i = 1, 2, \ldots, n.$$

The corresponding values of y at these points are denoted by

$$y(x_i) = y_i = y(x_0 + ih), \quad i = 0, 1, 2, \ldots, n.$$

From Eqs. (8.54) and (8.55), values of $y'(x)$ and $y''(x)$ at the point $x = x_i$ can now be written as

$$y_i' = \frac{y_{i+1} - y_{i-1}}{2h} + \mathrm{O}(h^2)$$

and

$$y_i'' = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + \mathrm{O}(h^2).$$

Satisfying the differential equation at the point $x = x_i$, we get

$$y_i'' + f_i y_i' + g_i y_i = r_i$$

Substituting the expressions for y_i' and y_i'', this gives

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + f_i \frac{y_{i+1} - y_{i-1}}{2h} + g_i y_i = r_i, \qquad i = 1, 2, \ldots, n-1,$$

where $y_i = y(x_i)$, $g_i = g(x_i)$, etc.

Multiplying through by h^2 and simplifying, we obtain

$$\left(1-\frac{h}{2}f_i\right)y_{i-1}+(-2+g_ih^2)y_i+\left(1+\frac{h}{2}f_i\right)y_{i+1}=r_ih^2, \qquad (8.56)$$

$$i=1,\ 2,\ \ldots,n-1$$

with

$$y_0 = a \quad \text{and} \quad y_n = b \qquad (8.57)$$

Equation (8.56) with the conditions (8.57) comprise a tridiagonal system which can be solved by the method outlined in Section 7.5.9 of Chapter 7. The solution of this tridiagonal system constitutes an approximate solution of the boundary value problem defined by Eqs. (8.46) and (8.47).

To estimate the error in the numerical solution, we define the *local truncation error*, τ, by

$$\tau=\left(\frac{y_{i-1}-2y_i+y_{i+1}}{h^2}-y_i''\right)+f_i\left(\frac{y_{i+1}-y_{i-1}}{2h}-y_i'\right).$$

Expanding y_{i-1} and y_{i+1} by Taylor's series and simplifying, the above gives

$$\tau=\frac{h^2}{12}(y_i^{\text{iv}}+2f_iy_i''')+\mathrm{O}(h^4). \qquad (8.58)$$

Thus, the finite difference approximation defined by Eq. (8.56) has second-order accuracy for functions with continuous fourth derivatives on $[x_0, x_n]$. Further, it follows that $\tau\to 0$ as $h\to 0$, implying that greater accuracy in the result can be achieved by using a smaller value of h. In such a case, of course, more computational effort would be required since the number of equations become larger.

An easier way to improve accuracy is to employ *Richardson's deferred approach to the limit*, assuming that the $\mathrm{O}(h^2)$ error is proportional to h^2. This means that the error has the form

$$y(x_i)-y_i=h^2e(x_i)+\mathrm{O}(h^4) \qquad (8.59)$$

For extrapolation to the limit, we solve Eq. (8.56) twice, with the interval lengths h and $h/2$ respectively. Let the corresponding solutions of Eq. (8.56) be denoted by $y_i(h)$ and $y_i(h/2)$. For a point x_i common to both, we therefore have

$$y(x_i)-y_i(h)=h^2e(x_i)+\mathrm{O}(h^4) \qquad (8.60a)$$

and

$$y(x_i)-y_i\left(\frac{h}{2}\right)=\frac{h^2}{4}e(x_i)+\mathrm{O}(h^4) \qquad (8.60b)$$

from which we obtain

$$y(x_i)=\frac{4y_i(h/2)-y_i(h)}{3}. \qquad (8.61)$$

We have explained the method with simple boundary conditions (8.47) where the function values on the boundary are prescribed. In many applied problems, however, derivative boundary conditions may be prescribed, and this requires a modification of the procedures described above. The following examples illustrate the application of the finite-difference method.

Example 8.15 A boundary-value problem is defined by

$$y'' + y + 1 = 0, \qquad 0 \le x \le 1$$

where

$$y(0) = 0 \quad \text{and} \quad y(1) = 0.$$

With $h = 0.5$, use the finite-difference method to determine the value of $y(0.5)$.

This example was considered by Bickley [1968]. Its exact solution is given by

$$y(x) = \cos x + \frac{1 - \cos 1}{\sin 1} \sin x - 1,$$

from which, we obtain

$$y(0.5) = 0.139493927.$$

Here $nh = 1$. The differential equation is approximated as

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + y_i + 1 = 0$$

and this gives after simplification

$$y_{i-1} - (2 - h^2) y_i + y_{i+1} = -h^2, \qquad i = 1, 2, \ldots, n-1$$

which together with the boundary conditions $y_0 = 0$ and $y_n = 0$, comprises a system of $(n+1)$ equations for the $(n+1)$ unknowns $y_0, y_1, \ldots, y_n$.

Choosing $h = 1/2$ (i.e. $n = 2$), the above system becomes

$$y_0 - \left(2 - \frac{1}{4}\right) y_1 + y_2 = -\frac{1}{4}.$$

With $y_0 = y_2 = 0$, this gives

$$y_1 = y(0.5) = \frac{1}{7} = 0.142857142\ldots$$

Comparison with the exact solution given above shows that the error in the computed solution is 0.00336.

On the other hand, if we choose $h = 1/4$ (i.e. $n = 4$), we obtain the three equations:

$$y_0 - \frac{31}{16} y_1 + y_2 = -\frac{1}{16}$$

$$y_1 - \frac{31}{16} y_2 + y_3 = -\frac{1}{16}$$

$$y_2 - \frac{31}{16} y_3 + y_4 = -\frac{1}{16},$$

where $y_0 = y_4 = 0$. Solving the system we obtain

$$y_2 = y(0.5) = \frac{63}{449} = 0.140311804,$$

the error in which is 0.00082. Since the ratio of the two errors is about 4, it follows that the order of convergence is h^2.

These results show that the accuracy obtained by the finite-difference method depends upon the width of the subinterval chosen and also on the order of the approximations. As h is reduced, the accuracy increases but the number of equations to be solved also increases.

Example 8.16 Solve the boundary-value problem

$$\frac{d^2y}{dx^2} - y = 0$$

with

$$y(0) = 0 \quad \text{and} \quad y(2) = 3.62686.$$

The exact solution of this problem is $y = \sinh x$. The finite-difference approximation is given by

$$\frac{1}{h^2}(y_{i-1} - 2y_i + y_{i+1}) = y_i. \tag{i}$$

We subdivide the interval [0, 2] into four equal parts so that $h = 0.5$. Let the values of y at the five points be y_0, y_1, y_2, y_3 and y_4. We are given that

$$y_0 = 0 \quad \text{and} \quad y_4 = 3.62686.$$

Writing the difference equations at the three interval points (which are the unknowns), we obtain

$$\left.\begin{aligned} 4(y_0 - 2y_1 + y_2) &= y_1 \\ 4(y_1 - 2y_2 + y_3) &= y_2 \\ 4(y_2 - 2y_3 + y_4) &= y_3, \end{aligned}\right\} \tag{ii}$$

respectively. Substituting for y_0 and y_4 and rearranging, we get the system

$$\left.\begin{aligned} -9y_1 + 4y_2 &= 0 \\ 4y_1 - 9y_2 + 4y_3 &= 0 \\ 4y_2 - 9y_3 &= -14.50744. \end{aligned}\right\} \tag{iii}$$

The solution of (iii) is given in the table below.

x	*Computed value* of y	*Exact value* $y = \sinh x$	Error
0.5	0.52635	0.52110	0.00525
1.0	1.18428	1.17520	0.00908
1.5	2.13829	2.12928	0.00901

It is possible to obtain a better approximation for the value of $y(1.0)$ by extrapolation to the limit. For this we divide the interval [0, 2] into two subintervals with $h = 1.0$. The difference equation at the single unknown point y_1 is given by

$$y_0 - 2y_1 + y_2 = y_1$$

Using the values of y_0 and y_2, we obtain

$$y_1 = 1.20895.$$

Hence Eq. (8.61) gives

$$y(1.0) = \frac{4(1.18428) - 1.20895}{3} = 1.17606,$$

which is a better approximation since the error is now reduced to 0.00086.

8.10.2 Cubic Spline Method

We consider again the boundary-value problem defined by Eqs. (8.46) and (8.47). Let $s(x)$ be the cubic spline approximating the function $y(x)$ and let $s''(x_i) = M_i$. Then, at $x = x_i$ the differential equation given in Eq. (8.46) gives

$$M_i + f_i s'(x_i) + g_i y_i = r_i \tag{8.62}$$

But

$$s'(x_i -) = \frac{h}{3!}(2M_i + M_{i-1}) + \frac{1}{h}(y_i - y_{i-1}) \tag{8.63}$$

and

$$s'(x_i +) = -\frac{h}{3!}(2M_i + M_{i+1}) + \frac{1}{h}(y_{i+1} - y_i) \tag{8.64}$$

Substituting Eqs. (8.63) and (8.64) successively in Eq. (8.62), we obtain the equations

$$M_i + f_i\left[\frac{h}{6}(2M_i + M_{i-1}) + \frac{1}{h}(y_i - y_{i-1})\right] + g_i y_i = r_i \tag{8.65}$$

and

$$M_i + f_i\left[-\frac{h}{6}(2M_i + M_{i+1}) + \frac{1}{h}(y_{i+1} - y_i)\right] + g_i y_i = r_i. \tag{8.66}$$

Since y_0 and y_n are known, Eqs. (8.65) and (8.66) constitute a system of $2n$ equations in $2n$ unknowns, viz., $M_0, M_1, \ldots, M_n$, $y_1, y_2, \ldots, y_{n-1}$. It is, however, possible to eliminate the M_i and obtain a tridiagonal system for y_i (see, Albasiny and Hoskins [1969]). The following examples illustrate the use of the spline method.

Example 8.17 We first consider the problem discussed in Example 8.15, viz.,

$$y'' + y + 1 = 0, \quad y(0) = y(1) = 0 \tag{i}$$

If we divide the interval [0, 1] into two equal subintervals, then from Eq. (i) and the recurrence relations for M_i, we obtain

$$y(0.5) = \frac{3}{22} = 0.13636, \tag{ii}$$

and

$$M_0 = -1, \quad M_1 = -\frac{25}{22}, \quad M_2 = -1$$

Hence we obtain

$$s'(0) = \frac{47}{88}, \quad s'(1) = -\frac{47}{88}, \quad s'(0.5) = 0.$$

From the analytical solution of the problem (i), we observe that $y(0.5) = 0.13949$ and hence the cubic spline solution of the boundary-value problem has an error of 2.24% (see Bickley [1968]).

Example 8.18 Given the boundary-value problem

$$x^2 y'' + xy' - y = 0; \qquad y(1) = 1, \quad y(2) = 0.5$$

apply the cubic spline method to determine the value of $y(1.5)$.

The given differential equation is

$$y'' = -\frac{1}{x} y' + \frac{1}{x^2} y. \tag{i}$$

Setting $x = x_i$ and $y''(x_i) = M_i$, Eq. (i) becomes

$$M_i = -\frac{1}{x_i} y_i' + \frac{1}{x_i^2} y_i. \tag{ii}$$

Using the expressions given in Eqs. (8.63) and (8.64), we obtain

$$M_i = -\frac{1}{x_i}\left(-\frac{h}{3} M_i - \frac{h}{6} M_{i+1} + \frac{y_{i+1} - y_i}{h}\right) + \frac{1}{x_i^2} y_i, \qquad i = 0, 1, 2, \ldots, n-1. \tag{iii}$$

and

$$M_i = -\frac{1}{x_i}\left(\frac{h}{3} M_i + \frac{h}{6} M_{i-1} + \frac{y_i - y_{i-1}}{h}\right) + \frac{1}{x_i^2} y_i, \qquad i = 1, 2, \ldots, n. \tag{iv}$$

If we divide [1, 2] into two subintervals, we have $h = 1/2$ and $n = 2$. Then Eqs. (iii) and (iv) give

$$10M_0 - M_1 + 24y_1 = 36$$

$$16M_1 - M_2 - 32y_1 = -12$$

$$M_0 + 20M_1 + 16y_1 = 24$$

$$M_1 + 26M_2 - 24y_1 = -9$$

Eliminating M_0, M_1 and M_2 from these system of equation we obtain

$$y_1 = 0.65599.$$

Since the exact value of $y_1 = y(1.5) = 2/3$, the error in the computed value of y_1 is 0.01, which is about 1.5% smaller.

Example 8.19 Consider a boundary-value problem in which the boundary conditions involve derivatives

$$\frac{d^2y}{dx^2} = y \tag{i}$$

with

$$y'(0) = 0 \quad \text{and} \quad y(1) = 1 \tag{ii}$$

The analytical solution of this problem is given by

$$y = \frac{\cosh x}{\cosh 1} \tag{iii}$$

In order to compare the finite-difference and spline methods, we solve this problem by both the methods. For the finite-difference solution, we write

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = y_i \tag{iv}$$

We divide the interval [0, 1] into two equal parts such that $h = 1/2$. Setting $i = 0$ and $i = 1$, Eq. (iv) gives

$$y_{-1} - 2y_0 + y_1 = \frac{1}{4}y_0 \tag{v}$$

and

$$y_0 - 2y_1 + y_2 = \frac{1}{4}y_1 \tag{vi}$$

From formula (8.54), we have

$$y_0' = \frac{y_1 - y_{-1}}{2h} \quad \text{or} \quad y_1 - y_{-1} = 2hy_0' \tag{vii}$$

Using the boundary conditions $y_0' = 0$ and $y_2 = 1$, Eqs. (v), (vi) and (vii) yield

$$y_1 = \frac{36}{49} = 0.9376.$$

The exact value of $y(0.5)$ is 0.7310 so that the finite-difference solution has an error of 0.2066.

For the spline solution, we have

$$y_{i-1} + 4y_i + y_{i+1} = \frac{6}{h^2}(y_{i-1} - 2y_i + y_{i+1}) \tag{viii}$$

With $h = 1/2$, we obtain

$$y_0 + 4y_1 + y_2 = 24\,(y_0 - 2y_1 + y_2)$$

Since $y_2 = 1$, the above equation becomes

$$y_0 + 4y_1 = 24\,(y_0 - 2y_1) + 23$$

or, equivalently

$$52y_1 = 23y_0 + 23 \tag{ix}$$

For the derivative boundary condition, we use Eq. (8.64) and obtain

$$y_0' = 0 = -\frac{1}{6}M_0 - \frac{1}{12}M_1 + 2\,(y_1 - y_0)$$

Since $M_0 = y_0$ and $M_1 = y_1$, the above equation gives

$$2y_0 + y_1 = 24\,(y_1 - y_0) \tag{x}$$

Equations (ix) and (x) yield

$$y_1 = y\,(0.5) = \frac{598}{823} = 0.7266.$$

Thus the error in the cubic spline solution is 0.0044. This example demonstrates the superiority of the cubic spline method over the finite difference method when the boundary value problem contains derivative boundary conditions.

8.10.3 Galerkin's Method

This method, also called the *weighted residual* method, uses *trial functions* (or approximating functions) which satisfy the boundary conditions of the problem. The trial function is substituted in the given differential equation and the result is called the *residual*. The integral of the product of this residual and a weighted function, taken over the domain, is then set to zero which yields a system of equations for the unknown parameters in the trial functions.

Let the boundary value problem be defined by

$$y'' + p(x)y' + q(x)y = f(x) \qquad a < x < b \tag{8.67}$$

with the boundary conditions

$$\left.\begin{aligned} p_0 y(a) + q_0 y'(a) &= r_0 \\ p_1 y(b) + q_1 y'(b) &= r_1 \end{aligned}\right\} \tag{8.68}$$

Let the approximate solution be given by

$$t(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{8.69}$$

where $\phi_i(x)$ are called *base functions.* Substituting for $t(x)$ in Eq. (8.67), we obtain a residual. Denoting this residual by $R(t)$,

we obtain

$$R(t) = t'' + p(x)t' + q(x)t - f(x) \tag{8.70}$$

Usually the base functions $\phi_i(x)$ are chosen as weight functions. We, therefore, have

$$I = \int_a^b \phi_i(x)\, R(t)\, dx = 0, \tag{8.71}$$

which yields a system of equations for the parameters α_i. When α_i are known, $t(x)$ can be calculated from Eq. (8.69).

Example 8.20 Solve the boundary value problem defined by

$$y'' + y + x = 0, \qquad 0 < x < 1$$

with the conditions

$$y(0) = y(1) = 0.$$

Let

$$t(x) = \alpha_1 \phi_1(x)$$

Since both the boundary conditions must be satisfied by $t(x)$, we choose

$$\phi_1(x) = x(1 - x).$$

Substituting for $t(x)$ in the given differential equation, we obtain

$$R(t) = t'' + t + x.$$

Hence we have

$$I = \int_0^1 (t'' + t + x)\, \alpha_1 x\, (1 - x)\, dx = 0$$

$$\Rightarrow \qquad \int_0^1 (t'' + t + x)\, x(1 - x)\, dx = 0 \tag{i}$$

Now,

$$\int_0^1 t'' x(1 - x)\, dx = [t' x(1 - x)]_0^1 - \int_0^1 t'(1 - 2x)\, dx,$$

on integrating by parts.

$$= -\int_0^1 t'(1 - 2x)\, dx,$$ since the first term vanishes.

$$= -\left[\{t(1 - 2x)\}_0^1 - \int_0^1 t(-2)\, dx \right]$$

$$= -2\int_0^1 t\, dx,$$ since $t = 0$ at $x = 0$ and $x = 1$.

Hence (i) simplifies to

$$-2\int_0^1 t\,dx + \int_0^1 tx(1-x)dx + \int_0^1 x^2(1-x)dx = 0$$

$$\Rightarrow \qquad -2\int_0^1 \alpha_1 x(1-x)dx + \int_0^1 \alpha_1 x^2(1-x)^2 dx + \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = 0$$

$$\Rightarrow \qquad \alpha_1 = \frac{5}{18} = 0.2778, \text{ an simplification.}$$

Then a first approximation to the solution is

$$y(0.5) = \frac{5}{18}(0.5)(0.5) = 0.06944.$$

The exact solution to the given boundary value problem is

$$y(x) = \frac{\sin x}{\sin 1} - x,$$

which means that our solution has an error of 0.0003.

The above approximation can be improved by assuming that

$$t(x) = \alpha_1 x(1-x) + \alpha_2 x^2(1-x).$$

Proceeding as above, we obtain

$$\alpha_1 = 0.1924 \text{ and } \alpha_2 = 0.1707.$$

It is clear that by adding more terms to $t(x)$, we can obtain the result to the desired accuracy.

EXERCISES

8.1. Given

$$\frac{dy}{dx} = 1 + xy, \quad y(0) = 1,$$

obtain the Taylor series for $y(x)$ and compute $y(0.1)$, correct to four decimal places.

8.2 Show that the differential equation

$$\frac{d^2y}{dx^2} = -xy, \quad y(0) = 1 \text{ and } y'(0) = 0,$$

has the series solution

$$y = 1 - \frac{x^3}{3!} + \frac{1\times 4}{6!}x^6 - \frac{1\times 4\times 7}{9!}x^9 + \cdots$$

8.3 If

$$\frac{dy}{dx} = \frac{1}{x^2 + y} \quad \text{with } y(4) = 4,$$

compute the values of $y(4.1)$ and $y(4.2)$ by Taylor's series method.

8.4 Use Picard's method to obtain a series solution of the problem given in Problem 8.1 above.

8.5 Use Picard's method to obtain $y(0.1)$ and $y(0.2)$ of the problem defined by

$$\frac{dy}{dx} = x + yx^4, \quad y(0) = 3.$$

8.6 Using Euler's method, solve the following problems:

(a) $\frac{dy}{dx} = \frac{3}{5}x^3 y,\ y(0) = 1$ (b) $\frac{dy}{dx} = 1 + y^2,\ y(0) = 0$

8.7 Compute the values of $y(1.1)$ and $y(1.2)$ using Taylor's series method for the solution of the problem

$$y'' + y^2 y^1 = x^3,\ y(1) = 1 \text{ and } y'(1) = 1.$$

8.8 Find, by Taylor's series method, the value of $y(0.1)$ given that

$$y'' - xy' - y = 0,\ y(0) = 1 \text{ and } y'(0) = 0.$$

8.9 Using Picard's method, find $y(0.1)$, given that

$$\frac{dy}{dx} = \frac{y - x}{y + x} \quad \text{and} \quad y(0) = 1.$$

8.10 Using Taylor's series, find $y(0.1)$, $y(0.2)$ and $y(0.3)$ given that

$$\frac{dy}{dx} = xy + y^2, \quad y(0) = 1.$$

8.11 Given the differential equation

$$\frac{dy}{dx} = x^2 + y$$

with $y(0) = 1$, compute $y(0.02)$ using Euler's modified method.

8.12 Solve, by Euler's modified method, the problem

$$\frac{dy}{dx} = x + y,\ y(0) = 0.$$

Choose $h = 0.2$ and compute $y(0.2)$ and $y(0.4)$.

8.13 Given the problem

$$\frac{dy}{dx} = f(x, y) \text{ and } y(x_0) = y_0,$$

an approximate solution at $x = x_0 + h$ is given by the third order Runge–Kutta formula

$$y(x_0 + h) = y_0 + \frac{1}{6}(k_1 + 4k_2 + k_3) + R_4,$$

where

$$k_1 = hf(x_0, y_0), \quad k_2 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right)$$

and

$$k_3 = hf(x_0 + h, y_0 + 2k_2 - k_1).$$

Show that R_4 is of order h^4.

8.14 Write an algorithm to implement Runge–Kutta fourth order formula for solving an initial value problem.

Find $y(0,1)$, $y(0.2)$ and $y(0.3)$ given that

$$y' = 1 + \frac{2xy}{1+x^2}, \quad y(0) = 0$$

8.15 Use Runge–Kutta fourth order formula to find $y(0.2)$ and $y(0.4)$ given that

$$y' = \frac{y^2 - x^2}{y^2 + x^2}, \quad y(0) = 1.$$

8.16 Solve the initial value problem defined by

$$\frac{dy}{dx} = \frac{3x + y}{x + 2y}, \quad y(1) = 1$$

and find $y(1.2)$ and $y(1.4)$ by the Runge–Kutta fourth order formula.

8.17 State Adam's predictor-corrector formulae for the solution of the equation

$$y' = f(x, y), \; y(x_0) = y_0.$$

Given the problem

$$y' + y = 0, \; y(0) = 1,$$

find $y(0.1)$, $y(0.2)$, and $y(0.3)$ by Runge–Kutta fourth order formula and hence obtain $y(0.4)$ by Adam's formulae.

8.18 Given the initial value problem defined by

$$\frac{dy}{dx} = y(1 + x^2), \quad y(0) = 1$$

find the values of y for $x = 0.2, 0.4, 0.6, 0.8$ and 1.0 using the Euler, the modified Euler and the fourth order Runge–Kutta methods. Compare the computed values with the exact values.

8.19 State Milne's predictor-corrector formulae for the solution of the problem

$$y' = f(x, y), \; y(x_0) = y_0.$$

Given the initial value problem defined by

$$y' = y^2 + xy, \; y(0) = 1,$$

find, by Taylor's series, the values of $y(0.1)$, $y(0.2)$ and $y(0.3)$. Use these values to compute $y(0.4)$ by Milne's formulae.

8.20 Using Milne's formulae, find $y(0.8)$ given that

$$\frac{dy}{dx} = x - y^2, \quad y(0) = 0, \quad y(0.2) = 0.02,$$

$$y(0.4) = 0.0795 \quad \text{and } y(0.6) = 0.1762.$$

8.21 Explain what is meant by a fourth order formula. Discuss this with reference to the solution of the problem

$$\frac{dy}{dx} = 3x + \frac{y}{2}, \quad y(0) = 1$$

by Runge–Kutta fourth order formula.

8.22 Use Taylor's series method to solve the system of differential equations

$$\frac{dx}{dt} = y - t, \quad \frac{dy}{dt} = x + t$$

with $x = 1$, $y = 1$ when $t = 0$, taking $\Delta x = \Delta t = 0.1$.

8.23 Using fourth order Runge–Kutta method, compute the value of $y(0.2)$ given that

$$\frac{d^2y}{dx^2} + y = 0$$

with $y(0) = 1$ and $y'(0) = 0$.

8.24 Given that

$$y'' - xy' + 4y = 0, \quad y(0) = 3, \; y'(0) = 0,$$

compute the value of $y(0.2)$ using Runge–Kutta fourth order formla.

8.25 Solve the boundary value problem defined by

$$y'' - y = 0, \quad y(0) = 0, \quad y(1) = 1,$$

by finite difference and cubic spline methods. Compare the solutions obtained at $y(0.5)$ with the exact value. In each case, take $h = 0.5$ and $h = 0.25$.

8.26 *Shooting method* This is a popular method for the solution of two-point boundary value problems. If the problem is defined by

$$y'' = f(x), \quad y(x_0) = 0 \text{ and } y(x_1) = A,$$

then it is first transformed into the initial value problem

$$y'(x) = z, \quad z'(x) = f(x),$$

with $y(x_0) = 0$ and $z(x_0) = m_0$, where m_0 is a guess for the value of $y'(x_0)$. Let the solution corresponding to $x = x_1$ be Y_0. If Y_1 is the value obtained by another guess m_1 for $y'(x_0)$, then Y_0 and Y_1 are related linearly. Thus, linear interpolation can be carried out between the values (m_0, y_0) and (m_1, y_1).

Obviously, the process can be repeated till we obtain a value for $y(x_1)$ which is close to A.

Apply the shooting method to solve the boundary value problem

$$y'' = y(x),\ y(0) = 0 \text{ and } y(1) = 1.$$

8.27 Fyfe [1969] discussed the solution of the boundary value problem defined by

$$y'' + \frac{4x}{1+x^2}y' + \frac{2}{1+x^2}y = 0, \quad y(0) = 1 \text{ and } y(2) = 0.2.$$

Solve this problem by cubic spline method first with $h = 1$ and then with $h = 1/2$ to determine the value of $y(1)$. Compare your results with the exact values of $y(1)$ obtained from the analytical solution $y = 1/(1 + x^2)$.

8.28 *Method of Linear Interpolation* Let the boundary value problem be defined by

$$y'' + f(x)y' + g(x)y = p(x),$$
$$y(x_0) = y_0 \quad \text{and} \quad y(x_n) = y_n.$$

Set up the finite difference approximation of the differential equation and solve the algebraic equations using the initial value y_0 and assuming a value, say Y_0, for $y(x_1)$. Again, we assume another value for $y(x_1)$, say Y_1 and then compute the values of $y_2, y_3, \ldots, y_{n-1}$ and y_n. We, thus, have two sets of values of $y(x_1)$ and $y(x_n)$. Now we use linear interpolation formula to compute the value of $y(x_1)$ for which $y(x_n) = y_n$. The process is repeated until we obtain the value of $y(x_n)$ close to the given boundary condition (see Problem 8.26).

Solve the boundary value problem defined by

$$y'' + xy' - 2y = 0,\ y(0) = 1 \text{ and } y(1) = 2$$

using the method of Linear interpolation.

8.29 Using Galerkin's method, compute the value of $y(0.5)$ given that

$$y'' + y = x^2, \quad 0 < x < 1, \quad y(0) = 0 \text{ and } y(1) = 0.$$

8.30 Solve Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 2, \quad 0 \le x, y \le 1$$

with $u = 0$ on the boundary C of the square region $0 \le x \le 1$, $0 \le y \le 1$.

Answers to Exercises

8.1 1.1053

8.2 $1 - \frac{x^3}{3!} + \frac{1\times 4}{6!}x^6 - \frac{1\times 4\times 7}{9!}x^9 + \cdots$

8.3 4.005, 4.0098

8.4 $1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8} + \cdots$

8.5 3.005, 3.0202

8.6 (a) 1.0006, (b) $y_1 = 1.0000$, $y_2 = 0.201$, $y_3 = 0.3020$

8.7 1.1002, 1.2015

8.8 1.005012

8.9 1.0906

8.10 1.11686, 1.27730, 1.5023

8.11 1.0202

8.12 0.0222, 0.0938

8.14 0.1006, 0.2052, 0.3176

8.15 0.19598, 1.3751

8.16 1.2636, 1.532

8.17 $y(0.1) = 0.90484$, $y(0.2) = 0.81873$, $y(0.3) = 0.7408$
$y(0.4) = 0.6806$ (Exact value = 0.6703).

8.18 $x = 0.2$ 0.024664 (Euler)
0.003014 (Modified Euler)
0.000003 (Runge–Kutta)
$x = 1.0$ 0.776885 (Euler)
0.12157 (Modified Euler)
0.000273 (Runge–Kutta)

8.19 $y(0.1) = 1.1169$, $y(0.2) = 1.2773$, $y(0.3) = 1.5023$,
$y(0.4) = 1.8376$.

8.20 $y^p(0.8) = 0.3049$, $y^c(0.8) = 0.30460$

8.21 $h = 0.2$, error = 0.00000110
$h = 0.1$, error = 0.0000007

8.22 $x_1 = 1.1003$, $y(0.1) = y_1 = 1.1100$

8.23 0.980067 (Exact value = 0.980066)

8.24 $y(0.2) = 2.762239$ (Exact value = 2.7616)
$z(0.2) = -2.360566$ (Exact value = –2.368)

8.25 Exact value of $y(0.5) = 0.443409$
(a) 0.443674, (b) 0.443140

8.26 $y'(0) = 2.8$

8.27 (a) $h = 1, \quad y_1 = 0.4$

$h = \frac{1}{2}, \quad y_2 = 0.485714$

(b) $h = 1, \quad y_1 = 0.542373$

$h = \frac{1}{2}, \quad y_2 = 0.5228$

8.28 $y_1 = 1.0828, \quad y_2 = 1.2918, \quad y_3 = 1.6282, \quad y_4 = 1.99997$

8.29 $y(0.5) = -0.041665$ (Exact value = –0.04592)

8.30 $u(x, y) = -\frac{5}{2}xy(x - 1)(y - 1)$

Chapter 9

Numerical Solution of Partial Differential Equations

9.1 INTRODUCTION

Partial differential equations occur in many branches of applied mathematics, for example, hydrodynamics, elasticity, quantum mechanics and electromagnetic theory. The analytical treatment of these equations is a rather involved process since it requires application of advanced mathematical techniques. On the other hand, it is generally easier to produce sufficiently approximate solutions by simple and efficient numerical methods. There exist several numerical methods for the solution of partial differential equations; for example, finite difference methods, spline methods, finite element methods, integral equation methods, etc. Of these, only the finite difference methods have become popular and are more gainfully employed than others. In this chapter, we discuss these methods, very briefly, and apply them to solve simple problems. We also consider the application of cubic splines to parabolic and hyperbolic equations.

The general second order linear partial differential equation is of the form

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Fu = G,$$

which can be written as

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \tag{9.1}$$

where A, B, C, D, E, F and G are all functions of x and y. Equations of the form (9.1) can be classified with respect to the sign of the discriminant

$$\Delta = B^2 - 4AC, \tag{9.2}$$

where Δ is computed at any point in the (x, y) plane. Equation (9.1) is said to be *elliptic, parabolic* or *hyperbolic* according as $\Delta < 0$, $\Delta = 0$ or $\Delta > 0$.

For example,

$$u_{xx} + u_{yy} = 0 \quad \text{(Laplace equation) is elliptic} \tag{9.3}$$

$$u_{xx} - u_{yy} = 0 \quad \text{(Wave equation) is hyperbolic} \tag{9.4}$$

$$u_t = u_{xx} \quad \text{(heat conduction equation) is parabolic} \tag{9.5}$$

In the study of partial differential equations, usually three types of problems arise:

(i) *Dirichlet's Problem* Given a continuous function f on the boundary C of a region R, it is required to find a function $u(x, y)$, satisfying the Laplace equation in R, i.e., to find $u(x, y)$ such that

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \text{ in } R, \\ \text{and} \qquad u &= f \text{ on } C. \end{aligned}\right\} \tag{9.6}$$

(ii) *Cauchy's Problem.*

$$\left.\begin{aligned} u_{tt} - u_{xx} &= 0 \quad \text{for} \quad t > 0 \\ u(x, 0) &= f(x) \\ \text{and} \qquad \frac{\partial u(x,0)}{\partial t} &= g(x) \end{aligned}\right\} \tag{9.7}$$

where $f(x)$ and $g(x)$ are arbitrary.

(iii)

$$\left.\begin{aligned} u_t &= u_{xx} \quad \text{for} \quad t > 0 \\ \text{and} \qquad u(x, 0) &= f(x) \end{aligned}\right\} \tag{9.8}$$

In partial differential equations, the form of the equation is always associated with a particular type of boundary conditions. In this case, the problem is said to be *well-defined* (or well-posed). The problems defined in Eqs. (9.6) to (9.8) are well-posed. If, however, we associate Laplace equation with Cauchy boundary conditions, the problem is said to be *ill-posed*. Thus, the problem defined by

$$\left.\begin{aligned} u_{xx} + u_{yy} &= 0 \\ u(x, 0) &= f(x) \\ \text{and} \qquad u_y(x, 0) &= g(x) \end{aligned}\right\} \tag{9.9}$$

is an ill-posed problem.

9.2 LAPLACE'S EQUATION

The equation defined by

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{9.10}$$

is called *Laplace's equation.* It occurs in all problems involving potential functions and is of elliptic type. To derive this equation, we consider a heated plate which is insulated everywhere except at its edges where the temperature is constant. Assuming that the xy-plane coincides with one rectangular face PQRS (see Fig. 9.1), we find that the quantity of heat entering the face PS in time Δt

$$= -k\alpha\, \Delta y \left[\frac{\partial u}{\partial x}\right]_x \Delta t,$$

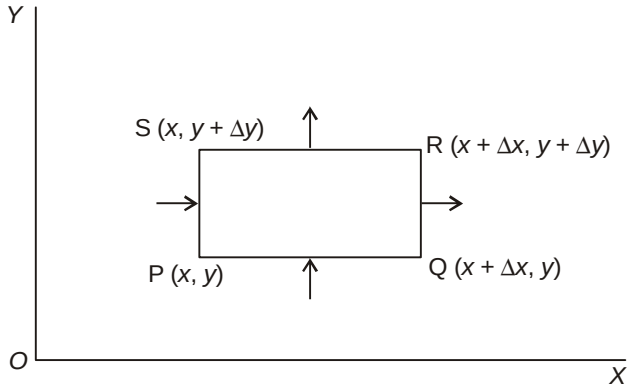


Figure 9.1 xy-plane coincides with a rectangular face PQRS.

where α is the thickness of the plate, $u(x, t)$ is the temperature at a distance x at time t and k is the conductivity of the material of the plate. Similarly, the amount of heat leaving the face QR in time Δt is

$$-k\alpha\, \Delta y \left[\frac{\partial u}{\partial x}\right]_{x+\Delta x} \Delta t$$

From the above two expressions, we obtain the gain of heat during time Δt

$$= k\alpha\, \Delta y \left(\left[\frac{\partial u}{\partial x}\right]_{x+\Delta x} - \left[\frac{\partial u}{\partial x}\right]_x\right) \Delta t$$

In the same way, we obtain the gain of heat from the faces PQ and SR in time Δt as

$$k\alpha\, \Delta x \left(\left[\frac{\partial u}{\partial x}\right]_{y+\Delta y} - \left[\frac{\partial u}{\partial y}\right]_y\right) \Delta t$$

Hence the total gain of heat in the plate

$$= k\alpha\ \Delta x \Delta y \left(\frac{\left[\frac{\partial u}{\partial x}\right]_{x+\Delta x} - \left[\frac{\partial u}{\partial x}\right]_{x}}{\Delta x} + \frac{\left[\frac{\partial u}{\partial y}\right]_{y+\Delta y} - \left[\frac{\partial u}{\partial y}\right]_{y}}{\Delta y} \right) \Delta t$$

This heat raises the temperature in the plate which is equal to

$$\rho\alpha s\ \Delta x\ \Delta y\ \Delta u,$$

where s is the specific heat and ρ is the density of the material. Hence we have

$$\rho\alpha s\ \Delta x \Delta y \Delta u = k\alpha\ \Delta x \Delta y \left(\frac{\left[\frac{\partial u}{\partial x}\right]_{x+\Delta x} - \left[\frac{\partial u}{\partial x}\right]_{x}}{\Delta x} + \frac{\left[\frac{\partial u}{\partial y}\right]_{y+\Delta y} - \left[\frac{\partial u}{\partial y}\right]_{y}}{\Delta y} \right) \Delta t \quad (9.11)$$

Dividing both sides of Eq. (9.11) by Δt and proceeding to the limit, we obtain

$$\frac{\partial u}{\partial t} = \frac{k}{\rho s}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

$$= a^2 \cdot \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right), \quad (9.12)$$

where $a^2 = \frac{k}{\rho s}$. Equation (9.12) gives the temperature distribution in the plate in the *transient state.* When the temperature $u(x, t)$ is constant, i.e., at the *steadystate* condition, $\frac{\partial u}{\partial t} = 0$ and Eq. (9.12) reduces to Laplace's equation (9.10).

Using the method of separation of variables, it can be shown that Eq. (9.10) possesses the solutions

$$u(x, y) = (c_1 e^{kx} + c_2 e^{-kx})(c_3 \cos ky + c_4 \sin ky) \quad (9.13)$$

$$u(x, y) = (c_5 \cos kx + c_6 \sin kx)(c_7 e^{ky} + c_8 e^{-ky}) \quad (9.14)$$

The proper form of solution has to be chosen depending upon the physical conditions of the problem.

9.3 FINITE-DIFFERENCE APPROXIMATIONS TO DERIVATIVES

Let the (x, y) plane be divided into a network of rectangles of sides $\Delta x = h$ and $\Delta y = k$ by drawing the sets of lines

$$x = ih, \qquad i = 0, 1, 2, \ldots$$

$$y = jk, \qquad j = 0, 1, 2, \ldots$$

The points of intersection of these families of lines are called *mesh* points, *lattice* points or *grid* points. Then, we have

$$u_x = \frac{u_{i+1,\,j} - u_{i,\,j}}{h} + O(h) \tag{9.15}$$

$$= \frac{u_{i,\,j} - u_{i-1,\,j}}{h} + O(h) \tag{9·16}$$

$$= \frac{u_{i+1,\,j} - u_{i-1,\,j}}{2h} + O(h^2) \tag{9·17}$$

and

$$u_{xx} = \frac{u_{i-1,\,j} - 2u_{i,\,j} + u_{i+1,\,j}}{h^2} + O(h^2) \tag{9.18}$$

where

$$u_{i,\,j} = u(ih,\ jk) = u(x, y)$$

Similarly, we have the approximations

$$u_y = \frac{u_{i,\,j+1} - u_{i,\,j}}{k} + O(k) \tag{9.19}$$

$$= \frac{u_{i,\,j} - u_{i,\,j-1}}{k} + O(k) \tag{9·20}$$

$$= \frac{u_{i,\,j+1} - u_{i,\,j-1}}{2k} + O(k^2) \tag{9·21}$$

and

$$u_{yy} = \frac{u_{i,\,j-1} - 2u_{i,\,j} + u_{i,\,j+1}}{k^2} + O(k^2) \tag{9.22}$$

We can now obtain the *finite-difference analogues* of partial differential equations by replacing the derivatives in any equation by their corresponding difference approximations given above. Thus, the Laplace equation in two dimensions, namely

$$u_{xx} + u_{yy} = 0$$

has its finite-difference analogue

$$\frac{1}{h^2}(u_{i+1,\,j} - 2u_{i,\,j} + u_{i-1,\,j}) + \frac{1}{k^2}(u_{i,\,j+1} - 2u_{i,\,j} + u_{i,\,j-1}) = 0. \tag{9.23}$$

If $h = k$, this gives

$$u_{i,j} = \frac{1}{4}(u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1}), \tag{9.24}$$

which shows that the value of u at any point is the mean of its values at the four neighbouring points. This is called the *standard five-point formula* [*see* Fig. 9.2(a)], and is written

$$u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1} - 4u_{i,\,j} = 0 \tag{9.25}$$

By expanding the terms on the right side of Eq. (9.24) by Taylor's series, it can be shown that

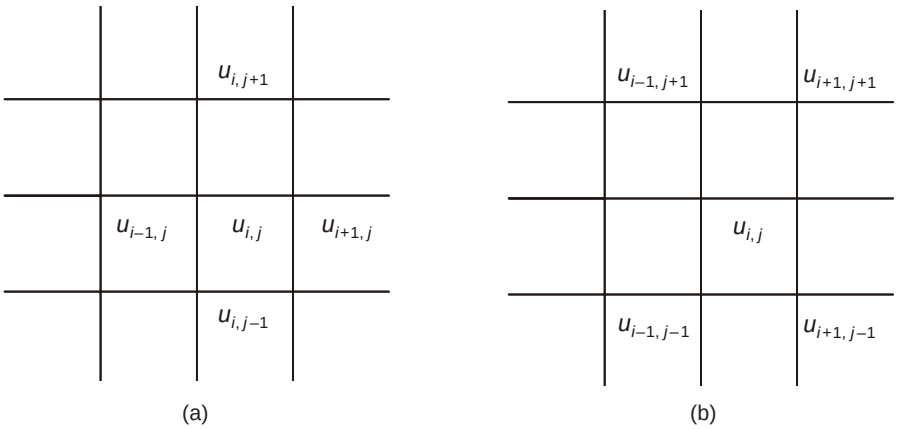
$$\begin{aligned} u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1} - 4u_{i,\,j} &= h^2(u_{xx} + u_{yy}) - \frac{1}{6}h^4 u_{xxyy} + \mathrm{O}(h^6) \\ &= -\frac{1}{6}h^4 u_{xxyy} + \mathrm{O}(h^6) \end{aligned} \tag{9.26}$$

Instead of formula given in Eq. (9.24), we may also use the formula

$$u_{i,\,j} = \frac{1}{4}(u_{i-1,\,j-1} + u_{i+1,\,j-1} + u_{i+1,\,j+1} + u_{i-1,\,j+1}) \tag{9.27}$$

which uses the function values at the diagonal points [see Fig. 9.2(b)], and is therefore called the *diagonal five-point formula*. This is perfectly valid since it is well-known that the Laplace equation remains invariant when the coordinate axes are rotated through 45°. Expanding the terms on the right side of Eq. (9.27) by Taylor's series, it can be shown that

$$u_{i-1,\,j-1} + u_{i+1,\,j-1} + u_{i+1,\,j+1} + u_{i-1,\,j+1} - 4u_{i,\,j} = \frac{2}{3}h^4 u_{xxyy} + \mathrm{O}(h^6) \tag{9.28}$$



(a) (b)

Figure 9.2 Approximations to Laplace's equation.

Neglecting terms of the order h^6, it follows from Eqs. (9.26) and (9.28) that the error in the diagonal formula is four times that in the standard formula. Hence, in all computations we should prefer to use the standard five-point formula, whenever possible.

Eliminating the term containing h^4 from both Eqs. (9.26) and (9.28), we obtain the *nine-point formula*

$$u_{i-1,\,j-1} + u_{i+1,\,j-1} + u_{i+1,\,j+1} + u_{i-1,\,j+1}$$
$$+ 4(u_{i+1,\,j} + u_{i-1,\,j} + u_{i,\,j+1} + u_{i,\,j-1}) - 20u_{i,\,j} = 0 \tag{9.29}$$

It is clear that the error in this formula is of order h^6. In a similar manner, the finite-difference analogues of Eqs. (9.4) and (9.5) can be obtained.

In this chapter, we consider those partial differential equations which can be replaced by the finite-difference analogues. These analogues, or *difference equations*, are then used as approximations to the concerned partial differential equations. Our general procedure is, therefore, to replace the partial differential equation by a finite-difference analogue and then obtain the solution at the mesh points. In the next section, we will discuss Laplace's equation which is generally solved by reduction to a system of algebraic equations. In Section 9.4, we will discuss methods for the numerical solution of the parabolic Eq. (9.5). Hyperbolic equations are considered in Section 9.8.

9.4 SOLUTION OF LAPLACE'S EQUATION

We wish to solve Laplace's equation

$$u_{xx} + u_{yy} = 0 \tag{9.30}$$

in a bounded region R with boundary C. As in Dirichlet's problem, let the value of u be specified everywhere on C. For simplicity, let R be a square region so that it can be divided into a network of small squares of side h. Let the values of $u(x, y)$ on the boundary C be given by c_i and let the interior mesh points and the boundary points be as in Fig. 9.3.

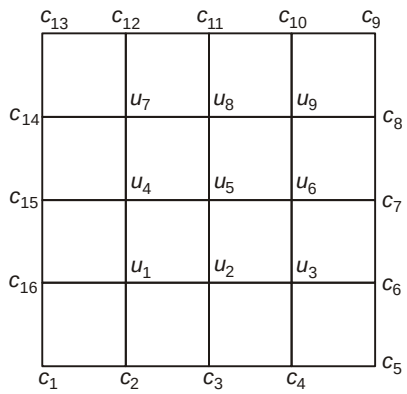


Figure 9.3 Interior mesh points and boundary points.

Then, as shown in the previous section, Eq. (9.30) can be replaced by either the standard five-point formula, viz. Eq. (9.25); or the diagonal five-point formula given in Eq. (9.27). The approximate function values at the interior mesh points can now be computed according to the scheme: we first use the diagonal five-point formula Eq. (9.27) and compute u_5, u_7, u_9, u_1 and u_3 in this order. Thus, we obtain

$$u_5 = \frac{1}{4}(c_1 + c_5 + c_9 + c_{13}); \qquad u_7 = \frac{1}{4}(c_{15} + u_5 + c_{11} + c_{13});$$

$$u_9 = \frac{1}{4}(u_5 + c_7 + c_9 + c_{11}); \qquad u_1 = \frac{1}{4}(c_1 + c_3 + u_5 + c_{15});$$

$$u_3 = \frac{1}{4}(c_3 + c_5 + c_7 + u_5).$$

We then compute, in the order, the remaining quantities, viz., u_8, u_4, u_6 and u_2 by the *standard five-point formula* (9.25). Thus, we have

$$u_8 = \frac{1}{4}(u_5 + u_9 + c_{11} + u_7); \qquad u_4 = \frac{1}{4}(u_1 + u_5 + u_7 + c_{15});$$

$$u_6 = \frac{1}{4}(u_3 + c_7 + u_9 + u_5); \qquad u_2 = \frac{1}{4}(c_3 + u_3 + u_5 + u_1).$$

When once all the $u_i, (i = 1, 2, 3, \ldots, 9)$ are computed, their accuracy can be improved by any of the iterative methods described below.

9.4.1 Jacobi's Method

Let $u_{i,j}^{(n)}$ denotes the nth iterative value of $u_{i,j}$. An iterative procedure to solve Eq. (9.25) is

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n)} + u_{i,j+1}^{(n)}] \tag{9.31}$$

for the interior mesh points. This is called the *point Jacobi method*.

9.4.2 Gauss–Seidel Method

The method uses the latest iterative values available and scans the mesh points systematically from left to right along successive rows. The iterative formula is:

$$u_{i,j}^{(n+1)} = \frac{1}{4}[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}] \tag{9.32}$$

It can be shown that the Gauss–Seidel scheme converges twice as fast as the Jacobi scheme. This method is also referred to as *Liebmann's method*.

9.4.3 Successive Over Relaxation (SOR) Method

Equation (9.32) can be written as

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \frac{1}{4}\left[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)} - 4u_{i,j}^{(n)}\right]$$

$$= u_{i,j}^{(n)} + \frac{1}{4}R_{i,j}$$

which shows that $(1/4)R_{i,j}$ is the change in the value of $u_{i,j}$ for one Gauss–Seidel iteration. In the SOR method, a larger change than this is given to $u_{i,j}^{(n)}$, and the iteration formula is written as

$$u_{i,j}^{(n+1)} = u_{i,j}^{(n)} + \frac{1}{4}\omega R_{i,j}$$

$$= \frac{1}{4}\omega\left[u_{i-1,j}^{(n+1)} + u_{i+1,j}^{(n)} + u_{i,j-1}^{(n+1)} + u_{i,j+1}^{(n)}\right] + (1-\omega)\,u_{i,j}^{(n)}$$

$$= \omega\, u_{i,j}^{(n+1)} + (1-\omega)\,u_{i,j}^{(n)} \tag{9.33}$$

The rate of convergence of Eq. (9.33) depends on the choice of ω, which is called the *accelerating factor* and lies between 1 and 2.

The percentage error in the value u_{ij} is given by

$$\left|\varepsilon_{ij}\right| = \left|\frac{u_{i,j}^{(n+1)} - u_{i,j}^{(n)}}{u_{ij}^{(n+1)}}\right| \times 100\% \tag{9.34}$$

It was shown by B.A. Carré that for $\omega = 1.875$, the rate of convergence of Eq. (9.33) is twice as fast as that when $\omega = 1$, and for $\omega = 1.9$, the rate of convergence is 40 times greater than that when $\omega = 1$. In general, however, it is difficult to estimate the best value of ω. The following examples illustrate the methods of solution.

Example 9.1 Solve Laplace's equation for the square region shown in Fig. 9.4, the boundary values being as indicated.

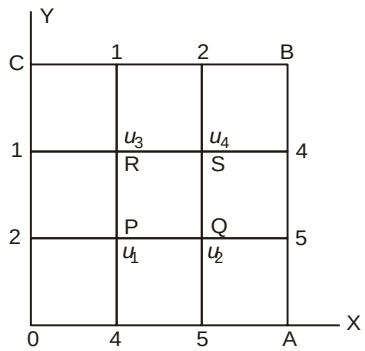


Figure 9.4

It is seen from the figure that the boundary values are symmetric about the diagonal AC. Hence, $u_1 = u_4$ and we need find only u_1, u_2 and u_3. The standard five-point formula applied at the point P gives

$$u_2 + u_3 + 2 + 4 - 4u_1 = 0.$$

Hence we have

$$u_1 = \frac{1}{4}(u_2 + u_3 + 6).$$

The iteration formula is therefore

$$u_1^{(n+1)} = \frac{1}{4}\left[u_2^{(n)} + u_3^{(n)} + 6\right].$$

Similarly, the iteration formulae at the points Q and R are given by

$$u_2^{(n+1)} = \frac{1}{2}u_1^{(n+1)} + \frac{5}{2},$$

and

$$u_3^{(n+1)} = \frac{1}{2}u_1^{(n+1)} + \frac{1}{2}.$$

For the first iteration, let $u_2 = 5$ (since it is nearer to the value $u = 5$), and $u_3^{(0)} = 1$. Hence

$$u_1^{(1)} = \frac{1}{4}(5 + 1 + 6) = 3,$$

$$u_2^{(1)} = \frac{1}{2}(3) + \frac{5}{2} = 4,$$

$$u_3^{(1)} = \frac{1}{2}(3) + \frac{1}{2} = 2.$$

For the second iteration, we have

$$u_1^{(2)} = \frac{1}{4}(4 + 2 + 6) = 3,$$

$$u_2^{(2)} = \frac{1}{2}(3) + \frac{5}{2} = 4,$$

and

$$u_3^{(2)} = \frac{1}{2}(3) + \frac{1}{2} = 2.$$

Since the values are unchanged, we conclude that $u_1 = 3$, $u_2 = 4$, $u_3 = 2$ and $u_4 = 3$.

Example 9.2 Solve the equation $u_{xx} + u_{yy} = 0$ in the domain of Fig. 9.5, below by (a) Jacobi's method, (b) Gauss–Seidel's method, and (c) SOR method.

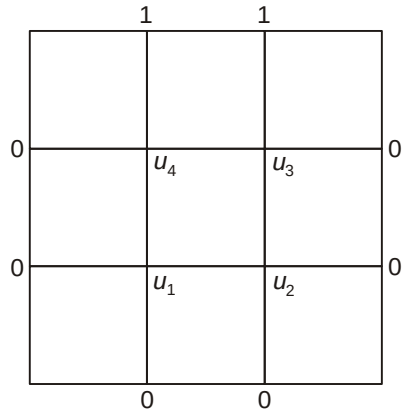


Figure 9.5

(a) To start *Jacobi's iteration process*, we obtain the approximate values of u_1, u_2, u_3 and u_4 as follows:

$$u_1^{(1)} = \frac{1}{4}(0+0+0+1) = 0.25;$$

$$u_2^{(1)} = \frac{1}{4}(0+0+0+1) = 0.25;$$

$$u_3^{(1)} = \frac{1}{4}(1+1+0+0) = 0.5;$$

$$u_4^{(1)} = \frac{1}{4}(1+1+0+0) = 0.5.$$

The iterations have been continued using Eq. (9.31), and seven successive iterates are given below:

u_1	u_2	u_3	u_4
0.1875	0.1875	0.4375	0.4375
0.15625	0.15625	0.40625	0.40625
0.14062	0.14062	0.39062	0.39062
0.13281	0.13281	0.38281	0.38281
0.12891	0.12891	0.37891	0.37891
0.12695	0.12695	0.37695	0.37695
0.12598	0.12598	0.37598	0.37598

(b) *Gauss–Seidel method:* Five successive iterates are given below:

u_1	u_2	u_3	u_4
0.25	0.3125	0.5625	0.46875
0.21875	0.17187	0.42187	0.39844
0.14844	0.13672	0.38672	0.38086
0.13086	0.12793	0.37793	0.37646
0.12646	0.12573	0.37573	0.37537

(c) *SOR method:* With $\omega = 1.1$, three successive iterates obtained by using Eq. (9.33) are given below.

u_1	u_2	u_3	u_4
0.275	0.35062	0.35062	0.35062
0.16534	0.10683	0.38183	0.37432
0.11785	0.12181	0.37216	0.37341

Example 9.3 Solve Laplace's equation for Fig. 9.6 given below:

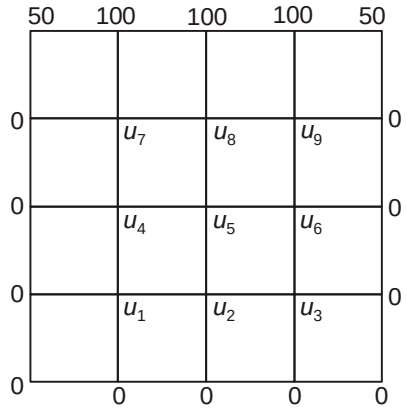


Figure 9.6

We first compute the quantities u_5, u_7, u_9, u_1 and u_3 by using the diagonal five-point formula given in Eq. (9.27). Thus, we obtain

$$u_5^{(1)} = 25.00; \qquad u_7^{(1)} = 42.75; \qquad u_9^{(1)} = 43.75;$$

$$u_1^{(1)} = 6.25; \qquad u_3^{(1)} = 6.25.$$

We now compute u_8, u_4, u_6 and u_2 successively by using the standard five-point formula given in Eq. (9.25)

$$u_8^{(1)} = 53.12; \qquad u_4^{(1)} = 18.75;$$

$$u_6^{(1)} = 18.75; \qquad u_2^{(1)} = 9.38.$$

We have thus obtained the first approximations of all the nine mesh points and we can now use one of the iterative formulae given in Section 9.4. We give below the first-four iterates obtained by using the Gauss–Seidel formula:

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
7.03	9.57	7.08	18.94	25.10	18.98	43.02	52.97	42.99
7.13	9.83	7.20	18.81	25.15	18.84	42.94	52.77	42.90
7.16	9.88	7.18	18.81	25.08	18.79	42.89	52.72	42.88
7.17	9.86	7.16	18.78	25.04	18.77	42.88	52.70	42.87

Example 9.4 Solve the Poisson equation

$$u_{xx} + u_{yy} = -10\,(x^2 + y^2 + 10).$$

in the domain of Fig. 9.7.

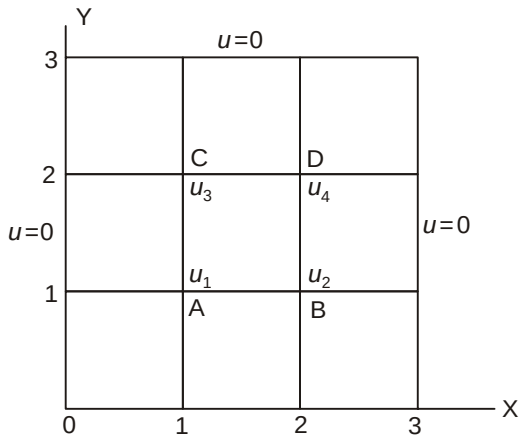


Figure 9.7

Let the values of u at the four grid points, A, B, C, D be u_1, u_2, u_3, u_4, respectively. Let the grid points be defined by $x = ih$, $y = jh$, where $h = 1$, $i, j = 0, 1, 2, 3$. At the point A, $i = 1$, $j = 1$. The standard five-point formula applied at the point A gives

$$u_2 + u_3 + 0 + 0 - 4u_1 = -10\,(1 + 1 + 10)$$

i.e.,

$$u_1 = \frac{1}{4}\,(u_2 + u_3 + 120) \tag{i}$$

Again, the standard five-point formula applied at the point B gives

$$u_1 + u_4 + 0 + 0 - 4u_2 = -10\,(4 + 1 + 10)$$

i.e.,

$$u_2 = \frac{1}{4}\,(u_1 + u_4 + 150) \tag{ii}$$

Similarly, the standard five-point formula applied at the points C and D gives, respectively

$$u_3 = \frac{1}{4}\,(u_1 + u_4 + 150) \tag{iii}$$

and

$$u_4 = \frac{1}{4}\,(u_2 + u_3 + 180) \tag{iv}$$

From (ii) and (iii), it is seen that $u_2 = u_3$ and so we need to find only u_1, u_2 and u_4 from (i), (ii) and (iv). The iteration formulae are therefore given by

$$u_1^{(n+1)} = \frac{1}{2} u_2^{(n)} + 30$$

$$u_2^{(n+1)} = \frac{1}{4}\left[u_1^{(n+1)} + u_4^{(n)} + 150\right]$$

$$u_4^{(n+1)} = \frac{1}{2} u_2^{(n+1)} + 45.$$

For the first iteration, we assume that $u_2^{(0)} = u_4^{(0)} = 0$. Hence we obtain

$$u_1^{(1)} = 30,$$

$$u_2^{(1)} = \frac{1}{4}(30 + 0 + 150) = 45$$

$$u_4^{(1)} = \frac{1}{2}(45) + 45 = 67.5.$$

For the second iteration, we have

$$u_1^{(2)} = \frac{1}{2} u_2^{(1)} + 30 = \frac{1}{2}(45) + 30 = 52.5$$

$$u_2^{(2)} = \frac{1}{4}\left[u_1^{(2)} + u_4^{(1)} + 150\right] = \frac{1}{4}[52.5 + 67.5 + 150] = 67.5$$

$$u_4^{(2)} = \frac{1}{2}\left[u_2^{(2)}\right] + 45 = 78.75.$$

For the third iteration, we obtain

$$u_1^{(3)} = \frac{1}{2} u_2^{(2)} + 30 = \frac{1}{2}(67.5) + 30 = 63.75$$

$$u_2^{(3)} = \frac{1}{4}\left[u_1^{(3)} + u_4^{(2)} + 150\right] = \frac{1}{4}[63.75 + 78.75 + 150] = 73.125.$$

$$u_4^{(3)} = \frac{1}{2} u_2^{(3)} + 45 = \frac{1}{2}(73.125) + 45 = 81.5625.$$

The fourth iteration gives

$$u_1^{(4)} = \frac{1}{2} u_2^{(3)} + 30 = \frac{1}{2}(73.125) + 30 = 66.5625$$

$$u_2^{(4)} = \frac{1}{4}\left[u_1^{(4)} + u_4^{(3)} + 150\right] = \frac{1}{4}[66.5625 + 81.5625 + 150] = 74.53125$$

$$u_4^{(4)} = \frac{1}{2} u_2^{(4)} + 45 = \frac{1}{2}(74.53125) + 45 = 82.2656.$$

For the fifth iteration, we obtain

$$u_1^{(5)} = \frac{1}{2}u_2^{(4)} + 30 = \frac{1}{2}(74.53125) + 30 = 67.2656$$

$$u_2^{(5)} = \frac{1}{4}\left[u_1^{(5)} + u_4^{(4)} + 150\right] = \frac{1}{4}[67.2656 + 82.2656 + 150] = 74.8828$$

$$u_4^{(5)} = \frac{1}{2}u_2^{(5)} + 45 = \frac{1}{2}(74.8828) + 45 = 82.4414.$$

The sixth iteration gives

$$u_1^{(6)} = \frac{1}{2}u_2^{(5)} + 30 = \frac{1}{2}(74.8828) + 30 = 67.4414.$$

$$u_2^{(6)} = \frac{1}{4}\left[u_1^{(6)} + u_4^{(5)} + 150\right] = \frac{1}{4}[67.4414 + 82.4414 + 150] = 74.9707.$$

$$u_4^{(6)} = \frac{1}{2}u_2^{(6)} + 45 = \frac{1}{2}(74.9707) + 45 = 82.4854.$$

From the last two iterates, we conclude that

$$u_1 = 67, \quad u_2 = u_3 = 75, \quad \text{and} \quad u_4 = 83.$$

9.4.4 ADI Method

This is an efficient method for the numerical solution of elliptic partial differential equations and was proposed by Peaceman and Rachford. It is quite general but, for easy understanding, we demonstrate its applicability with reference to the Laplace equation in two dimensions. For more details, the reader is referred to Isaacson and Keller [1966].

We consider Laplace's equation in two dimensions, viz.,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{9.35}$$

and the standard five-point formula

$$u_{i-1,\,j} + u_{i+1,\,j} + u_{i,\,j-1} + u_{i,\,j+1} - 4u_{i,\,j} = 0 \tag{9.25}$$

The use of formula given in Eq. (9.25) involves the solution of a system of algebraic equations, whose coefficient matrix, for $n = 6$, is of the form

$$A = \begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 \\ 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix} \tag{9.36}$$

The general form of such a system is given by

$$B = \begin{bmatrix} T & I & & & 0 & \\ I & T & I & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & I & T & I \\ & & & & I & T \end{bmatrix}, \tag{9.37}$$

where T is a tridiagonal matrix of the form

$$T = \begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix} \tag{9.38}$$

System A is called a *block tridiagonal* system and such systems are solved by Gaussian elimination or, in the case of large systems, by Gauss–Seidel iterations. But tridiagonal systems of the type of Eq. (9.38) are much easier to solve than block tridiagonal systems. Hence the question arises as to whether we can obtain directly tridiagonal systems in the numerical solution of Laplace's equation. Peaceman and Rachford showed that this is possible and their method of procedure, called the *alternating direction implicit* method (or the ADI method) is described below.

We rearrange Eq. (9.25) in either of two ways:

$$u_{i-1,\,j} - 4u_{i,\,j} + u_{i+1,\,j} = -u_{i,\,j-1} - u_{i,\,j+1} \tag{9.39}$$

or

$$u_{i,\,j-1} - 4u_{i,\,j} + u_{i,\,j+1} = -u_{i-1,\,j} - u_{i+1,\,j} \tag{9.40}$$

The ADI is an *iteration* method and Eqs. (9.39) and (9.40) are used as iteration formulae

$$u_{i-1,\,j}^{(r+1)} - 4u_{i,\,j}^{(r+1)} + u_{i+1,\,j}^{(r+1)} = -u_{i,\,j-1}^{(r)} - u_{i,\,j+1}^{(r)} \tag{9.41}$$

and

$$u_{i\,j-1}^{(r+2)} - 4u_{i,\,j}^{(r+2)} + u_{i\,j+1}^{(r+2)} = -u_{i-1,\,j}^{(r+1)} - u_{i+1,\,j}^{(r+1)} \tag{9.42}$$

Equation (9.41) is used to compute function values at all internal mesh points along rows and Eq. (9.42) those along columns. For $j = 1, 2, 3, \ldots, n-1$, Eq. (9.41) yields a tridiagonal system of equations and can easily be solved. Similarly, for $i = 1, 2, 3, \ldots, n-1$, Eq. (9.42) also yields a tridiagonal system of equations.

In the ADI method, formulae (9.41) and (9.42) are used alternately. For example, for the first row $j = 1$, and Eq. (9.41) gives

$$u_{i-1,1}^{(r+1)} - 4u_{i,1}^{(r+1)} + u_{i+1,1}^{(r+1)} = -u_{i,0}^{(r)} - u_{i,2}^{(r)}, \qquad (i = 1, 2, 3, \ldots, n-1) \tag{9.43}$$

Together with the boundary conditions, Eq. (9.43) represents a tridiagonal system of equations and are easily solved for $u_{i,1}^{(r+1)}$. We next put $j = 2$ and obtain the values of $u_{i,2}^{(r+1)}$ on the second row. The process is repeated for all the rows, viz. up to $j = n-1$. We next alternate the direction, i.e. we use Eq. (9.42) to compute $u_{i,j}^{(r+2)}$. It is easy to see that at every stage we will be solving a tridiagonal system of equations. Example 9.5 demonstrates the method of solution.

Example 9.5 Solve Laplace's equation, $u_{xx} + u_{yy} = 0$, in the domain of Fig. 9.8 (see Example 9.2).

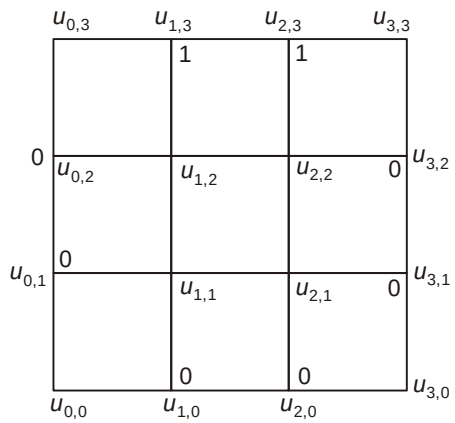


Figure 9.8

To apply formulae given in Eqs. (9.41) and (9.42), we relabel the internal mesh points, as in Fig. 9.8.

To start the iterations, we set $r = 0$. For the first row, $j = 1$. Then, Eq. (9.41) gives

$$u_{i-1,1}^{(1)} - 4u_{i,1}^{(1)} + u_{i+1,1}^{(1)} = -u_{i,0}^{(0)} - u_{i,2}^{(0)}. \tag{i}$$

With $i = 1$ and $i = 2$, this gives two equations

$$u_{0,1}^{(1)} - 4u_{1,1}^{(1)} + u_{2,1}^{(1)} = -u_{1,0}^{(0)} - u_{1,2}^{(0)}$$

and

$$u_{1,1}^{(1)} - 4u_{2,1}^{(1)} + u_{3,1}^{(1)} = -u_{2,0}^{(0)} - u_{2,2}^{(0)}.$$

Substituting the boundary values and assuming that $u_{1,2}^{(0)} = 1$ and $u_{2,2}^{(0)} = 1$, the above equations yield

$$u_{1,1}^{(1)} = u_{2,1}^{(1)} = \frac{1}{3} = 0.3333.$$

For computing the function values on the second row, we set $j = 2$ in (9.41) to obtain

$$u_{i-1,2}^{(1)} - 4u_{i,2}^{(1)} + u_{i+1,2}^{(1)} = -u_{i,1}^{(0)} - u_{i,3}^{(0)} \tag{ii}$$

With $i = 1$ and $i = 2$, Eq. (ii) gives

$$u_{0,2}^{(1)} - 4u_{1,2}^{(1)} + u_{2,2}^{(1)} = -u_{1,1}^{(0)} - u_{1,3}^{(0)}$$

and

$$u_{1,2}^{(1)} - 4u_{2,2}^{(1)} + u_{3,2}^{(1)} = -u_{2,1}^{(0)} - u_{2,3}^{(0)}$$

Substituting the boundary values and solving the above, we obtain

$$u_{1,2}^{(1)} = u_{2,2}^{(1)} = \frac{1}{3} = 0.3333.$$

Having completed the computations on the two rows, we now alternate the direction and compute the function values on the columns, starting with the first one. For this, we use Eq. (9.42) with $r = 0$. Setting $i = 1$, Eq. (9.42) becomes

$$u_{1,j-1}^{(2)} - 4u_{1,j}^{(2)} + u_{1,j+1}^{(2)} = -u_{0,j}^{(1)} - u_{2,j}^{(1)} \tag{iii}$$

Putting $j = 1$ and $j = 2$ in the above, we obtain the equations

$$u_{1,0}^{(2)} - 4u_{1,1}^{(2)} + u_{1,2}^{(2)} = -u_{0,1}^{(1)} - u_{2,1}^{(1)}$$

and

$$u_{1,1}^{(2)} - 4u_{1,2}^{(2)} + u_{1,3}^{(2)} = -u_{0,2}^{(1)} - u_{2,2}^{(1)}.$$

Substituting the boundary values and solving the above equations, we obtain

$$u_{1,1}^{(2)} = \frac{8}{45} = 0.1778 \quad \text{and} \quad u_{1,2}^{(2)} = \frac{17}{45} = 0.3778$$

To compute the values on the second column, we now set $i = 2$ in Eq. (9.42)

$$u_{2,j-1}^{(2)} - 4u_{2,j}^{(2)} + u_{2,j+1}^{(2)} = -u_{1,j}^{(1)} - u_{3,j}^{(1)} \tag{iv}$$

Putting $j = 1$ and $j = 2$ in the above, we obtain the equations

$$u_{2,0}^{(2)} - 4u_{2,1}^{(2)} + u_{2,2}^{(2)} = -u_{1,1}^{(1)} - u_{3,1}^{(1)}$$

and

$$u_{2,1}^{(2)} - 4u_{2,2}^{(2)} + u_{2,3}^{(2)} = -u_{1,2}^{(1)} - u_{3,2}^{(1)}$$

Substituting the boundary values in the above two equations and solving them, we obtain

$$u_{2,1}^{(2)} = 0.1778 \quad \text{and} \quad u_{2,2}^{(2)} = 0.3778$$

The iterations are continued to improve the function values obtained first on the rows, then on the columns, and so on. The reader is advised to continue these computations for the next iteration.

9.5 HEAT EQUATION IN ONE DIMENSION

The heat equation in one dimension is a typical parabolic partial differential equation and is a time variable problem. Equation (9.12), derived in Section 9.2, models two-dimensional heat conduction in a plate. Instead of a plate, if we consider a long thin insulated rod and equate the amount of heat absorbed to the difference between the amount of heat entering a small element and that leaving the element in time Δt, we obtain the partial differential equation

$$\frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2} \tag{9.44}$$

where

$$\alpha^2 = \frac{k}{s\rho} \tag{9.45}$$

In Eq. (9.45), k is the coefficient of conductivity of the material, ρ is its density and s is its specific heat. Analytical solutions of Eq. (9.44), obtained by the method of separation of variables are given by

$$\left.\begin{aligned} u(x, t) &= e^{-p^2\alpha^2 t}(c_1 \cos px + c_2 \sin px) \\ u(x, t) &= e^{p^2\alpha^2 t}\,(c_2 e^{px} + c_3 e^{-px}) \end{aligned}\right\} \tag{9.46}$$

From Eq. (9.46), the appropriate form of solution should be chosen depending upon the boundary conditions given. It is clear that to solve Eq. (9.44), we need one initial condition and two boundary conditions. In the sequel, we shall discuss the *finite difference* and *cubic spline approximations* to this equation.

9.5.1 Finite-difference Approximations

We divide the (x, t) plane into smaller rectangles by means of the sets of lines

$$x = ih, \quad i = 0, 1, 2, \ldots$$

$$t = kl, \quad k = 0, 1, 2, \ldots$$

where $h = \Delta x$ and $l = \Delta t$. Denoting $u(ih, kl) = u_i^k$, we have

$$\frac{\partial u}{\partial t} \approx \frac{u_i^{k+1} - u_i^k}{l} \tag{9.47}$$

and

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) \tag{9.48}$$

Equation (9.44) is replaced by the finite difference analogue

$$\frac{u_i^{k+1} - u_i^k}{l} = \alpha^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2},$$

which simplifies to

$$u_i^{k+1} = \lambda u_{i-1}^k + u_{i+1}^k + (1 - 2\lambda) u_i^k, \tag{9.49}$$

where

$$\lambda = \frac{\alpha^2 l}{h^2} \tag{9.50}$$

In Eq. (9.49), u_i^{k+1} is expressed *explicitly* in terms of u_{i-1}^k, u_{i+1}^k and u_i^k. Hence it is called the *explicit* formula for the solution of one-dimensional heat equation. It can be shown that Eq. (9.49) is valid only for $0 \le \lambda \le \frac{1}{2}$, which is called the *stability condition* for the explicit formula.

If we set $\lambda = \frac{1}{2}$ in Eq. (9.49), we obtain the simple formula

$$u_i^{k+1} = \frac{1}{2}\left(u_{i-1}^k + u_{i+1}^k\right) \tag{9.51}$$

which is called *Bender–Schmidt recurrence formula*. It is clear that Eqs. (9.49) and (9.51) have limited application because of the restriction on the values of λ. A formula which does not have any restriction on λ is that due to Crank and Nicolson. In Eq. (9.44), if we replace $\frac{\partial^2 u}{\partial x^2}$ by the average of its finite difference approximations on the kth and $(k+1)$th time levels, we obtain

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{2h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$$

Hence, Eq. (9.44) is approximated by

$$\frac{u_i^{k+1} - u_i^k}{l} = \frac{\alpha^2}{2h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)$$

which simplifies to

$$-\lambda u_{i-1}^{k+1} + (2 + 2\lambda) u_i^{k+1} - \lambda u_{i+1}^{k+1} = \lambda u_{i-1}^k + (2 - 2\lambda) u_i^k + \lambda u_{i+1}^k \tag{9.52}$$

On the left side of Eq. (9.52) we have three unknowns and on the right side, all are known quantities. This is called *Crank–Nicolson formula* for the one-dimensional heat equation and it is an *implicit formula.*

It is convergent *for all finite values of* λ. If there are N internal mesh points on each time row, then Eq. (9.52) gives N simultaneous equations for the N unknowns. In a similar way, values of u on all time rows can be calculated.

Example 9.6 Use the Bender–Schmidt formula to solve the heat conduction problem

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}$$

with the conditions $u(x, 0) = 4x - x^2$ and $u(0, t) = u(4, t) = 0$.

Setting $h = 1$, we see that $l = 1$ when $\lambda = \frac{1}{2}$.

Now, the initial values are

$$u(0, 0) = 0, \quad u(1, 0) = 3,$$
$$u(2, 0) = 4, \quad u(3, 0) = 3$$
$$\text{and} \quad u(4, 0) = 0.$$

Further, $u(0, t) = u(4, t) = 0$.

For $l = 1$, Bender–Schmidt formula gives

$$u_1^1 = \frac{1}{2}(0 + 4) = 2,$$

$$u_2^1 = \frac{1}{2}(3 + 3) = 3,$$

$$u_3^1 = \frac{1}{2}(4 + 0) = 2.$$

Similarly, for $l = 2$, we obtain

$$u_1^2 = \frac{1}{2}(0 + 3) = 1.5,$$

$$u_2^2 = \frac{1}{2}(2 + 2) = 2,$$

$$u_3^2 = \frac{1}{2}(3 + 0) = 1.5.$$

Continuing in this way, we obtain

$$u_1^3 = 1, \qquad u_2^3 = 1.5, \qquad u_3^3 = 1,$$

$$u_1^4 = 0.75, \qquad u_2^4 = 1, \qquad u_3^4 = 0.75,$$

$$u_1^5 = 0.5, \qquad u_2^5 = 0.75, \qquad u_3^5 = 0.5, \text{ and so on.}$$

Example 9.7 Solve the heat conduction problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions $u(x, 0) = \sin \pi x$, $0 \le x \le 1$, and $u(0, t) = u(1, t) = 0$. Use Bender–Schmidt's and Crank–Nicolson formulae to compute the value of $u(0.6, 0.04)$ and compare the results with the exact value.

The exact solution of this problem is given by

$$u(x, t) = e^{-\pi^2 t} \sin \pi x,$$

so that the exact value of $u(0.6, 0.04)$ is 0.6408.

(a) *Bender–Schmidt formula*

Let $h = 0.2$. Then $l = \lambda h^2 = \frac{1}{2}(0.04) = 0.02$.

The initial values of u are

$$u_0^0 = 0, \qquad u_1^0 = 0.5878, \qquad u_2^0 = 0.9510,$$

$$u_3^0 = 0.9510, \qquad u_4^0 = 0.5878, \qquad u_5^0 = 0.$$

Then Bender–Schmidt formula gives

$$u_1^1 = \frac{1}{2}(0.9510) = 0.4755, \qquad u_2^1 = \frac{1}{2}(0.5878 + 0.9510) = 0.7694,$$

$$u_3^1 = 0.7694, \qquad u_4^1 = 0.4755.$$

Also,

$$u_1^2 = \frac{1}{2}(0.7694) = 0.3847, \qquad u_2^2 = 0.62245,$$

$$u_3^2 = 0.62245, \qquad u_4^2 = 0.3847.$$

Therefore, $u(0.6, 0.04) = u_3^2 = 0.6224$, the error in which is 0.0184.

(b) *Crank–Nicolson formula*

Let $h = 0.2$ and $l = 0.04$, so that $\lambda = 1$.

For $\lambda = 1$, Crank–Nicolson formula becomes

$$-u_{i-1}^{k+1} + 4u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k + u_{i+1}^k \tag{i}$$

Putting $k = 0$ in (i), we obtain

$$-u_{i-1}^1 + 4u_i^1 - u_{i+1}^1 = u_{i-1}^0 + u_{i+1}^0$$

Corresponding to $i = 1, 2, 3$, and 4, we obtain the four equations

$$4u_1^1 - u_2^1 = 0.9510$$

$$-u_1^1 + 4u_2^1 - u_3^1 = 1.5388$$

$$-u_2^1 + 4u_3^1 - u_4^1 = 1.5388$$

$$-u_3^1 + 4u_4^1 = 0.9510$$

By symmetry, we have

$$u_1^1 = u_4^1 \quad \text{and} \quad u_2^1 = u_3^1$$

Solving the above system, we obtain

$$u_2^1 = u_3^1 = 0.6460$$

Hence, $u(0.6,\ 0.04) \approx 0.6460$, the error in which is 0.0052.

Example 9.8 Solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions

$$u(x,\ 0) = 0, \quad u(0,\ t) = 0 \quad \text{and} \quad u(1,\ t) = t.$$

Using Crank–Nicolson scheme, find the value of $u\left(\frac{1}{2}, \frac{1}{8}\right)$ taking successively (i) $h = \frac{1}{2}$, $l = \frac{1}{8}$, (ii) $h = \frac{1}{4}$, $l = \frac{1}{8}$. Compare the results obtained with the exact value of $u\left(\frac{1}{2}, \frac{1}{8}\right) = 0.01878$.

(i) $h = \frac{1}{2}$, $l = \frac{1}{8}$. Then $\lambda = \frac{1}{2}$.

Crank–Nicolson scheme gives

$$-u_{i-1}^{k+1} + 6u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k + 2u_i^k + u_{i+1}^k.$$

Setting $k = 0$ and $i = 1$ in the above equation, we obtain

$$-u_0^1 + 6u_1^1 - u_2^1 = u_0^0 + 2u_1^0 + u_2^0 = 0.$$

$$u_1^1 = \frac{1}{48}, \quad \text{since } u_0^1 = 0 \text{ and } u_2^1 = \frac{1}{8}.$$

$$= 0.02083 \text{ (error} = 0.00205).$$

(ii) $h = \frac{1}{4}$, $l = \frac{1}{8}$. Then $\lambda = 2$.

Therefore, Crank–Nicolson scheme gives

$$-u_{i-1}^{k+1} + 3u_i^{k+1} - u_{i+1}^{k+1} = u_{i-1}^k - u_i^k + u_{i+1}^k$$

With $k = 0$, we obtain

$$-u_{i-1}^1 + 3u_i^1 - u_{i+1}^1 = u_{i-1}^0 - u_i^0 + u_{i+1}^0 = 0, \ i = 1, 2, 3.$$

Corresponding to i = 1, 2 and 3, we obtain the three equations

$$-3u_1^1 - u_2^1 = 0$$
$$u_1^1 - 3u_2^1 + u_3^1 = 0$$
$$u_2^1 - 3u_3^1 = -\frac{1}{8}.$$

Solving the above system, we obtain

$$u_2^1 = u\left(\frac{1}{2}, \frac{1}{8}\right) \approx \frac{1}{56} = 0.01786 \text{ (error} = 0.00092).$$

9.6 ITERATIVE METHODS FOR THE SOLUTION OF EQUATIONS

The iterative methods discussed in Section 9.4 can be applied to solve the finite-difference equations obtained in the preceding section. In the Crank–Nicolson method, the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

is replaced by the finite-difference equation

$$(1+r)u_{i,\,j+1} = u_{i,\,j} + \frac{1}{2}r(u_{i-1,\,j+1} + u_{i+1,\,j} + u_{i+1,\,j+1} + u_{i-1,\,j} - 2u_{i,\,j}) \tag{9.53}$$

where $r = k/h^2$.

In Eq. (9.53), the unknowns are $u_{i,\,j+1}, u_{i-1,\,j+1}$ and $u_{i+1,\,j+1}$, and all others are known since they were already computed at the jth step. Hence, dropping the j's and setting

$$c_i = u_{i,\,j} + \frac{1}{2}r(u_{i-1,\,j} - 2u_{i,\,j} + u_{i+1,\,j}) \tag{9.54}$$

Eq. (9.53) can be written as

$$u_i = \frac{r}{2(1+r)}(u_{i-1} + u_{i+1}) + \frac{c_i}{1+r} \tag{9.55}$$

From Eq. (9.55), we obtain the iteration formula

$$u_i^{(n+1)} = \frac{r}{2(1+r)}[u_{i-1}^{(n)} + u_{i+1}^{(n)}] + \frac{c_i}{1+r}, \tag{9.56}$$

which expresses the $(n+1)$th iterate in terms of the nth iterates only, and is known as *Jacobi's iteration formula.*

It can be seen from Eq. (9.56) that at the time of computing $u_i^{(n+1)}$, the latest value of u_{i-1}, namely $u_{i-1}^{(n+1)}$, is already available. Hence, the convergence of Jacobi's iteration formula can be improved by replacing $u_{i-1}^{(n)}$ in formula given in Eq. (9.56) by its latest value available, namely by $u_{i-1}^{(n+1)}$. Accordingly, we obtain the formula

$$u_i^{(n+1)} = \frac{r}{2(1+r)}[u_{i-1}^{(n+1)} + u_{i+1}^{(n)}] + \frac{c_i}{1+r} \tag{9.57}$$

which is called the *Gauss–Seidel iteration formula*. It can be shown that Eq. (9.57) converges for all finite values of r and that it converges twice as fast as Jacobi's scheme.

Equation (9.57) can be rewritten as

$$u_i^{(n+1)} = u_i^{(n)} + \left\{\frac{r}{2(1+r)}[u_{i-1}^{(n+1)} + u_{i+1}^{(n)}] + \frac{c_i}{1+r} - u_i^{(n)}\right\}$$

from which it is clear that the expression within the curly brackets is the difference between the nth and $(n+1)$th iterates. If we take the difference to be ω times this expression, we then obtain

$$u_i^{(n+1)} = u_i^{(n)} + \omega\left\{\frac{r}{2(1+r)}[u_{i-1}^{(n+1)} + u_{i+1}^{(n)}] + \frac{c_i}{1+r} - u_i^{(n)}\right\} \tag{9.58}$$

which is called the *successive over-relaxation* (or SOR) method. ω is called the *relaxation factor* and it lies, generally, between 1 and 2.

Example 9.9 Solve

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the initial condition $u = \sin \pi x$ at $t = 0$ for $0 \le x \le 1$ and $u = 0$ at $x = 0$ and $x = 1$ for $t > 0$, by the *Gauss–Seidel method*.

We choose $h = 0.2$ and $k = 0.02$ so that $r = k/h^2 = 1/2$. Equation (9.57) therefore becomes

$$u_i^{(n+1)} = \frac{1}{6}[u_{i-1}^{(n+1)} + u_{i+1}^{(n)}] + \frac{2}{3}c_i \tag{i}$$

Let the values of u at the interior mesh points on the row corresponding to $t = 0.02$ be u_1, u_2, u_3, u_4, as shown in Fig. 9.9.

Applying formula given in Eq. (i) at the four interior mesh points, we obtain successively

$$u_1^{(n+1)} = \frac{1}{6}[0 + u_2^{(n)}] + \frac{2}{3}\left[0.5878 + \frac{1}{4}(0 - 2\times 0.5878 + 0.9511)\right]$$

$$= \frac{1}{6}u_2^{(n)} + 0.3544 \tag{ii}$$

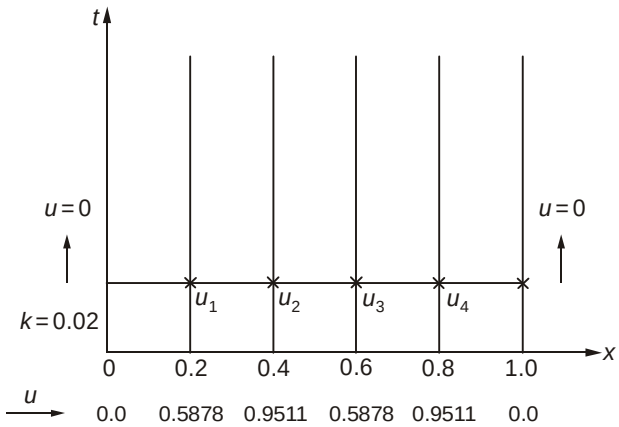


Figure 9.9

$$u_2^{(n+1)} = \frac{1}{6}[u_1^{(n+1)} + u_3^{(n)}] + \frac{2}{3}\left[0.9511 + \frac{1}{4}(0.5878 - 2\times 0.9511 + 0.9511)\right]$$

$$= \frac{1}{6}[u_1^{(n+1)} + u_3^{(n)}] + 0.5736 \qquad \text{(iii)}$$

$$u_3^{(n+1)} = \frac{1}{6}[u_2^{(n+1)} + u_4^{(n)}] + \frac{2}{3}\left[0.9511 + \frac{1}{4}(0.9511 - 2\times 0.9511 + 0.5878)\right]$$

$$= \frac{1}{6}[u_2^{(n+1)} + u_4^{(n)}] + 0.5736 \qquad \text{(iv)}$$

$$u_4^{(n+1)} = \frac{1}{6}[u_3^{(n+1)} + 0] + \frac{2}{3}\left[0.5878 + \frac{1}{4}(0.9511 - 2\times 0.5878 + 0.0)\right]$$

$$= \frac{1}{6}u_3^{(n+1)} + 0.3544 \qquad \text{(v)}$$

Equations (ii), (iii), (iv) and (v) can now be used to obtain better approximations for u_1, u_2, u_3 and u_4, respectively. The table below gives the successive iterates of u_1, u_2, u_3 and u_4 corresponding to $t = 0.02$.

x	0.0	0.2	0.4	0.6	0.8	1.0
$u(x)$	0.0	0.5878	0.9511	0.9511	0.5878	0.0
$n = 0$	0.0	0.5878	0.9511	0.9511	0.5878	0.0
$n = 1$	0.0	0.5129	0.8176	0.8078	0.4890	0.0
$n = 2$	0.0	0.4907	0.7900	0.7868	0.4855	0.0
$n = 3$	0.0	0.4861	0.7858	0.7855	0.4853	0.0
$n = 4$	0.0	0.4854	0.7854	0.7854	0.4853	0.0
$n = 5$	0.0	0.4853	0.7854	0.7854	0.4853	0.0

The symmetry of the solution about $x = 0.5$ is quite clear in the above table. The analytical solution of the problem is given by $u = e^{-\pi^2 t} \sin \pi x$ and the exact values of u for $x = 0.2$ and $x = 0.4$ are respectively 0.4825 and 0.7807. The percentage error in both the solutions is about 0.6%, and the error can be reduced by taking a finer mesh. The reader should check some of the figures given in the table.

9.7 APPLICATION OF CUBIC SPLINE

We consider again the initial boundary value problem defined by

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{9.59}$$

with the conditions

$$\left.\begin{aligned} u(x, 0) &= g(x), \quad 0 \le x \le a \\ u(0, t) &= f_1(t) \\ \text{and} \quad u(a, t) &= f_2(t) \end{aligned}\right\} \tag{9.60}$$

where $g(x)$, $f_1(t)$ and $f_2(t)$ are given functions. As earlier, we denote $\Delta x = h$, $\Delta t = l$ and $u(ih, kl) = u_i^k$. If $S_k(x)$ denotes the cubic spline approximating the function values u_i^k, then Eq. (9.59) is approximated by

$$\frac{u_i^{k+1} - u_i^k}{l} = (1 - \theta)\, M_i^k + \theta M_i^{k+1} \tag{9.61}$$

where

$$M_i^k = S_k''(x_i).$$

Using the recurrence relations for the spline second derivatives at the kth and $(k+1)$th time levels, both M_i^k and M_i^{k+1} in Eq. (9.61) can be eliminated. When this is done, we obtain the following finite difference approximation to the one-dimensional heat equation

$$\begin{aligned} &(1 - 6r\theta)\,(u_{i-1}^{k+1} + u_{i+1}^{k+1}) + (4 + 12r\theta)\, u_i^{k+1} \\ &\qquad = [1 + 6r(1 - \theta)]\,(u_{i-1}^k + u_{i+1}^k) + [4 - 12r(1 - \theta)]\, u_i^k, \end{aligned} \tag{9.62}$$

where $r = \dfrac{l}{h^2}$ and $i = 1, 2, \ldots, n - 1$.

Equation (9.62) is due to Papamichael and Whiteman. It is a general implicit representation of Eq. (9.59) and reduces to the explicit and Crank–Nicolson formulae for particular choices of θ. It can be verified that Eq. (9.62) reduces to the explicit formula for $\theta = \dfrac{1}{6r}$, whereas, the choice $\theta = \dfrac{1}{2} + \dfrac{1}{6r}$ leads to the Crank–Nicolson formula.

The following are the computational steps for solving the problem defined by Eqs. (9.59) and (9.60) using Eq. (9.62).

(i) Determine M_i^0 from the relation

$$M_{i-1}^0 + 4M_i^0 + M_{i+1}^0 = \frac{6}{h^2}(g_{i-1} - 2g_i + g_{i+1}), \quad (i = 1, 2, \ldots, n-1) \tag{9.63}$$

with

$$M_0^0 = f_1''(0) \quad \text{and} \quad M_n^0 = f_2''(0).$$

(ii) Solve the system in Eq. (9.62) for $k = 0$ with $u_0^1 = f_1(l)$ and $u_n^1 = f_2(l)$. This gives the values of $u_1^1, u_2^1, \ldots, u_{n-1}^1$.

(iii) Compute M_1^1 (i = 1, 2, …, n – 1) using the recurrence relation with $M_0^1 = f_1''(1)$ and $M_n^1 = f_2''(1)$.

At this stage, we have computed the spline solution at $t = l$. Obviously, this procedure can be repeated to compute the solution at $t = 2l$, $3l$, ….

9.8 WAVE EQUATION

The wave equation is defined by the boundary value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial t^2} \tag{9.64}$$

with the boundary conditions

$$\left.\begin{aligned} u(x, 0) &= f(x), \\ u_t(x, 0) &= \phi(x) \\ u(0, t) &= \psi_1(t) \\ u(1, t) &= \psi_2(t) \end{aligned}\right\} \tag{9.65}$$

for $0 \le t \le T$. This equation is of hyperbolic type and models the transverse vibrations of a stretched string. As earlier, we use the following difference approximations for the derivatives

$$u_{xx} = \frac{1}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) + \mathrm{O}(h^2) \tag{9.66}$$

and

$$u_{tt} = \frac{1}{l^2}\left(u_i^{k-1} - 2u_i^k + u_i^{k+1}\right) + \mathrm{O}(l^2) \tag{9.67}$$

where $x = ih$, $t = kl$, and $u(x, t) = u(ih, kl) = u_i^k$.
Further, $u_t(x, t)$ is approximated by

$$u_t(x, t) = \frac{u_i^{k+1} - u_i^{k-1}}{2l} + \mathrm{O}(l^2) \tag{9.68}$$

Substituting from Eqs. (9.66) and (9.67) in Eq. (9.64), we obtain

$$\frac{1}{l^2}\left(u_i^{k-1} - 2u_i^k + u_i^{k+1}\right) = \frac{c^2}{h^2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right).$$

Setting $\alpha = \dfrac{cl}{h}$ in the above and rearranging the terms, we have

$$u_i^{k+1} = -u_i^{k-1} + \alpha^2(u_{i-1}^k + u_{i+1}^k) + 2(1-\alpha^2)\,u_i^k \tag{9.69}$$

Equation (9.69) shows that the function values at the kth and (k – 1)th time levels are required to determine those at the $(k+1)$th time level. Such difference schemes are called *three level difference schemes* compared to the two level schemes derived in the parabolic case.* Formula (9.69) holds good if $\alpha < 1$, which is the condition for stability.

There exist implicit finite difference schemes for the equation given by Eq. (9.64). Two such schemes are

$$\frac{u_i^{k-1} - 2u_i^k + u_i^{k+1}}{l^2} = \frac{c^2}{2h^2}\left(u_{i-1}^{k-1} - 2u_i^{k-1} + u_{i+1}^{k-1} + u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right) \tag{9.70}$$

and

$$\frac{u_i^{k-1} - 2u_i^k + u_i^{k+1}}{l^2} = \frac{c^2}{4h^2}\Big[\left(u_{i-1}^{k-1} - 2u_i^{k-1} + u_{i+1}^{k-1}\right) + 2\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right)$$
$$+ \left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)\Big] \tag{9.71}$$

Equations (9.70) and (9.71) hold good for all values of $\dfrac{cl}{h}$. The use of formula given in Eq. (9.69) is demonstrated in the following examples.

Example 9.10 Solve the equation $u_{tt} = u_{xx}$ subject to the following conditions

$$u(0,\ t) = 0, \qquad u(1,\ t) = 0, \quad t > 0$$

and

$$\frac{\partial u}{\partial t}(x, 0) = 0,\ u(x, 0) = \sin^3(\pi x), \qquad 0 \le x \le 1.$$

This problem has an exact solution given by

$$u(x, t) = \frac{3}{4}\sin \pi x \cos \pi t - \frac{1}{4}\sin 3\pi x \cos 3\pi t.$$

Let $h = 0.25$ and $l = 0.2$. Then $\alpha = 0.8 < 1$. The given conditions are

$$u_0^k = 0,\ u_4^k = 0,\ u_i^0 = \sin^3(\pi i h), \qquad i = 1, 2, 3, 4.$$

Also,

$$\frac{\partial u(x, 0)}{\partial t} = 0 \Rightarrow u_i^1 - u_i^{-1} = 0$$
$$\Rightarrow u_i^{-1} = u_i^1.$$

*A three level scheme for solving parabolic equations in one dimension may be found in Sastry [1976].

With $\alpha = 0.8$, the explicit formula becomes

$$u_i^{k+1} = -u_i^{k-1} + 0.64\left(u_{i-1}^k + u_{i+1}^k\right) + 2(0.36)u_i^k.$$

Setting $k = 0$, the above gives:

$$u_i^1 = -u_i^{-1} + 0.64\left(u_{i-1}^0 + u_{i+1}^0\right) + 0.72u_i^0.$$

$$\Rightarrow u_i^1 = 0.32\left(u_{i-1}^0 + u_{i+1}^0\right) + 0.36u_i^0, \quad \text{since } u_i^{-1} = u_i^1.$$

Therefore,

$$\begin{aligned} u_1^1 &= 0.32\left(u_0^0 + u_2^0\right) + 0.36u_1^0 \\ &= 0.32\,(0+1) + 0.36(0.3537) \\ &= 0.4473\,(\text{error} = 0.0365) \end{aligned}$$

Similarly

$$u_2^1 = 0.5867\ (\text{error} = 0.0571)$$

and

$$u_3^1 = 0.4473\,(\text{error} = 0.0365)$$

The computations can be continued for $k = 1, 2, \ldots$

Example 9.11 Solve the boundary value problem defined by $u_{tt} = 4u_{xx}$ subject to the conditions.

$$u(0, t) = 0 = u(4, t),\ u_t(x, 0) = 0,\ u(x, 0) = 4x - x^2.$$

Let

$$h = 1 \text{ and } \alpha = 1 \text{ so that } l = 0.5.$$

We have

$$u_0^k = u_4^k = 0 \text{ for all } k.$$

Since

$$u_t(x, 0) = 0, \text{ we obtain } u_i^{-1} = u_i^1.$$

Further,

$$u(x, 0) = 4x - x^2$$

$$\Rightarrow u_i^0 = 4i - i^2, \text{ since } h = 1.$$

Then,

$$u_0^0 = 0,\ u_1^0 = 3,\ u_2^0 = 4,\ u_3^0 = 3 \text{ and } u_4^0 = 0.$$

For $\alpha = 1$, the explicit scheme becomes

$$u_i^{k+1} = -u_i^{k-1} + u_{i-1}^k + u_{i+1}^k \tag{i}$$

Now, for $k = 0$, Eq. (i) gives

$$u_i^1 = -u_i^{-1} + u_{i-1}^0 + u_{i+1}^0$$

$$\Rightarrow u_i^1 = \frac{1}{2}\left(u_{i-1}^0 + u_{i+1}^0\right), \quad \text{since } u_i^{-1} = u_i^1.$$

Hence

$$u_1^1 = \frac{1}{2}\left(u_0^0 + u_2^0\right) = 2, \quad u_2^1 = \frac{1}{2}(3+3) = 3, \quad u_3^1 = \frac{1}{2}(4+0) = 2.$$

Similarly, we obtain

$$u_1^2 = 0, \quad u_2^2 = 0, \quad u_3^2 = 0.$$

and the succeeding time rows can be built up.

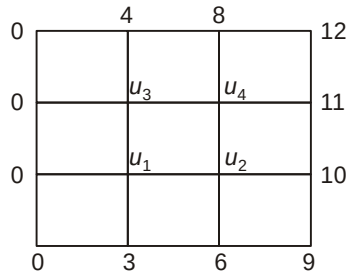
9.8.1 Software for Partial Differential Equations

Several packages are available for solving partial differential equations; but, these are often limited to the solution of simple problems involving two- and three-dimensional cases. IMSL possesses routines for solving Poisson's equation in two and three dimensions and also systems of partial differential equations in one dimension. It is well known that MATLAB has excellent display capabilities and these can be used, with advantage, for visualisation of two-dimensional spatial problems.

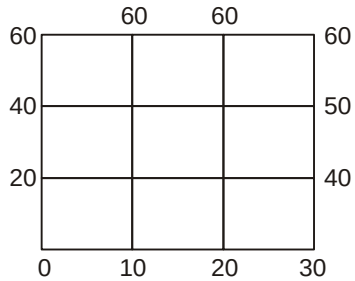
EXERCISES

The function $u(x, y)$ satisfies Laplace's equation at all points within the squares given below and has the boundary values as indicated. Compute a solution correct to two decimal places by the finite difference method (Problems 9.1–9.6).

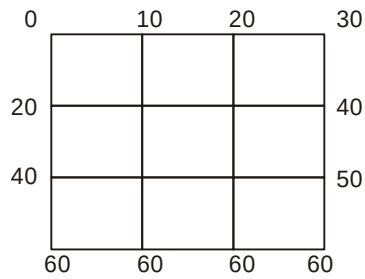
9.1



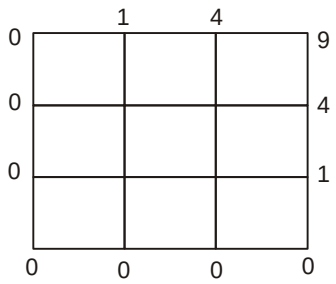
9.2



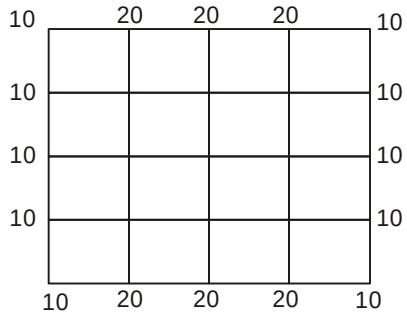
9.3



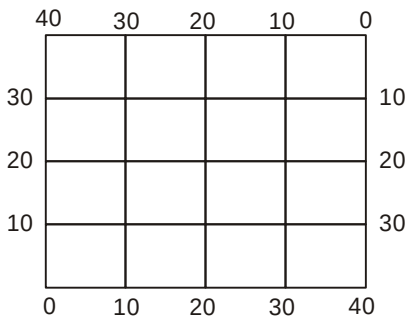
9.4



9.5



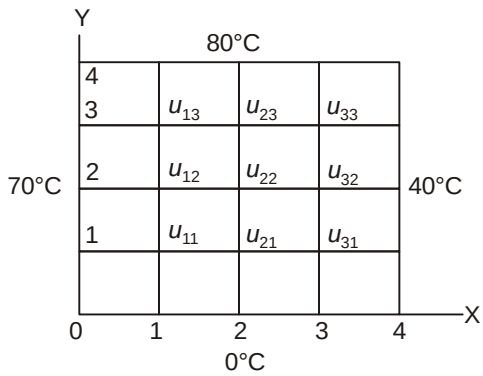
9.6



9.7 Solve

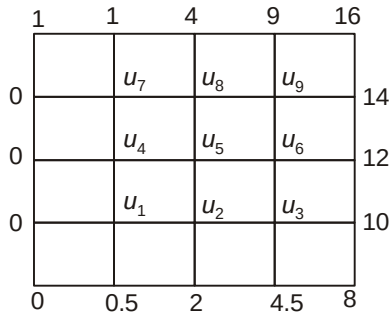
$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

for the temperature of the heated plate for the square region shown in the figure below. Use successive over relaxation with $\omega = 1.5$ and determine the temperatures at the internal mesh points upto the third iteration. Give an estimate of the per cent error in the value of u_{22}.



9.8 Solve Laplace's equation with $h = 1/3$ over the boundary of a square of unit length with $u(x, y) = 9x^2y^2$ on the boundary.

9.9 Write down the finite difference analogue of the equation $u_{xx} + u_{yy} = 0$ and solve it for the square region given below.

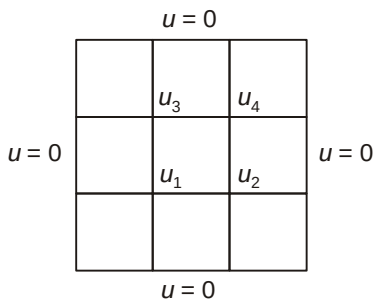


With $h = k = 1.0$, use the Gauss–Seidel method to compute, correct two decimal places, values of u at the internal mesh points.

9.10 Solve Poisson's equation

$$\nabla^2 u = 8x^2y^2$$

for the square grid shown below ($h = 1$).



9.11 Use the Bender–Schmidt formula to solve the problem

$$u_t = 5u_{xx},$$

$$u(0,\ t) = 0, \qquad u(5,\ t) = 60,$$

$$\text{and } u(x,\ 0) = \begin{cases} 20x, & 0 \le x \le 3 \\ 60, & 3 \le x \le 5 \end{cases}$$

With $h = 1$ and $l = 0.5$, compute the values of $u(x,\ 0.5)$, $u(x,\ 1.0)$ and $u(x,\ 1.5)$ for $1 \le x \le 4$.

9.12 Solve the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

with the conditions

$$u(x,\ 0) = \sin x,\ 0 \le x \le \pi$$

$$\text{and } u(0,\ t) = u(\pi,\ t) = 0,\ t > 0.$$

Compare the values obtained by Bender–Schmidt, Crank–Nicolson and Cubic Spline formulae for $u\left(\frac{\pi}{2}, \frac{\pi^2}{16}\right)$.

9.13 Use the explicit formula to solve the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

with the conditions $u(0, t) = u(5, t) = 0$ and $u(x, 0) = x^2(25 - x^2)$. With $h = 1$ and $l = 0.5$, tabulate the values of u_i^k for $i = 0, 1, 2, 3, 4, 5$ and $k = 0, 1, 2$.

9.14 Solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

subject to the conditions $u(x, 0) = 0 < u(0, t) = 0$ and $u(1, t) = t$. With $h = \frac{1}{4}$ and $l = \frac{1}{16}$, compute the value of $u\left(\frac{1}{2}, \frac{1}{8}\right)$ using Crank–Nicolson formula.

9.15 Solve the heat conduction equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

with the conditions $u(x, 0) = \sin x$, $0 \le x \le \pi$, $\frac{\partial u}{\partial t}(0, t) = e^{-t}$ and $\frac{\partial u}{\partial x}(\pi, t) = -e^{-t}$. Using Crank–Nicolson and cubic spline formulae. Taking $h = \frac{\pi}{2}$ and $l = \frac{\pi^2}{4\sqrt{20}}$, compute $u\left(\frac{\pi}{2}, \frac{\pi^2}{4\sqrt{20}}\right)$. Compare with the exact value.

9.16 Derive a cubic spline finite difference formula for the solution of the equation

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial u}{\partial x} + cu \quad (a,\ b,\ c \text{ being constants})$$

with the conditions $u(x, 0) = g(x)$, $0 \le x \le a$ and $\frac{\partial u}{\partial t}(0, t) = f_1(t)$ and $\frac{\partial u}{\partial t}(a, t) = f_2(t)$.

9.17 Du Fort–Frankel formula: The formula

$$u_i^{k+1} = \frac{1-2r}{1+2r}u_i^{k-1} + \frac{2r}{1+2r}\left(u_{i-1}^k + u_{i+1}^k\right)$$

where $r = \frac{l}{h^2}$, is known as Du Fort–Frankel formula for the solution of the one-dimensional heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$. It is a three level

formula since we require function values on $(k-1)$th and kth time levels in order to compute those on the $(k+1)$th time level. Solve the equation given above with the conditions $u(x, 0) = \sin x$, $0 \le x \le \pi$. and $u(0, t) = u(\pi, t) = 0$

using both the explicit and Du Fort–Frankel formulae with $h = \frac{\pi}{4}$ and $l = \frac{\pi^2}{32}$. Estimate the value of $u\left(\frac{\pi}{2}, \frac{\pi^2}{16}\right)$ and compare with the exact value.

9.18 Order of Accuracy of Finite Difference Formulae: The explicit formula

$$u_i^{k+1} = r\left(u_{i-1}^k + u_{i+1}^k\right) + (1-2r)\, u_i^k$$

is an approximation to the one-dimensional heat equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$. Expanding both sides of this formula by Taylor's series, we obtain

$$u_i^{k+1} - r\left(u_{i-1}^k + u_{i+1}^k\right) - (1-2r)u_i^k = \left(\frac{l^2}{2} - \frac{lh^2}{12}\right)\left[\frac{\partial^2 u}{\partial t^2}\right]_{i,k} + \cdots$$

In such a case, we say that the *local order of accuracy* of the explicit formula is $O(l + h^2)$. This is also called the *local truncation error* of the formula. Show that the local truncation error of the Crank–Nicolson formula is $O(l^2 + h^2)$.

9.19 Show that the local truncation error of formula given in Eq. (9.69) is given by

$$\frac{k^2h^2}{12}\left(\alpha^2 - 1\right)\frac{\partial^4 u}{\partial x^4} + \frac{k^2h^4}{360}\left(\alpha^4 - 1\right)\frac{\partial^6 u}{\partial x^6} + \cdots$$

9.20 Solve the equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

with the conditions $u(0, t) = u(1, t) = 0$, $u(x, 0) = \sin \pi x$, and $\frac{\partial u}{\partial t}(x, 0) = 0$. Taking $h = 0.25$ and $l = 0.2$, compute $u(0.5, 0.4)$ in two time steps and compare your result with the exact solution given by $u(x, t) = \sin \pi x \cos \pi t$.

Answers to Exercises

9.1 $u_1 = 3.33, \quad u_2 = 6.67, \quad u_3 = 3.67, \quad u_4 = 7.33.$

9.2 $u_1 = 26.7, \quad u_2 = 33.3, \quad u_3 = 43.3, \quad u_4 = 46.7.$

9.3 $u_1^{(5)} = 43.33, \quad u_2^{(5)} = 46.66, \quad u_3^{(5)} = 26.66, \quad u_4^{(5)} = 33.33.$

9.4 $u_1^{(6)} = 0.498, \quad u_2^{(6)} = 0.9998, \quad u_3^{(6)} = 0.9998, \quad u_4^{(6)} = 2.4999.$

9.5 12th iteration values

$u_1 = 15, \quad u_2 = 16.25, \quad u_3 = 15$

$u_4 = 13.75, \quad u_5 = 15, \quad u_6 = 13.75$

$u_7 = 15, \quad u_8 = 16.25, \quad u_9 = 15.$

9.6 $u_1 = 15, \quad u_2 = 20, \quad u_3 = 25, \quad u_4 = 20,$

$u_5 = 20, \quad u_6 = 20, \quad u_7 = 25, \quad u_8 = 20, \quad u_9 = 15.$

9.7 $u_{11}^{(1)} = 26.25, \quad u_{21}^{(1)} = 9.84375, \quad u_{31}^{(1)} = 10.691406$

$u_{12}^{(1)} = 36.09375, \quad u_{22}^{(1)} = 17.226562, \quad u_{32}^{(1)} = 25.469238$

$u_{13}^{(1)} = 69.785156, \quad u_{23}^{(1)} = 62.629394, \quad u_{33}^{(1)} = 78.036987.$

Second iteration values: Error in $u_{22}^{(2)} = 65\%$

Third iteration: Error in $u_{22}^{(3)} = 13\%$.

9.8 See Problem 9.4.

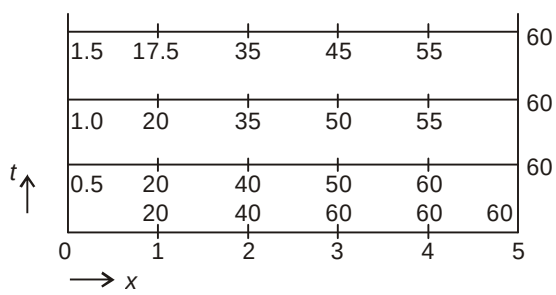
9.9 $u_1 = 1.57, u_2 = 3.70, \quad u_3 = 6.57,$

$u_4 = 2.06, u_5 = 4.69, \quad u_6 = 8.06,$

$u_7 = 2.09, u_8 = 4.92 \quad u_9 = 9.00.$

9.10 $u_{11} = -13, \quad u_{21} = u_{12} = -22, \quad u_{22} = -43.$

9.11



9.12 (a) Error = 0.03964,
(b) Error = 0.00728,
(c) Error = 0.00728.

9.13

$k \backslash i$	0	1	2	3	4	5
0	0	24.0	84.0	144.0	144.0	0
1	0	42.0	78.0	78.0	57.0	0
2	0	39.0	60.0	67.5	39.0	0

9.14 $u_1^2 = 0.005899, \quad u_2^2 = 0.019132, \quad u_3^2 = 0.052771.$
Exact values are

$$u\left(\frac{1}{4}, \frac{1}{8}\right) = 0.00541, \quad u\left(\frac{1}{2}, \frac{1}{8}\right) = 0.01878, \quad u\left(\frac{3}{4}, \frac{1}{8}\right) = 0.05240.$$

9.15 Exact solution is

$$u(x, t) = e^{-t} \sin x.$$

$$u_1^1 = 0.659726 \text{ (error = 0.0837)}$$

Spline solution

$$u_1^1 = 0.6415 \quad \text{(error = 0.0655)}$$

9.17 Exact solution is

$$u\left(\frac{\pi}{2}, t\right) = e^{-t}$$

Explicit formula:

$$k = 0\text{: Error in } u_4^1 = 0.0011,$$

$$k = 1\text{: Error in } u_4^2 = 0.0020,$$

$$k = 2\text{: Error in } u_4^3 = 0.0028.$$

Du Fort–Frankel formula

$$\text{Error in } u_4^2 = 0.0016$$

9.20 $u_2^2 \approx u(0.5, 0.4) = 0.320469,$
Exact $u(0.5, 0.4) = 0.309017.$

Chapter 10

Numerical Solution of Integral Equations

10.1 INTRODUCTION

Any equation in which the unknown function appears under the integral sign is known as an *integral equation.* Such equations arise in the formulation of physical problems. A few examples are given below.

(a) *Abel's Integral Equation*

$$f(x) = \int_0^x K(x, t)\phi(t)\,dt \tag{10.1}$$

This equation arises in the problem of finding the path of a particle which is constrained to move under gravity in a vertical plane. In Eq. (10.1), $\phi(t)$ is the unknown function, whereas, $f(x)$ and $K(x, t)$ are given functions.

(b) *Love's Equation*

$$y(x) + \int_{-1}^{1} K(x, s)\, y(s)\, ds = 1 \tag{10.2}$$

This equation occurs in the problem of determining the capacity of a circular plate condenser. In this case, the unknown function $y(x)$ appears both outside and inside the integral sign.

(c) *Vandrey's Equation*

$$v(s) = \phi(s) - \frac{\lambda}{\pi}\int_0^L K(s, \sigma)v(\sigma)\, d\sigma, \quad 0 \le s \le L \tag{10.3}$$

This equation occurs in fluid dynamics while calculating the pressure distribution on the surface of a body of revolution moving in a fluid.

Integral equations are classified into two main types. *Volterra* integral equations and *Fredholm* integral equations. Again, there are integral equations of the first and second kinds. If the unknown function appears outside the integral sign also, then it is called an integral equation of the second kind. Equations (10.2) and (10.3) are Fredholm integral equations of the second kind. On the other hand, Eq. (10.1), where the unknown function appears only under the integral sign, is called Volterra integral equation of the first kind.

If $\phi(s) = 0$ in Eq. (10.3), then the equation is said to be *homogeneous;* otherwise it is *non-homogeneous.* For non-homogeneous equations, λ is a numerical parameter whereas for homogeneous equations, it is an *eigenvalue problem* in which the objective is to determine those values of λ, called the eigenvalues, for which the integral equation possesses nontrival solutions called *eigenfunctions.*

In Eq. (10.1), $k(x, t)$ is called the *kernel* of the integral equation. If the kernel is bounded and continuous, then the integral equation is said to be nonsingular. If the range of integration is infinite, or if the kernel violates the above conditions, then the integral equation is said to be *singular*.

A solution of the integral equation is a function which, when substituted into the equation, reduces it to an identity.

Example 10.1 Show that the function $y(t) = 1 - t$ is a solution of the integral equation

$$t = \int_0^t e^{(t-u)} y(u)\, du$$

We have

$$\begin{aligned}\int_0^t e^{(t-u)} y(u)\, du &= \int_0^t e^{(t-u)}\,(1-u)\, du \\ &= e^t \int_0^t e^{-u}(1-u)\, du = e^t\left[-e^{-u}(1-u) + \int e^{-u}(-1) du\right]_0^t \\ &= t, \text{ on simplification.}\end{aligned}$$

Since integral equations occur in physical problems, it is usually difficult to solve them analytically and hence it would be necessary to adopt a numerical method of solution. Before presenting these methods, it would be instructive to demonstrate the relationship between integral equations and initial value problems. This is shown in the following examples.

Example 10.2 Transform the initial-value problem

$$y'' + y = 0, \qquad \text{with } y(0) = 0 \text{ and } y'(0) = 1 \tag{i}$$

into an equivalent integral equation.

Let

$$\frac{d^2y}{dx^2} = u(x) \tag{ii}$$

Integrating both sides of Eq. (ii) with respect to x, we obtain

$$\frac{dy}{dx} = \int_0^x u(x)\,dx + y'(0) = \int_0^x u(x)\,dx + 1,$$

on using the given condition. Integrating the above with respect to x, we get

$$y(x) = \int_0^x (x-t)\,u(t)\,dt + x \tag{iii}$$

Substituting Eqs. (ii) and (iii) in Eq. (i), we obtain

$$u(x) + \int_0^x (x-t)\,u(t)\,dt + x = 0$$

or

$$u(x) = -x + \int_0^x (t-x)\,u(t)\,dt,$$

which is a Volterra integral equation.

In deriving Eq. (iii), we have used the formula

$$\int_a^x\int_a^x \cdots \int_a^x f(x)\underbrace{dx\,dx\ldots dx}_{n \text{ times}} = \frac{1}{(n-1)!}\int_a^x (x-t)^{n-1} f(t)\,dt \tag{iv}$$

Example 10.3 Transform the integral equation

$$y(x) = x + \int_0^x [5 - 6(x-t)]\,y(t)\,dt$$

into an equivalent initial value problem.

We have

$$y(x) = x + \int_0^x [5 - 6(x-t)]\,y(t)\;dt \tag{i}$$

Clearly, $y(0) = 0$.

We use the following formula for differentiation under an integral sign

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)\,dt = \frac{db}{dx} f(x,b) - \frac{da}{dx} f(x,a) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt. \qquad \text{(ii)}$$

Using Eq. (ii) and differentiating both sides of Eq. (i), we obtain

$$y'(x) = 1 + 5y(x) + \int_0^x -6y(t)\,dt$$

$$= 1 + 5y(x) - 6\int_0^x y(t)\,dt \qquad \text{(iii)}$$

Therefore,

$$y'(0) = 1, \quad \text{since } y(0) = 0.$$

Differentiating Eq. (iii), we obtain

$$y''(x) = 5y'(x) - 6y(x), \quad \text{on using Eq. (ii)}$$

Hence we have

$$y''(x) - 5y'(x) + 6y(x) = 0$$

which is the required differential equation satisfying the conditions $y(0) = 0$ and $y'(0) = 1$.

10.2 NUMERICAL METHODS FOR FREDHOLM EQUATIONS

There exist several methods for the numerical solution of Fredholm integral equations of the second kind, e.g. method of degenerate kernels, method of successive approximations, collocation and product-integration methods, etc. We present a few of these methods, in a formal way, with simple examples. For error estimates and other details, the reader is referred to Atkinson [1967].

10.2.1 Method of Degenerate Kernels

We consider the integral equation

$$f(x) - \int_a^b K(x,t)\,f(t)\,dt = \phi x \qquad (10.4)$$

A kernel $K(x, t)$ is said to be *degenerate* if it can be expressed in the form

$$K(x,t) = \sum_{i=1}^{n} u_i(x)\,v_i(t) \qquad (10.5)$$

Substituting this in Eq. (10.4), We obtain

$$f(x) - \sum_{i=1}^{n} \int_a^b u_i(x)\, v_i(t)\; f(t)\, dt = \phi(x) \tag{10.6}$$

Setting

$$\int_a^b v_i\,(t)\; f(t)\, dt = A_i \tag{10.7}$$

Eq. (10.6) gives

$$f(x) = \sum_{i=1}^{n} A_i\, u_i(x) + \phi(x) \tag{10.8}$$

The constants A_i are still to be detemined, but substituting from Eq. (10.8) in Eq. (10.7), we get

$$\int_a^b v_i(t) \left[\sum_{j=1}^{n} A_j u_j(t) + \phi(t) \right] dt = A_i$$

or

$$\sum^{n} A_j \int_a^b v_i(t)\, u_j(t)\, dt + \int_a^b v_i(t)\, \phi(t)\, dt = A_i, \tag{10.9}$$

which represents a system of n equations in the n unknowns A_1, $A_2, \ldots,$ A_n. When the A_i are determined, Eq. (10.8) then gives $f(x)$.

Although the method is important in the theory of integral equations, it does not seem to be much useful in the numerical work, since the kernel is unlikely to have the simple form (10.5) in practical problems. In general, however, it is possible to take a partial sum of Taylor's series for the kernel. This is shown in Example 10.5.

Example 10.4 We consider the equation

$$f(x) - \lambda \int_0^{\pi/2} \sin x \cos t\; f(t)\, dt = \sin x.$$

Setting

$$\int_0^{\pi/2} \cos t\, f(t)\, dt = A, \tag{i}$$

the integral equation becomes

$$f(x) = \lambda A \sin x + \sin x = (\lambda A + 1) \sin x.$$

Substituting this in (i), we obtain

$$\int_0^{\pi/2} \cos t(\lambda A+1)\sin t\, dt = A,$$

which gives on simplification

$$A = \frac{1}{2-\lambda}.$$

Hence the solution of the integral equation is given by

$$f(x) = \frac{2}{2-\lambda}\sin x \qquad (\lambda \neq 2).$$

By direct substitution the reader should verify that this is the solution of the given integral equation.

Example 10.5 Solve the integral equation

$$f(x) = \frac{1}{2}(e^{-x}+3x-1) + \int_0^1 (e^{-xt^2}-1)x\, f(t)\, dt.$$

We have

$$\begin{aligned} K(x,t) &= (e^{-xt^2}-1)x \\ &= \left(1 - xt^2 + \frac{x^2t^4}{2} + \cdots - 1\right)x \\ &= -x^2t^2 + \frac{1}{2}x^3t^4, \end{aligned}$$

neglecting the other terms of the Taylor's series.

Hence the given integral equation becomes

$$\begin{aligned} f(x) &= \frac{1}{2}(e^{-x}+3x-1) + \int_0^1 (-x^2t^2 + \frac{1}{2}x^3t^4)\, f(t)\, dt \\ &= \frac{1}{2}(e^{-x}+3x-1) - x^2\int_0^1 t^2 f(t)dt + \frac{1}{2}x^3\int_0^1 t^4 f(t)\, dt \\ &= \frac{1}{2}(e^{-x}+3x-1) - k_1x^2 + \frac{1}{2}k_2x^3, \qquad \text{(i)} \end{aligned}$$

where

$$k_1 = \int_0^1 t^2 f(t)\, dt \qquad \text{(ii)}$$

and

$$k_2 = \int_0^1 t^4 f(t)\,dt. \tag{iii}$$

Substituting for $f(t)$ from (i) in (ii), we obtain

$$k_1 = \int_0^1 t^2 \left[\frac{1}{2}(e^{-t} + 3t - 1) - k_1 t^2 + \frac{1}{2}k_2 t^3\right] dt \tag{iv}$$

Since

$$\int_0^1 t^2 e^{-t} dt = 2 - \frac{5}{e},$$

Equation (iv) gives

$$\frac{6k_1}{5} - \frac{k_2}{12} = -\frac{5}{2e} + \frac{29}{24}. \tag{v}$$

Similarly, substituting for $f(t)$ in (iii) and simplifying, we obtain

$$\frac{k_1}{7} + \frac{15}{16}k_2 = -\frac{65}{2e} + \frac{243}{20}. \tag{vi}$$

Solution of (v) and (vi) is given by

$$k_1 = 0.2522 \quad \text{and} \quad k_2 = 0.1685.$$

Hence the solution of the given integral equation is

$$f(x) = \frac{1}{2}(e^{-x} + 3x - 1) - 0.2522x^2 + \frac{1}{2}(0.1685)x^3.$$

10.2.2 Method of Successive Approximations

Let the integral equation be given by

$$y(x) = f(x) + \lambda \int_a^b K(x, t)\, y(t)\, dt \tag{10.10}$$

where $f(x)$ is continuous in (a, b) and the kernel $K(x, t)$ is finite. We first approximate $y(x)$ in the integral in Eq. (10.10) by $y^{(0)}(x)$, and then determine $y^{(1)}(x)$ from the relation

$$y^{(1)}(x) = f(x) + \lambda \int_a^b K(x, t)\, y^{(0)}(t)\, dt \tag{10.11}$$

We, next, determine $y^{(2)}(x)$ from the relation

$$y^{(2)}(x) = f(x) + \lambda \int_a^b K(x, t)\, y^{(1)}(t)\, dt \tag{10.12}$$

Proceeding in this way, we construct a series of approximations $y^{(0)}(x)$, $y^{(1)}(x)$, $y^{(2)}(x)$, …, $y^{(n)}(x)$ such that

$$y^{(n)}(x) = f(x) + \lambda \int_a^b K(x,t)\, y^{(n-1)}(t)\, dt \tag{10.13}$$

Does the sequence of approximations $y^0(x)$, $y^{(1)}(x)$, …, $y^{(n)}(x)$, converge to the exact solution of Eq. (10.10)? The answer is yes, provided that certain conditions are satisfied. We state, without proof, that the sequence $y^{(0)}(x)$, $y^{(1)}(x)$, $y^{(2)}(x)$, … converges to the solution of Eq. (10.10), provided that

$$|\lambda| < \frac{1}{P} \quad \text{where } P = \left\{ \int_a^b \int_a^b [K(x,t)]^2\, dx\, dt \right\}^{1/2} \tag{10.14}$$

Example 10.6 Solve the integral equation

$$y(x) = x + \int_{-1}^{1} xt\; y(t)\, dt$$

by the method of successive approximations.
Here $\lambda = 1$ and $K(x, t) = xt$
Therefore,

$$P^2 = \int_{-1}^{1}\int_{-1}^{1} x^2 t^2\, dx\, dt = \frac{4}{9}$$

Hence $P = \dfrac{2}{3}$ and conditions (10.14) are satisfied.

With $y^{(0)}(x) = 1$, we obtain successively

$$y^{(1)}(x) = x + \int_{-1}^{1} xt \cdot dt = x, \qquad \text{since } \int_{-1}^{1} t\, dt = 0$$

$$y^{(2)}(x) = x + \int_{-1}^{1} xt \cdot t\, dt = x + x\left(\frac{2}{3}\right) = x\left(1 + \frac{2}{3}\right)$$

$$y^{(3)}(x) = x + \int_{-1}^{1} xt \cdot t\left(1 + \frac{2}{3}\right) dt$$

$$= x + x\left(1 + \frac{2}{3}\right)\frac{2}{3} = x\left[1 + \frac{2}{3} + \left(\frac{2}{3}\right)^2\right].$$

$$y^{(4)}(x) = x + \int_{-1}^{1} \left[xt \cdot t \left(1 + \frac{2}{3} + \frac{4}{9} \right) \right] dt$$

$$= x \left[1 + \frac{2}{3} + \left(\frac{2}{3} \right)^2 + \left(\frac{2}{3} \right)^3 \right]$$

Hence we obtain

$$y(x) = x \left[1 + \left(\frac{2}{3} \right) + \left(\frac{2}{3} \right)^2 + \left(\frac{2}{3} \right)^3 + \cdots \right]$$

$$= 3x.$$

It can be verified that $y = 3x$ is the exact solution of the given integral equation.

10.2.3 Quadrature Methods

We consider the integral equation in the form

$$f(x) - \int_a^b K(x, t)\, f(t)\, dt = \phi(x). \tag{10.15}$$

Since a definite integral can be closely approximated by a quadrature formula, we approximate the integral term in Eq. (10.15) by a formula of the form

$$\int_a^b F(x)\, dx = \sum_{m=1}^{n} A_m\, F(x_m), \tag{10.16}$$

where A_m and x_m are the weights and abscissae, respectively. Consequently, Eq. (10.15) can be written as

$$f(x) - \sum_{m=1}^{n} A_m K(x, t_m)\, f(t_m) = \phi(x), \tag{10.17}$$

where $t_1, t_2, \ldots, t_n$ are points in which the interval (a, b) is subdivided. Further, Eq. (10.17) must hold for all values of x in the interval (a, b); in particular, it must hold for $x = t_1, x = t_2, \ldots, x = t_n$. Hence we obtain

$$f(t_i) - \sum_{m=1}^{n} A_m K(t_i, t_m)\, f(t_m) = \phi(t_i), \quad i = 1, 2, \ldots, n. \tag{10.18}$$

which is a system of n linear equations in the n unknowns $f(t_1)$, $f(t_2)$, $\ldots$, $f(t_n)$. When the $f(t_i)$ are determined, Eq. (10.17) gives an approximation

for $f(x)$. Obviously, different types of quadrature formulae can be employed, and the following examples demonstrate the use of trapezoidal and Simpson's rules.

Example 10.7 Solve

$$f(x) - \int_0^1 (x+t)\, f(t)\, dt = \frac{3}{2}x - \frac{5}{6} \tag{i}$$

By direct substitution, it can be verified that the analytical solution is given by $f(x) = x - 1$. For the numerical solution, we divide the range [0, 1] into two equal subintervals so that $h = 1/2$. Applying the trapezoidal rule to approximate the integral term in (i), we obtain

$$f(x) - \frac{1}{4}\left[x\, f_0 + 2\left(x + \frac{1}{2}\right) f_1 + (x+1)\, f_2\right] = \frac{3}{2}x - \frac{5}{6}, \qquad \text{where } f_i = f(x_i).$$

Setting $x = t_i$, where $t_0 = 0$, $t_1 = 1/2$ and $t_2 = 1$, this gives the system of equations

$$12 f_0 - 3 f_1 - 3 f_2 = -10$$

$$-3 f_0 + 12 f_1 - 9 f_2 = -2$$

$$-3 f_0 - 9 f_1 + 6 f_2 = 8$$

The solution is

$$f_0 = -\frac{1}{2}, \quad f_1 = -\frac{5}{6}, \quad f_2 = \frac{1}{2}.$$

On the other hand, if we use Simpson's rule to approximate the integral term in (i), we obtain

$$f(x) - \frac{1}{6}\left[x\, f_0 + 4\left(x + \frac{1}{2}\right) f_1 + (x+1)\, f_2\right] = \frac{3}{2}x - \frac{5}{6} \tag{ii}$$

Setting $x = t_i$, we get

$$6 f_0 - 2 f_1 - f_2 = -5$$

$$-f_0 + 4 f_1 - 3 f_2 = -1$$

$$-f_0 - 6 f_1 + 4 f_2 = 4.$$

The solution of which is

$$f_0 = -1, \quad f_1 = -\frac{1}{2}, \quad f_2 = 0.$$

Using these values in (ii), we get

$$f(x) = \frac{1}{6}\left[-x + 4\left(x + \frac{1}{2}\right)\left(-\frac{1}{2}\right)\right] + \frac{3}{2}x - \frac{5}{6}$$

$$= x - 1, \text{ which is the exact solution.}$$

It should be noted that Simpson's rule gives exact result in this case since the integrand is a second-degree polynomial in t.

Example 10.8 The integral equation

$$y(x) + \int_{-1}^{1} K(x, s)\, y(s)\, ds = 1, \tag{i}$$

where

$$K(x, s) = \frac{1}{\pi} \frac{1}{1 + (x - s)^2} \tag{ii}$$

occurs in an electrostatics problem considered by Love [1949], and is called *Love's* equation. The analytical method of solution, suggested by Love, is somewhat laborious and various numerical methods were proposed. The simplest is to approximate the inegral term in (i) by the trapezoidal rule. For this we divide the interval (–1, 1) into n smaller intervals of width h, the ith point of subdivision being denoted by s_i, such that

$$s_i = -1 + ih, \qquad i = 0, 1, 2, \ldots, n$$

and $nh = 2$. Denoting $y(x_i)$ by y_i, Eq. (i) gives

$$y_i + \sum_{j=0}^{n-1} \int_{s_j}^{s_{j+1}} K(x_i, s)\; y(s)\, ds = 1.$$

Approximating the integral term by the trapezoidal rule, the above equation becomes

$$y_i + \sum_{j=0}^{n-1} \frac{h}{2}\Big[\, K(x_i, s_j) y(s_j) + K(x_i, s_{j+1}) y(s_{j+1}) \Big] = 1,$$

which can be rewritten as:

$$y_i + \frac{h}{2} K(x_i, s_0) y_0 + \frac{h}{2} K(x_i, s_n) y_n + h \sum_{j=1}^{n-1} K(x_i, s_j) y_j = 1 \tag{iii}$$

for $i = 0, 1, 2, \ldots, n$. Equation (iii) represents a system of $(n + 1)$ linear equations in $(n + 1)$ unknowns, viz., $y_0, y_1, \ldots, y_n$, and was solved on a

digital computer. The solution is *symmetric* and the computed values of $y(x)$ at $x = 0$ and $x = 1$ are given in the table below. For comparison, the exact values are also tabulated. To study the order of convergence of the method, computations were made with different values of n. The h^2-order of convergence of the trapezoidal rule is quite revealing.

x	*Exact y(x)*	*n*	*Computed y(x)*	*Error*	*Ratio*
0.0	0.65741	4	0.66026	0.00285	
		8	0.65812	0.00071	4
		16	0.65759	0.00018	4
		32	0.65746	0.00005	3.6
1.0	0.75572	4	0.75452	0.00120	
		8	0.75542	0.00030	4
		16	0.75564	0.00008	3.75
		32	0.75570	0.00002	4

10.2.4 Use of Chebyshev Series

We consider the Fredholm integral equation in the form

$$y(x) + \int_{-1}^{1} K(x, s)\, y(s)\, ds = f(x), \qquad (-1 \le x \le 1). \tag{10.19}$$

we write

$$y(x) = \sum_{r=0}^{N} a_r T_r(x) \tag{10.20}$$

and

$$f(x) = \sum_{r=0}^{N} f_r T_r(x) \tag{10.21}$$

Where $T_r(x)$ is the Chebyshev polynomial of degree r. Substituting Eqs. (10.20) and (10.21) in Eq. (10.19), and interchanging the order of integration and summation in the integral term, we obtain

$$\sum_{r=0}^{N} a_r T_r(x) + \sum_{j=0}^{N} a_j \int_{-1}^{1} K(x, s)\, T_j(s)\, ds = \sum_{r=0}^{N} f_r T_r(x). \tag{10.22}$$

Let us now assume that

$$\int_{-1}^{1} K(x, s) T_j(s)\, ds = \sum_{r=0}^{N} b_{jr} T_r(x). \tag{10.23}$$

Then, we can equate corresponding coefficients of $T_r(x)$ on both sides of Eq. (10.22). This gives a system of $(N+1)$ equations in $(N+1)$ unknowns a_r:

$$a_r + \sum_{j=0}^{N} a_j b_{jr} = f_r, \qquad r = 0, 1, 2, \ldots, N. \tag{10.24}$$

When the a_r are known, Eq. (10.20) provides the solution as a Chebyshev series. This method is due to Elliott [1963], but a variant of this method, due to El-Gendi [1969], gives better accuracy and is described below.
We consider the numerical quadrature of the definite integral

$$I = \int_{-1}^{1} f(x)\, dx, \tag{10.25}$$

where $f(x)$ is defined and well-behaved in $[-1, 1]$. We can write

$$f(x) = \sum_{r=0}^{N}{}'' a_r T_r(x), \tag{10.26}$$

where

$$a_r = \frac{2}{N} \sum_{j=0}^{N}{}'' f\left(\cos\frac{j\pi}{N}\right)\cos\frac{rj\pi}{N}. \tag{10.27}$$

Substituting Eqs. (10.26) and (10.27) in Eq. (10.25) and simplifying using the well-known relation (see Jain [1971]):

$$\int_{-1}^{1} T_{2j}(x)\, dx = \frac{2}{1-4j^2}, \tag{10.28}$$

we can write

$$\int_{-1}^{1} f(x)\, dx = \sum_{s=0}^{N} p_{Ns} f_s, \tag{10.29}$$

where for even N,

$$p_{Ns} = \frac{4}{N} \sum_{j=0}^{N/2}{}'' \frac{1}{1-4j^2}\cos\frac{2j\pi s}{N}, \qquad s = 1, 2, \ldots, N-1 \tag{10.30a}$$

and

$$p_{N,0} = p_{N,N} = \frac{1}{N^2-1} \tag{10.30b}$$

The integral term in Eq. (10.19) can also be approximated in the sameway to obtain the system of equations

$$[I + A][y] = [f], \tag{10.31}$$

where

$$a_{ij} = p_{Nj} K_{ij} \tag{10.32}$$

$$K_{ij} = K\left(-\cos\frac{i\pi}{N}, -\cos\frac{j\pi}{N}\right), \quad i, j = 0, 1, \ldots, N. \tag{10.33}$$

$$y_i = y\left(-\cos\frac{i\pi}{N}\right), \tag{10.34}$$

p_{Nj} being given by Eq. (10.30). The system (10.31) can now be solved to obtain directly the values of y (see Sastry [1975]).

Example 10.9 The integral equation

$$y(x) + \int_{-1}^{1} K(x, s)\, y(s)\, ds = 1 \tag{i}$$

where

$$K(x, s) = \frac{1}{\pi} \frac{d}{d^2 + (x - s)^2} \tag{ii}$$

and d is a positive real number, occurs in the problem of determining the capacity of a circular plate condenser and was considered by Love [1949]. He showed, by analytical methods, that there exists a unique, continuous, real and even solution, and that it can be expressed as a convergent series of the form

$$y(x) = 1 + \sum_{n=1}^{\infty} (-1)^n \int_{-1}^{1} K_n(x, s)\, ds, \tag{iii}$$

where the iterated kernels $K_n(x, s)$ are given by

$$\left.\begin{aligned} K_1(x, s) &= \frac{d}{\pi[d^2 + (x - s)^2]} \\ K_n(x, s) &= \int_{-1}^{1} K_{n-1}(x, t) K_1(t, s)\, dt. \end{aligned}\right\} \tag{iv}$$

This method of solution is somewhat laborious, and numerical solutions to this problem were found by several authors, e.g. Fox and Goodwin [1953], Young [1954], Elliott [1963], Wolfe [1969], El-Gendi [1969] and Phillips [1972]. All these authors, excepting Phillips, investigated the problem only for the case $d = 1.0$. Results for $d = 1.0$, using the trapezoidal rules were already given in Example 10.8. This method is unsuitable for smaller values of d. Thus, for example with 32 subdivisions and $d = 0.001$, the value obtained for $x = 0$ is 0.04782 compared to the true value 0.50015.

The table below summarizes the results obtained by the Chebyshev series method with $N = 8$ and $d = 1.0$, and it is clear that this method gives better accuracy than the trapezoidal method. However, this method too gives inaccurate results for smaller values of d.

Chebyshev Series Solution of Love's Equation with $d = 1.0$

$x_j = -\cos(j\pi/N)$	$y(x_j)$
0.0	0.65740981
0.38268	0.67248912
0.70711	0.70866017
0.92388	0.74265684
1.0	0.75571801

10.2.5 Cubic Spline Method

We know that in the interval $x_{j-1} \le x \le x_j$, $s(x)$ is given by

$$s(x) = M_{j-1}\frac{(x_j - x)^3}{6h} + M_j\frac{(x - x_{j-1})^3}{6h}$$

$$+\left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)\frac{(x_j - x)}{h} + \left(y_j - \frac{h^2}{6}M_j\right)\frac{(x - x_{j-1})}{h} \qquad (10.35)$$

where $M_j = s''(x_j)$, $y_j = y(x_j)$, and $x_j = x_0 + jh$, $j = 0, 1, \ldots, N$. If we now approximate the integral term in Eq. (10.19) by using Eq. (10.35), we obtain

$$y(x_i) + \sum_{j=1}^{N}\int_{s_{j-1}}^{s_j} K(x, s)\left[M_{j-1}\frac{(s_j - s)^3}{6h} + M_j\frac{(s - s_{j-1})^3}{6h}\right.$$

$$\left. + \left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)\frac{(s_j - s)}{h} + \left(y_j - \frac{h^2}{6}M_j\right)\frac{(s - s_{j-1})}{h}\right]ds$$

$$= f(x_i), \quad i = 0, 1, 2, \ldots, N \qquad (10.36)$$

Putting $s = s_{j-1} + ph$, the above equation simplifies to

$$y(x_i) + h\sum_{j=1}^{N} \int_0^1 K(x_i, s_{j-1} + ph)\left[M_{j-1}\frac{(1-p)^3 h^2}{6} + M_j \frac{p^3 h^2}{6}\right.$$

$$\left. + \left(y_{j-1} - \frac{h^2}{6}M_{j-1}\right)(1-p) + \left(y_j - \frac{h^2}{6}M_j\right)p\right]dp$$

$$= f(x_i), \quad i = 0, 1, 2, \ldots, N \tag{10.37}$$

In Eq. (10.37), the integrals

$$\int_0^1 K\left(x_i, s_{j-1} + ph\right) p^m \, dp, \qquad m = 0, 1, 2 \text{ and } 3, \tag{10.38}$$

have to be evaluated. This can be done either analytically (wherever possible) or alternatively, by numerical techniques. When these integrals are evaluated, Eq. (10.37) together with the relations

$$\left.\begin{aligned} &\frac{h}{6}M_{j-1} + \frac{2h}{3}M_j + \frac{h}{6}M_{j+1} = \frac{y_{j-1} - 2y_j + y_{j+1}}{h} \\ &\qquad\qquad j = 1, 2, \ldots, N-1 \\ \text{and} \qquad &M_0 = M_N = 0 \end{aligned}\right\} \tag{10.39}$$

will form a set of $(2N + 2)$ linear algebraic equations in $(2N + 2)$ unknowns, viz., $y_0, y_1, \ldots, y_N$, $M_0, M_1, \ldots, M_N$. As an example, we consider again Love's equation given in the previous example.

Example 10.10 In contrast with the previous methods, the spline method can be applied when the values of d are small. For this particular example, the integrals in Eq. (10.38) were calculated analytically. Thus for $m = 0$, we have

$$X_0 = \int_0^1 K(x_i, s_{j-1} + ph)\, dp = \frac{1}{\pi}\int_0^1 \frac{d}{d^2 + (x_i - s_{j-1} - ph)^2}\, dp$$

Putting $x_i = -1 + ih$ and $s_{j-1} = -1 + (j-1)h$, and evaluating the definite integral, we obtain

$$X_0 = \frac{1}{h\pi}\tan^{-1}\left[\frac{h/d}{1 + (h^2/d^2)(i-j)(i-j+1)}\right]$$

Similarly we obtain the results

$$X_1 = \int_0^1 K(x_i, s_{j-1} + ph)\, p\, dp$$

$$= \frac{d}{2\pi h^2}\left[\log \frac{d^2 + h^2(i-j)^2}{d^2 + h^2(i-j+1)^2}\right] + (i-j+1)\, X_0$$

$$X_2 = \int_0^1 K(x_i, s_{j-1} + ph)\, p^2\, dp$$

$$= \frac{d}{\pi h^2} - \left[\frac{d^2}{h^2} + (i-j+1)^2\right] X_0 + 2X_1\,(i-j+1)$$

$$X_3 = \int_0^1 K(x_i, s_{j-1} + ph)\, p^3\, dp$$

$$= \frac{d}{2\pi h^2}[5 + 4(i-j)] + \left[3(i-j+1)^2 - \frac{d^2}{h^2}\right] X_1$$

$$- 2(i-j+1)\left[\frac{d^2}{h^2} + (i-j+1)^2\right] X_0$$

The system of equations was solved by the Gauss–Seidel iteration method and a standard subroutine was used for this. The results are summarized in the following table for different values of d, and agree closely well with those obtained by Phillips [1972]. It was found that the method is unsuitable for finding the solution for larger values of d as the convergence is rather slow. Thus for $d = 1.0$ the value obtained with 500 iterations for $x = 1.0$ is 0.80692 compared to the true value 0.75572. For more computational results, see the paper by Sastry [1975].

Cubic Spline Solutions of Love's Equation

	y(x)		
x	*d* = 0.1	*d* = 0.01	*d* = 0.001
0.0	0.51261	0.50146	0.50015
0.2	0.51470	0.50158	0.50016
0.4	0.51858	0.50187	0.50019
0.6	0.52876	0.50261	0.50026
0.8	0.60688	0.51713	0.50271
1.0	0.78627	0.69641	0.67179

These results show that the spline method for the numerical solution of Fredholm integral equations is potentially useful. Its application to more complicated problems will have to be examined together with an estimation of error in the method.

10.3 SINGULAR KERNELS

If $K(s, t)$ is discontinuous or continuous but badly behaved, the integral equation is called a *singular* integral equaion and the quadrature methods, discussed earlier, should not be applied. We may, however, approximate the smooth part of the integrand by a simple function and then integrate the total new integrand exactly. Such formulae are called *generalized quadrature* formulae, also called *product integration* formulae.
We consider the integral equation

$$f(x) + \int_a^b K(x, t)\ f(t)\, dt = \phi(x), \qquad a \le x \le b. \tag{10.40}$$

Let $b - a = nh$ and $t_j = a + jh,\ j = 0, 1, \ldots, n$ so that $t_0 = a$ and $t_n = b$. Then Eq. (10.40) can be written as

$$f(x) + \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} K(x, t)\ f(t)\, dt = \phi(x). \tag{10.41}$$

We now approximate the integral term in Eq. (10.41) by the generalized trapezoidal rule discussed in Chapter 6, i.e. we replace $f(t)$ in the integral by the linear interpolating function $f_n(t)$ given by

$$f_n(t) = \frac{1}{h}[(t_{j+1} - t)\ f(t_j) + (t - t_j)\ f(t_{j+1})]. \tag{10.42}$$

Substituting Eq. (10.42) in Eq. (10.41), we obtain

$$f(x) + \frac{1}{h}\sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} K(x, t)\,[(t_{j+1} - t)\, f(t_j) + (t - t_j)\, f(t_{j+1})]\, dt = \phi(x).$$

Setting $t = t_j + ph$, this gives

$$f(x) + h\sum_{j=0}^{n-1} \int_0^1 [(1 - p)\, f(t_j) + p f(t_{j+1})]\, K(x, t_j + ph)\, dp = \phi(x).$$

If we now put $x = x_i$, $i = 0, 1, 2, \ldots, n$, we obtain

$$f_i + \sum_{j=0}^{n-1} (\alpha_{ij} f_j + \beta_{ij} f_{j+1}) = \phi_i, \qquad (i = 0, 1, 2, \ldots, n) \tag{10.43}$$

where

$$\left.\begin{aligned} \alpha_{ij} &= h\int_0^1 (1-p)\, K(x_i, t_j + ph)\, dp \\ \beta_{ij} &= h\int_0^1 pK(x_i, t_j + ph)\, dp \end{aligned}\right\} \tag{10.44}$$

and

$$f_i = f(x_i) = f(a + ih)$$

Equation (10.43) represents a system of $(n + 1)$ linear equations in $(n + 1)$ unknowns $f(t_0)$, $f(t_1), \ldots, f(t_n)$, and can therefore be solved. We illustrate the use of generalized quadrature with two numerical examples.

Example 10.11 We consider again Love's equation discussed in Example 10.8. The integrals α_{ij} and β_{ij} in Eq. (10.44) can be computed analytically. If we denote

$$\begin{aligned} X0(i, j) &= \int_0^1 K(x_i, t_j + ph)\, dp \\ &= \frac{1}{\pi}\int_0^1 \frac{1}{1 + h^2(i - j - p)^2}\, dp \\ &= \frac{1}{h\pi}\tan^{-1}\left[\frac{h}{1 + h^2(i - j)(i - j - 1)}\right] \end{aligned}$$

and

$$\begin{aligned} X1(i, j) &= \int_0^1 pK(x_i, t_j + ph)\, dp \\ &= \frac{1}{\pi}\int_0^1 \frac{p\, dp}{1 + h^2(i - j - p)^2} \\ &= \frac{1}{2h^2\pi}\log\left[\frac{1 + h^2(i - j - 1)^2}{1 + h^2(i - j)^2}\right] + (i - j)\, X0(i, j) \end{aligned}$$

then

$$\alpha_{ij} = h[X0(i, j) - X1(i, j)] \quad \text{and} \quad \beta_{ij} = hX1(i, j).$$

With $n = 4$, we obtain from Eq. (10.43) the equations

$$1.076 f_0 + 0.126 f_1 + 0.081 f_2 + 0.050 f_3 + 0.018 f_4 = 1.0$$

$$0.071 f_0 + 1.153 f_1 + 0.126 f_2 + 0.081 f_3 + 0.029 f_4 = 1.0$$

$$0.047 f_0 + 0.126 f_1 + 1.153 f_2 + 0.126 f_3 + 0.047 f_4 = 1.0$$

$$0.029 f_0 + 0.081 f_1 + 0.126 f_2 + 0.153 f_3 + 0.071 f_4 = 1.0$$

$$0.018 f_0 + 0.050 f_1 + 0.081 f_2 + 0.126 f_3 + 1.076 f_4 = 1.0$$

The solution of this system, which is centro-symmetric, was obtained on a digital computer. The computations were repeated for $n = 8$, 16 and 32 and the results, together with the exact values, are tabulated below:

x	*Exact y(x)*	*n*	*Computed y(x)*	*Error*
0.0	0.65741	4	0.65609	0.00132
		8	0.65708	0.00033
		16	0.65733	0.00008
		32	0.65739	0.00002
1.0	0.75572	4	0.75484	0.00088
		8	0.75550	0.00022
		16	0.75566	0.00006
		32	0.75570	0.00002

Comparison with the results obtained by the ordinary trapezoidal rule (see table of results in Example 10.8) shows that this rule gives better accuracy than the ordinary trapezoidal rule. The order of convergence is h^2 as in the latter rule.

The next example demonstrates the use of generalized quadrature in dealing with kernels having a logarithmic singularity.

Example 10.12 We consider now an example from fluid mechanics involving potential flow of an incompressible inviscid fluid.

In many fluid dynamics problems, it is necessary to calculate the pressure distribution on the surface of a body moving in a fluid. For a body of revolution in axial flow, Vandrey [1961] derived the linear integral equation

$$v(s) = 2x'(s) - \frac{1}{\pi}\int_0^L K(s,\sigma)\, v(\sigma)\, d\sigma, \quad 0 \le s \le L \tag{i}$$

where

$$\left.\begin{aligned} K(s,\sigma) &= \frac{1}{\sqrt{(x-\xi)^2+(y+\eta)^2}}\left\{\frac{x'y-y'(x-\xi)}{y}K(k)\right. \\ &\quad \left. -E(k)\left[\frac{x'y-y'(x-\xi)}{y}+2\eta\frac{x'(y-\eta)-y'(x-\xi)}{(x-\xi)^2+(y-\eta)^2}\right]\right\} \\ K_2 &= \frac{4y\eta}{(x-\xi)^2+(y+\eta)^2},\ x'=\frac{dx}{ds} \end{aligned}\right\} \quad \text{(ii)}$$

and $K(k)$ and $E(k)$ are complete elliptic integrals of the first and second kinds respectively with modulus k. In (i), $v(s)$ denotes the velocity distribution function on the body surface from which the pressure distribution can be found by Bernoulli's equation. Details of the problem and its reduction to a system of equations are given in the papers by Kershaw [1971] and Sastry [1973, 1976], where further references may be found. Using the expansions of $K(k)$ and $E(k)$ given in Dwight [1934], the kernel $K(s,\sigma)$ in (ii) can be split into the form:

$$K(s,\sigma) = P(s,\sigma)\log|s-\sigma| + Q(s,\sigma) \quad \text{(iii)}$$

where

$$\left.\begin{aligned} P(s,\sigma) &= -\frac{1}{\sqrt{(x-\xi)^2+(y+\eta)^2}}\left\{\frac{x'y-y'(x-\xi)}{y}\frac{2}{\pi}E(k_1)\right. \\ &\quad \left. -2\eta\frac{x'(y-\eta)-y'(x-\xi)}{(x-\xi)^2+(y-\eta)^2}\frac{2}{\pi}[K(k_1)-E(k_1)]\right\} \\ Q(s,\sigma) &= K(s,\sigma)-P(s,\sigma)\log|s-\sigma| \end{aligned}\right\} \quad \text{(iv)}$$

and

$$k_1^2 = 1-k^2 = \frac{(x-\xi)^2+(y-\eta)^2}{(x-\xi)^2+(y+\eta)^2}.$$

When $\sigma = s$, it is found that

$$\left.\begin{aligned} P(s,s) &= -\frac{x'}{2y} \\ Q(s,s) &= \frac{1}{2y}\left\{x'\left[-\frac{1}{2}\log(x'^2+y'^2)+\frac{1}{2}\log 4y^2+\log 4-1\right]-y\frac{x''y'-y''x'}{x'^2+y'^2}\right\} \end{aligned}\right\} \quad \text{(v)}$$

The method of generalized quadrature described in Chapter 6 can now be applied to reduce the integral equaion (i) to a system of linear algebraic equations.

The table below gives the numerical results for a cylinder. As the solution is centro-symmetric, the results are given only up to $s = 90°$. The computations are made with 20 subdivisions and the accuracy is quite good. For the sake of camparison, the accurate value 1.5 $\sin s$ is also tabulated. On running the program twice with $n = 10$ and $n = 20$, it was found that the order of convergence is two.

s (in deg)	*Accurate value of v(s)*	*Computed value*	*Error*
18	0.4635	0.4619	0.0016
36	0.8817	0.8816	0.0001
54	1.2135	1.2141	0.0006
72	1.4266	1.4275	0.0009
90	1.5000	1.5011	0.0011

For a numerical solution of this problem using Everett's formula, see Kershaw [1961].

10.4 METHOD OF INVARIANT IMBEDDING

This is a method of recent origin, being mainly due to the efforts of Kalaba and Ruspini [1969], and is applicable to Fredholm integral equations of the second kind

$$y(x) = g(x) + \int_0^a K(x, s)\, y(s)\, ds \tag{10.45a}$$

where

$$K(x, s) = \int_0^\infty f(xz)\, f(sz)\, w(z)\, dz \tag{10.45b}$$

In the method of invariant imbedding, Eq. (10.45) is first rewritten as a Volterra integral equation in the form

$$y(x, t) = g(x) + \int_0^t K(x, s)\, y(s, t)\, ds; \qquad 0 \le x \le t; \quad 0 \le t \le a. \tag{10.46}$$

An essential feature of the method is to convert the Volterra integral equation (10.46) into initial–value problems and then solve the initial-value problems by any of the standard techniques. The transformation to the initial-value problems involves a series of complicated mathematical manipulations and

the interested reader is referred to the original paper by Kalaba and Ruspini [1969]. We, however, demonstrate its applicability to a practical situation.

Example 10.13 We consider the problem proposed by Srivastava and Palaiya [1969] who have studied the distribution of thermal stresses in a semi-infinite solid containing a pennyshaped crack situated parallel to the free boundary. The free boundary of the solid is kept at zero temperature and in the axisymmetric case the problem is reduced to the solution of a Fredholm integral equation of the second kind

$$y(x) + \int_0^1 K(x, s)\, y(s)\, ds = -\frac{4}{\pi}, \tag{i}$$

where

$$K(x, s) = -\frac{2}{\pi} \int_0^\infty e^{-2\xi H} \cos \xi x \cos \xi s \, d\xi, \tag{ii}$$

in which $y(x)$ represents the non-dimensionalized stress distribution function and the integral equation was derived by assuming that the centre of the crack is at the origin; that the solid, which is isotropic and homogeneous, is divided into two domains: (i) the layer defined by $-H \le z \le 0$, and (ii) the half-plane $0 \le z \le \infty$; and that the temperature prescribed on the surface of the crack is constant. The derivation and physical details of the problem may be found in the above cited reference where the integral equation was solved by the classical iterative method for small values of the ratio of the radius of the crack to that of its distance from the free boundary, and for values of this ratio *nearer* unity, the equation was solved numerically by quadrature method.

For the numerical solution by the method of invariant imbedding, the radius of the crack is assumed to be of unit length and the integrals are approximated by using Gaussian quadrature. Then, the initial-value problems become:

$$\left.\begin{aligned} \frac{dR_{ik}(t)}{dt} &= \left[\cos(tA_k) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{mk}(t)\right] \\ &\quad \times \left[\cos(tA_i) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{im}(t)\right] \\ R_{ik}(0) &= 0 \end{aligned}\right\} \tag{iii}$$

and

$$\left.\begin{aligned}\frac{de_i(t)}{dt} &= \left[g(t) + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, e_m(t)\right] \\ &\quad \times \left[\cos tA_i + \sum_{m=1}^{N} \frac{2}{(1+a_m)^2} F_m W(A_m) \cos(tA_m)\, R_{im}(t)\right] \\ e_i(0) &= 0, \qquad 1 \le i \le N, \quad 0 \le t \le 1.\end{aligned}\right\} \quad \text{(iv)}$$

where

$$e_i(t) = e(A_i, t)$$

and finally,

$$y(x, t) = g(x) + \sum_{m=1}^{N} \frac{2F_m}{(1+a_m)^2} W(A_m) \cos xA_m\, e_m(t), \qquad 0 \le x \le t \le 1. \quad \text{(v)}$$

In Eqs. (iii) to (v), the notation

$$A_n = \frac{1-a_n}{1+a_n}$$

is used, a_m and F_m being the abscissae and weights of the N-point Gaussian quadrature formula defined by

$$\int_{-1}^{1} f(x)\, dx = \sum_{m=1}^{N} F_m f(a_m)$$

The Eqs. (iii) and (iv) have been solved using the fourth-order Runge–Kutta method, and the *five-point Gaussian formula*. The results are obtained on a digital computer and are given in the following table for different values of H:

H	x	$y(x)$
1.05	0.0	–1.7718
	1.0	–1.7013
1.1	0.0	–1.7450
	1.0	–1.6813
1.2	0.0	–1.6898
	1.0	–1.6464
1.3	0.0	–1.6599
	1.0	–1.6169
1.6667	0.0	–1.5618
	1.0	–1.5397

Although the method produces results which agree quite well with those obtained by Srivastava and Palaiya, it suffers with the serious disadvantage of being a complicated process and requiring an enormous amount of computing time.

A central idea of the method is to take full advantage of the ability of the modern highspeed digital computer to solve systems of ordinary differential equations with given initial conditions, and it therefore finds important applications in the numerical solution of integral equations occurring in radiative transfer, optimal filtering and multiple scattering.

EXERCISES

Verify whether the functions given below are solutions of the integral equations indicated against them (Problems 10.1–10.5):

10.1 $f(x) = 1$: $f(x) + \int_0^1 x(e^{xt} - 1)\, f(t)\, dt = e^x - x.$

10.2 $f(t) = e^t$: $f(t) + \lambda \int_0^1 \sin(tx) f(x)\, dx = 1.$

10.3 $f(x) = xe^x$: $f(x) = e^x \sin x + 2\int_0^x \cos(x-t) f(t)\, dt.$

10.4 $\phi(x) = x$: $\phi(x) = \dfrac{15x-2}{18} + \int_1^3 (x+t)\phi(t)\, dt.$

10.5 $f(x) = x - 1$: $f(x) = \int_0^1 (x+t)\, f(t)\, dt + \dfrac{3}{2}x - \dfrac{5}{6}.$

Transform the following initial value problems into equivalent integral equations (Problems 10.6–10.8):

10.6 $y'' + y = \cos x, \quad y(0) = 0, \quad y'(0) = 0.$

10.7 $x^2y'' - xy' + y = x^2, \quad y(1) = 1, \quad y'(1) = 0.$

10.8 $y'' = 1 - xy, \quad y(0) = y'(0) = 0.$

Solve the following integral equations with degenerate kernels (Problems 10.9–10.12):

10.9 $f(x) - \lambda \int_{-\pi/4}^{\pi/4} \tan s\, f(s)\, ds = \cot x.$

10.10 $f(x) - \lambda \int_0^{\pi/2} \sin x \cos t\, f(t)\, dt = \sin x.$

10.11 $f(x) - \lambda \int_0^{\pi} \sin(x-u) f(u)\, du = \cos x.$

10.12 $f(x) = \lambda \int_0^{2\pi} \sin(x+t) f(t)\, dt + x.$

10.13 Solve the integral equations in Problems 10.4 and 10.5 by Simpson's 1/3-rule. In each case, divide the range into two equal subintervals and approximate to the solution. Compare your results with the exact solution.

Answers to Exercises

10.1 Satisfies

10.2 Does not satisfy

10.3 Satisfies

10.4 Satisfies

10.5 $f(x) = x - 1$ satisfies the integral equation

10.6 $y(x) = -\int_0^{x} (x-t) y(t)\, dt - \cos x + 1$

10.7 $\phi(x) = 1 - \dfrac{1}{x^2} + \dfrac{1}{x^2} \int_1^{x} t\, \phi(t)\, dt$

10.8 $y(x) = \dfrac{x^2}{2} - \int_0^{x} (x-t) y(t)\, dt$

10.9 $f(x) = \lambda \dfrac{\pi}{2} + \cot x$

10.10 $f(x) = \dfrac{2 \sin x}{2 - \lambda}$

10.11 $f(x) = \dfrac{2\pi\lambda \sin x}{4 + \pi^2 \lambda^2} + \dfrac{4 \cos x}{4 + \pi^2 \lambda^2}$

10.12 $f(x) = 2\pi\lambda \left[\dfrac{\pi \sin x}{\lambda^2 \pi^2 - 1} + \dfrac{\cos x}{\lambda^2 \pi^2 - 1} \right] + x, \quad \lambda^2 \pi^2 \neq 1.$

10.13 (a) $f_0 = 0$, $f_1 = \dfrac{1}{2}$, $f_2 = 1$, (b) $f_0 = -1$, $f_1 = -\dfrac{1}{2}$, $f_2 = 0$.

Chapter 11

The Finite Element Method

11.1 INTRODUCTION

In Chapters 8 and 9 we discussed finite difference methods for the solution of boundary-value problems defined by ordinary and partial differential equations. We now describe another class of methods for the solution of such problems, known as the *finite element methods*. A full discussion of these methods is outside the scope of this book—as normally this does not form part of an introductory course on numerical methods. We give here only a brief presentation so as to enable the reader to know that such methods exist. The discussion includes an elementary formulation of the method with simple applications to ordinary differential equations. For details, the reader is referred to the excellent book by Reddy [1985].

The basic idea behind the finite element method is to replace a continuous function by means of piecewise polynomials. Such an approximation, called the *piecewise polynomial approximation*, will be discussed in Section 11.1.2. The reader is already aware of the importance of polynomial approximations in numerical analysis. These are used in the numerical solution of practical problems where the exact functions are difficult to obtain or cumbersome to use. The idea of piecewise polynomial approximation is also not new to the reader, since the cubic spline already discussed, belongs to this class of polynomials.

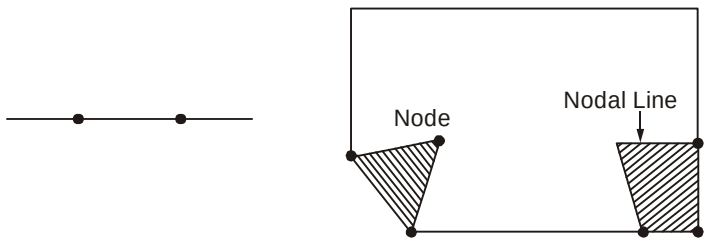
In engineering applications, several approximate methods of solution are used and the reader is familiar with a few of them, e.g. the method of least squares, method of collocation, etc. In Section 11.2, we discuss two important methods of approximation, viz., the Rayleigh–Ritz method and the Galerkin

technique. Rayleigh developed the method to solve certain vibration problems and Ritz provided a mathematical basis for it and also applied it to more general problems. Whereas the Rayleigh–Ritz method is based on the existence of a *functional* (see Section 11.1.1), the Galerkin technique uses the governing equations of the problem and minimizes the error of the approximate solution. The latter does not require a functional. A disadvantage of both these methods is that higher-order polynomials have to be used to obtain reasonable accuracy.

The finite element method, described in the present chapter, is one of the most important numerical applications of the Rayleigh–Ritz and Galerkin methods. Its mathematical software is quite popular and used extensively in the solution of many practical problems of engineering and applied science. In the finite element method, the domain of integration is subdivided into a number of smaller regions called *elements* and over each of these elements the continuous function is approximated by a suitable piecewise polynomial. To obtain a better approximation one need not use higher-order polynomials but only use a finer subdivision, i.e. increase the number of elements.

In practice, several types of elements are in use, the type used being largely dependent upon the geometrical shape of the region under consideration. In two-dimensional problems, the elements used are triangles, rectangles and quadrilaterals. For three-dimensional problems, tetrahedra, hexahedra and parallelopiped elements are used. Since our attempt in this chapter is only to introduce the finite element method, we restrict its application to the solution of simple one-dimensional problems (see Section 11.4.1).

Examples of typical finite elements are shown in Fig. 11.1.



(a) Line element (b) Triangular and Quadrilateral elements

Figure 11.1 Typical finite elements.

11.1.1 Functionals

The concept of a functional is required to understand the Rayleigh–Ritz method, which will be discussed in the next section. This concept arises in the study of variational principles, which occur widely in physical and other problems. Mathematically, a variational principle consists in determining the extreme value of the integral of a typical function, say $f(x, y, y')$. Here the integrand is a function of the coordinates and their derivatives and the

integration is performed over a region. Consider, for example, the integral defined by

$$I(y) = \int_a^b f(x, y, y')\, dx, \tag{11.1}$$

where $y(x)$ satisfies the boundary conditions $y(a) = y(b) = 0$.

The integrand f is integrated over the one-dimensional domain x. I is said to be a functional and is defined as a function which transforms a function y into a real number, the value of the definite integral in Eq. (11.1). From calculus of variations we know that a necessary condition for $I(y)$ to have an extremum is that $y(x)$ must satisfy the Euler–Lagrange differential equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0. \tag{11.2*}$$

Similarly, for functionals of the form

$$I(y) = \int_a^b f(x, y, y', y'')\, dx \tag{11.3}$$

the Euler–Lagrange equation takes the form

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial f}{\partial y''}\right) = 0. \tag{11.4}$$

The Euler–Lagrange equation (11.2) has several solutions and the one which satisfies the given boundary conditions is selected. Thus, one determines the functional so that it takes on an extremum value from a set of permissible functions. This is the central problem of a variational principle. An important point here is that an extremum may not exist. In other words, a variational principle may exist, but an extremum may not exist. Furthermore, not all differential equations have a variational principle. These difficulties are serious and therefore impose limitations on the application of the variational principle to the solution of engineering problems.

Many problems arising in physics and engineering are modelled by boundary-value problems and initial boundary-value problems. Frequently, these equations are equivalent to the problem of the minimization of a functional which can be interpreted in terms of the total energy of the given system. In any physical situation, therefore, the functional is obtained from a consideration of the total energy explicitly. Mathematically, however, it would be useful to be able to determine the functional from the governing differential equation itself. This is illustrated below with an example.

*For example, see Sastry [2004].

Example 11.1 Find the functional for the boundary-value problem defined by

$$\frac{d^2y}{dx^2} = f(x) \tag{i}$$

and

$$y(a) = y(b) = 0. \tag{ii}$$

We have

$$\delta \int_a^b f\, y\, dx = \int_a^b f\, \delta y\, dx$$

$$= \int_a^b \frac{d^2y}{dx^2} \delta y\, dx, \text{ since } f(x) = \frac{d^2y}{dx^2}$$

$$= \left[\frac{dy}{dx}\delta y\right]_a^b - \int_a^b \frac{dy}{dx}\frac{d}{dx}(\delta y)\, dx, \text{ on integrating by parts}$$

$$= -\int_a^b \frac{dy}{dx}\frac{d}{dx}(\delta y)\, dx, \text{ since } \delta y(a) = \delta y(b) = 0$$

$$= -\int_a^b \frac{dy}{dx}\delta\left(\frac{dy}{dx}\right) dx, \text{ since } \frac{d}{dx}(\delta y) = \delta\left(\frac{dy}{dx}\right)$$

$$= -\int_a^b \frac{1}{2}\delta\left(\frac{dy}{dx}\right)^2 dx$$

$$= -\delta \int_a^b \frac{1}{2}\left(\frac{dy}{dx}\right)^2 dx.$$

Hence

$$\delta \int_a^b \left[fy + \frac{1}{2}\left(\frac{dy}{dx}\right)^2 \right] dx = 0.$$

It follows that a unique solution of the problem (i) to (ii) exists at a minimum value of the integral defined by

$$I(v) = \int_a^b \left[fv + \frac{1}{2}\left(\frac{dv}{dx}\right)^2 \right] dx. \tag{iii}$$

A quicker way of finding the functional of a boundary value problem is the following (See Reddy [1985]).

Let $v(x)$ be a function satisfying the essential boundary conditions, viz. $v(a) = v(b) = 0$. Multiply the differential equation written in the form

$$-y'' + f(x) = 0$$

by v and integrate with respect to x. We then obtain

$$\begin{aligned} 0 &= -\int_a^b vy''\,dx + \int_a^b vf\,dx \\ &= [-vy']_a^b + \int_a^b v'y'\,dx + \int_a^b vf\,dx \\ &= \int_a^b (v'y' + vf)\,dx \end{aligned}$$

Finally, substitute $y = v$ in the above and multiply the *bilinear* terms by 1/2. We then obtain the required functional

$$I(v) = \int_a^b \left[\frac{1}{2}v'^2 + vf\right] dx,$$

which is the same as Eq. (iii) obtained earlier.

By definition, therefore, the integral in (iii) represents the required functional of the problem. In a similar way, functionals of other boundary-value and initial boundary-value problems can be derived.

It is outside the scope of this book to deal extensively with the determination of functionals corresponding to boundary-value problems. We list below some familiar boundary-value problems with their associated functionals and these would be useful in understanding the problems discussed in this chapter.

(i)
$$\frac{d^2y}{dx^2} = f(x),\ y(a) = y(b) = 0 \tag{11.5}$$

$$I(v) = \int_a^b v(2f - v'')\,dx. \tag{11.6}$$

(ii)
$$\frac{d^2y}{dx^2} + ky = x^2,\ 0 < x < 1;\ y(0) = 0,\ \left(\frac{dy}{dx}\right)_{x=1} = 1 \tag{11.7}$$

$$I(v) = \frac{1}{2}\int_0^1 \left[\left(\frac{dv}{dx}\right)^2 - kv^2 + 2vx^2\right] dx - v(1). \tag{11.8}$$

(iii) $$x^2 y'' + 2xy' = f(x),\ y(a) = y(b) = 0 \tag{11.9}$$

$$I(v) = \int_a^b v\left[2f - \frac{d}{dx}(x^2 y')\right] dx. \tag{11.10}$$

(iv) $\nabla^2 u = 0,\ u = 0$ on the boundary C of R. (11.11)

$$I(v) = \iint_R \frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dx\, dy. \tag{11.12}$$

(v) $\nabla^2 u = -f,\ u = 0$ on the boundary C of R. (11.13)

$$I(v) = \iint_R \left\{\frac{1}{2}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] - uf\right\} dx\, dy. \tag{11.14}$$

(vi) $$\left.\begin{aligned} EI\frac{d^4 y}{dx^4} + ky &= f(x),\ 0 < x < l \\ y = 0 = \frac{d^2 y}{dx^2} &\text{ at } x = 0,\ l \end{aligned}\right\} \tag{11.15}$$

$$I(v) = \frac{1}{2}\int_0^l \left[EI\left(\frac{d^2 v}{dx^2}\right)^2 + kv^2 - 2vf\right] dx. \tag{11.16}$$

11.1.2 Base Functions

Suppose we wish to approximate a real-valued function $f(x)$ over a finite interval $[a, b]$. A usual approach is to divide $[a, b]$ into a number of subintervals $[x_i,\ x_{i+1}]$, $i = 0, 1, 2, \ldots, n-1$, where $x_0 = a$ and $x_n = b$, and to interpolate linearly between the values of $f(x)$ at the end points of each subinterval. In $[x_i,\ x_{i+1}]$, the linear approximating function is given by

$$l_i(x) = \frac{1}{h_i}[(x_{i+1} - x) f_i + (x - x_i) f_{i+1}], \tag{11.17}$$

where $h_i = x_{i+1} - x_i$. From this, we construct the piecewise linear interpolating function over $[x_0, x_n]$ by the formula

$$P(x) = \sum_{i=0}^{n} \phi_i(x)\ f_i \tag{11.18}$$

where

$$\left.\begin{aligned}
\phi_0(x) &= \begin{cases} (x_1 - x)/h_0 & x_0 \le x \le x_1 \\ 0, & x_1 \le x \le x_n \end{cases} \\
\phi_i(x) &= \begin{cases} (x - x_{i-1})/h_{i-1}, & x_{i-1} \le x \le x_i \\ (x_{i+1} - x)/h_i, & x_i \le x \le x_{i+1} \\ 0, & x \ge x_{i+1} \end{cases} \\
\phi_n(x) &= \begin{cases} 0, & x_0 \le x \le x_{n-1} \\ (x - x_{n-1})/h_{n-1} & x_{n-1} \le x \le x_n \end{cases}
\end{aligned}\right\} \tag{11.19}$$

The functions $\phi_i(x)$, $i = 1, 2, \ldots, n$ are called *base functions* or *shape functions*. It is easily seen that the base functions $\phi_i(x)$ are identically zero except for the range $[x_{i-1}, x_{i+1}]$ with $\phi_i(x_i) = 1$.

Other types of base functions such as piecewise Hermite polynomials, cubic splines, etc., are also used in the literature but these will not be considered in this book.

11.2 METHODS OF APPROXIMATION

In this section we discuss two methods of approximation, viz. the *Rayleigh–Ritz* and *Galerkin* methods. As mentioned earlier, the former method is based on the existence of a functional which is then minimized. The second technique is due to Galerkin who proposed it as an error minimization method. It belongs to a wider class of methods called *weighted residual methods*. An advantage of the Galerkin method is that it works with the governing equations of the problem and does not require a functional.

Both the methods have a common feature in that they seek an approximate solution in the form of a linear combination of base functions. Nevertheless, they differ from each other in choosing the base functions.

11.2.1 Rayleigh–Ritz Method

In this method we do not obtain the actual minimum but only an approximate solution as nearer the actual solution as the base functions allow. To obtain a good approximation, therefore, the choice of the base functions is important and to improve the approximation, the number of base functions should be increased.

We explain this method by considering second-order boundary-value problem defined by

$$y'' + p(x)y + q(x) = 0, \qquad y(a) = y(b) = 0. \tag{11.20}$$

The functional for the above problem is given by

$$I(v) = \int_a^b \left[\left(\frac{dv}{dx} \right)^2 - pv^2 - 2qv \right] dx = 0. \tag{11.21}$$

From the definition of the functional we know that if $y(x)$, the solution of Eq. (11.20), is substituted in Eq. (11.21), then the integral I will be minimum. Since we do not know the solution of Eq. (11.20), we try with an approximate solution and determine the parameters of the approximation so that the integral is minimum. This is the central idea of the Rayleigh–Ritz method. Now, let

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x) \tag{11.22}$$

be an approximate solution where the base functions, $\phi_i(x)$, are linearly independent and satisfy the boundary conditions given in Eq. (11.20), i.e. let

$$\phi_i(a) = 0 \quad \text{and} \quad \phi_i(b) = 0. \tag{11.23}$$

Substituting for v in Eq. (11.21), we obtain

$$I(\alpha_1, \alpha_2, \ldots \alpha_n) = \int_a^b \left\{ \left[\frac{d}{dx} \sum \alpha_i \phi_i(x) \right]^2 - p\left[\sum \alpha_i \phi_i(x) \right]^2 - 2q \sum \alpha_i \phi_i(x) \right\} dx = 0. \tag{11.24}$$

For minimum, we have

$$\frac{\partial I}{\partial \alpha_1} \delta\alpha_1 + \frac{\partial I}{\partial \alpha_2} \delta\alpha_2 + \cdots + \frac{\partial I}{\partial \alpha_n} \delta\alpha_n = 0. \tag{11.25}$$

Since the $\delta\alpha_i$ are arbitrary, Eq. (11.25) gives

$$\frac{\partial I}{\partial \alpha_i} = 0, \qquad i = 1, 2, \ldots, n. \tag{11.26}$$

If I is a quadratic function of y and dy/dx, then Eq. (11.26) will be linear in α_i and can be solved easily.

We state, without proof, that the Rayleigh–Ritz method converges to the actual solution of the problem provided that the functions ϕ_i are linearly independent and satisfy at least the essential boundary conditions of the problem. The following examples illustrate the method of procedure.

Example 11.2 We consider the two-point boundary-value problem defined by

$$y'' + x = 0, \quad 0 < x < 1, \qquad y(0) = y(1) = 0. \tag{i}$$

From Eq. (11.6), we have

$$I(v) = \int_0^1 v(-2x - v'')\, dx = -\int_0^1 v(2x + v'')\, dx. \tag{ii}$$

Let

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x) \tag{iii}$$

where

$$\phi_i(0) = \phi_i(1) = 0 \qquad \text{for all } i. \tag{iv}$$

Substituting (iii) in (ii), we obtain

$$I(v) = -\int_0^1 \left[\sum_{i=1}^{n} \alpha_i \phi_i(x)\right]\left[2x + \sum_{j=1}^{n} \alpha_j\, \phi_j''(x)\right] dx. \tag{v}$$

For convenience, we set

$$p_i = \int_0^1 x\, \phi_i(x)\, dx \tag{vi}$$

and

$$q_{ij} = \int_0^1 \phi_i(x)\phi_j''(x)\, dx$$

$$= [\phi_i(x)\, \phi_j'(x)]_0^1 - \int_0^1 \phi_i'(x)\, \phi_j'(x)\, dx, \text{ on integrating by parts}$$

$$= -\int_0^1 \phi_i'(x)\, \phi_j'(x)\, dx, \tag{vii}$$

using boundary conditions (iv).

Then Eq. (v) becomes

$$I(v) = -2\sum_{i=1}^{n} \alpha_i p_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j q_{ij}$$

Hence $\partial I/\partial \alpha_i = 0$ gives

$$2p_i + 2\sum_{j=1}^{n} \alpha_j q_{ij} = 0, \qquad (i = 1, 2, \ldots, n). \tag{viii}$$

We wish to find an approximate solution with $n = 2$ and we therefore choose $\phi_1(x) = x(1-x)$ and $\phi_2(x) = x^2(1-x)$, so that the boundary conditions (iv) are satisfied.

Now, from (vi), we have

$$p_1 = \int_0^1 x^2(1-x)\,dx = \frac{1}{12}$$

and

$$p_2 = \int_0^1 x^3(1-x)\,dx = \frac{1}{20}.$$

Also, $\phi_1'(x) = 1-2x$ and $\phi_2'(x) = 2x-3x^2$. Equation (vii) gives

$$q_{11} = -\int_0^1 (1-2x^2)\,dx = -\frac{1}{3}$$

$$q_{12} = -\int_0^1 (1-2x)(2x-3x^2)\,dx = -\frac{1}{6} = q_{21}, \text{ by symmetry}$$

$$q_{22} = -\int_0^1 (2x-3x^2)^2\,dx = -\frac{2}{15}.$$

Equations (viii) now give

$$4\alpha_1 + 2\alpha_2 = 1 \quad \text{and} \quad 10\alpha_1 + 8\alpha_2 = 3,$$

whose solution is $\alpha_1 = \alpha_2 = 1/6$. Hence

$$v(x) = \frac{1}{6}x(1-x) + \frac{1}{6}x^2(1-x) = \frac{1}{6}x(1-x^2).$$

It can be verified that this is the exact solution of the problem (i).

Example 11.3 Solve the boundary-value problem defined by

$$y'' + y = -x, \qquad 0 < x < 1 \tag{i}$$

with

$$y(0) = y(1) = 0 \tag{ii}$$

The exact solution of the problem (i) and (ii) is given by

$$y(x) = \frac{\sin x}{\sin 1} - x. \tag{iii}$$

To find the approximate solution by the Rayleigh–Ritz method, we take the functional in the form

$$I(v) = \int_0^1 (vv'' + v^2 + 2vx)\,dx. \tag{iv}$$

Let an approximate solution be given by

$$v(x) = \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{v}$$

where

$$\phi_i(0) = \phi_i(1) = 0 \text{ for all } i. \tag{vi}$$

Substituting for v in (iv), we obtain

$$I(v) = \int_0^1 \left[\sum_{i=1}^{n} \alpha_i \phi_i(x) \sum_{j=1}^{n} \alpha_j \phi_j''(x) + \sum_{i=1}^{n} \alpha_i \phi_i(x) \sum_{j=1}^{n} \alpha_j \phi_j(x) + 2x \sum_{i=1}^{n} \alpha_i \phi_i(x) \right] dx \tag{vii}$$

As in the previous example, we let

$$p_i = \int_0^1 x\phi_i(x)\, dx \tag{viii}$$

and

$$q_{ij} = \int_0^1 \phi_i(x)\phi_j''(x)\, dx = -\int_0^1 \phi_i'(x)\phi_j'(x)\, dx. \tag{ix}$$

Further, let

$$r_{ij} = \int_0^1 \phi_i(x)\phi_j(x)\, dx. \tag{x}$$

Equation (vii) now becomes

$$I(v) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j q_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j r_{ij} + 2 \sum_{i=1}^{n} \alpha_i p_i \tag{xi}$$

For minimum, we, therefore, have

$$\frac{\partial I}{\partial \alpha_i} = 2 \sum_{j=1}^{n} \alpha_j q_{ij} + 2 \sum_{j=1}^{n} \alpha_j r_{ij} + 2p_i = 0,$$

which simplifies to

$$\sum_{j=1}^{n} \alpha_j (q_{ij} + r_{ij}) = -p_i, \qquad i = 1, 2, \ldots, n. \tag{xii}$$

To obtain an approximate solution, we take $n = 2$. Then, Eq. (xii) becomes

$$\left.\begin{aligned} \alpha_1(q_{11} + r_{11}) + \alpha_2(q_{12} + r_{12}) = -p_1 \\ \alpha_1(q_{21} + r_{21}) + \alpha_2(q_{22} + r_{22}) = -p_2 \end{aligned}\right\} \tag{xiii}$$

Choosing $\phi_1(x) = x(1-x)$ and $\phi_2(x) = x^2(1-x)$, we then obtain

$$p_1 = \int_0^1 x^2(1-x)\,dx = \frac{1}{12};$$

$$p_2 = \int_0^1 x^3(1-x)\,dx = \frac{1}{20};$$

$$q_{11} = -\int_0^1 (1-2x)^2\,dx = -\frac{1}{3};$$

$$q_{12} = -\int_0^1 (1-2x)(2x-3x^2)\,dx = -\frac{1}{6};$$

$$q_{22} = -\int_0^1 (2x-3x^2)^2\,dx = -\frac{2}{15};$$

$$r_{11} = \int_0^1 x^2(1-x)^2\,dx = \frac{1}{30};$$

$$r_{12} = \int_0^1 x^3(1-x)^2\,dx = \frac{1}{60};$$

$$r_{22} = \int_0^1 x^4(1-x)^2\,dx = \frac{1}{105}.$$

Equation (xiii) now give

$$18\alpha_1 + 9\alpha_2 = 5$$
$$63\alpha_1 + 52\alpha_2 = 21.$$

Solving, we obtain

$$\alpha_1 = 0.1924 \quad \text{and} \quad \alpha_2 = 0.1707.$$

Hence the approximation is given by

$$y = x(1-x)\,(0.1924 + 0.1707x)$$

Example 11.4 Solve the boundary-value problem defined by

$$y'' - x = 0 \tag{i}$$

and

$$y(0) = 0,\ y'(1) = -\frac{1}{2} \tag{ii}$$

by the Rayleigh–Ritz method.

In this case, one of the boundary conditions is essential and the other natural. Also, the exact solution of the problem is given by

$$y(x) = \frac{x^3}{6} - x. \tag{iii}$$

The functional for this problem is given by

$$I(v) = \int_0^1 (v'^2 + 2vx)\, dx + v(1). \tag{iv}$$

Let

$$v(x) = \alpha_1 x + \alpha_2 x^2 \tag{v}$$

be an approximate solution so that $v(x)$ satisfies the essential boundary condition, viz., $v(0) = 0$. Then $v'(x) = \alpha_1 + 2\alpha_2 x$ and Eq. (iv) gives

$$I(v) = \int_0^1 [(\alpha_1 + 2\alpha_2 x)^2 + 2x(\alpha_1 x + \alpha_2 x^2)]\, dx + \alpha_1 + \alpha_2 \tag{vi}$$

Hence,

$$\left.\begin{aligned} \frac{\partial I}{\partial \alpha_1} &= 0 = \int_0^1 [2\,(\alpha_1 + 2\alpha_2 x) + 2x^2]\, dx + 1 \\ \frac{\partial I}{\partial \alpha_2} &= 0 = \int_0^1 [2\,(\alpha_1 + 2\alpha_2 x)\, 2x + 2x^3]\, dx + 1. \end{aligned}\right\} \tag{vii}$$

Simplification gives the two equations

$$\alpha_1 + \alpha_2 = -\frac{5}{6} \quad \text{and} \quad \alpha_1 + \frac{4}{3}\alpha_2 = -\frac{3}{4}, \tag{viii}$$

whose solution is $\alpha_1 = -13/12$ and $\alpha_2 = 1/4$.
The approximate solution is given by

$$y^{(1)} = -\frac{13}{12}x + \frac{1}{4}x^2.$$

The student should compare this with the exact solution.

11.2.2 Galerkin's Method

The Rayleigh–Ritz method discussed in Section 11.2.1 is a powerful technique for the solution of boundary-value problems. It has, however, the disadvantage of requiring the existence of a functional which is not always possible to obtain. In fact, not all differential equations have a variational principle. Most

engineering problems are expressed in terms of certain governing equations and boundary conditions, and not in terms of a functional. Galerkin's method belongs to a wider class of methods called the *weighted residual methods*. In this method, an approximating function called the *trial function* (which satisfies all the boundary conditions) is substituted in the given differential equation and the result is called the *residual* (the result will not be zero since we have substituted an approximating function). The residual is then weighted and the integral of the product, taken over the domain, is then set to zero. It can be shown that if the Euler–Lagrange equation corresponding to a functional coincides with the differential equation of the problem, then both the Rayleigh–Ritz and Galerkin methods yield the same system of equations.

See Section 8.10.3 for application of this method to solve two-point boundary value problems.

11.3 APPLICATION TO TWO-DIMENSIONAL PROBLEMS

The application of the Rayleigh–Ritz and Galerkin methods to two-dimensional problems, although straightforward, is more complicated because of the increase in the number of parameters to be determined. We illustrate the application of Ritz method with an example.

Example 11.5 We consider Poisson's equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=k, \qquad 0\le x,\ y\le 1 \tag{i}$$

with $u=0$ on the boundary C of the region S.
The functional for the above problem is given by

$$I(v)=\iint_S v\left(2k-\frac{\partial^2 v}{\partial x^2}-\frac{\partial^2 v}{\partial y^2}\right)dx\,dy, \tag{ii}$$

where v vanishes on the boundary C. Let

$$v(x,y)=\alpha xy(x-1)(y-1) \tag{iii}$$

be a first approximation to u. Clearly, v satisfies the boundary conditions, i.e. $v=0$ on the boundary C. The derivatives are given by

$$\left.\begin{aligned} &\frac{\partial v}{\partial x}=\alpha y(y-1)(2x-1); \quad \frac{\partial v}{\partial y}=\alpha x(x-1)(2y-1);\\ &\frac{\partial^2 v}{\partial x^2}=2\alpha y(y-1); \qquad \frac{\partial^2 v}{\partial y^2}=2\alpha x(x-1). \end{aligned}\right\} \tag{iv}$$

Substituting for v in Eq. (ii), we obtain

$$I(v) = \int_0^1 \int_0^1 \alpha xy\,(x-1)\,(y-1)\,[2k - 2\alpha y(y-1) - 2\alpha x(x-1)]\,dx\,dy. \qquad \text{(v)}$$

Let

$$\left.\begin{aligned} a &= \int_0^1 \int_0^1 xy\,(x-1)\,(y-1)\,dx\,dy = \frac{1}{36} \\ b &= \int_0^1 \int_0^1 xy^2\,(x-1)\,(y-1)^2\,dx\,dy = -\frac{1}{180} \\ c &= \int_0^1 \int_0^1 x^2 y\,(x-1)^2\,(y-1)\,dx\,dy = -\frac{1}{180}. \end{aligned}\right\} \qquad \text{(vi)}$$

Equation (v) now simplifies to

$$I(v) = 2k\alpha a - 2\alpha^2 b - 2\alpha^2 c.$$

Hence

$$\frac{\partial I}{\partial \alpha} = 0 = 2ka - 4\alpha b - 4\alpha c.$$

Thus

$$\alpha = \frac{ak}{2\,(b+c)} = -\frac{5}{4}k, \quad \text{using (vi).}$$

It follows that the required approximation for u is given by

$$u \approx v = -\frac{5}{4}kxy\,(x-1)\,(y-1).$$

The student should verify that the Galerkin method gives the same solution as above.

11.4 FINITE ELEMENT METHOD

The Rayleigh–Ritz and Galerkin methods, discussed in the previous sections, cannot be applied directly for obtaining the global approximate solutions of engineering problems. An important reason for this is the difficulty associated with the choice of trial functions (satisfying the boundary conditions) particularly for complicated boundaries. This means that the application is restricted to problems with a simple geometry. Another reason is that very high-order polynomials have to be used to obtain global solutions with a reasonable accuracy. In the finite element method, the ideas of both the Rayleigh–Ritz and Galerkin methods are used in such a way that the above mentioned difficulties are overcome.

In the finite element method the region of interest is subdivided into a finite number of subregions, called the *elements*, and over each element the variational formulation of the given differential equation is constructed using simple functions for approximations. The individual elements are then assembled and the equations for the whole problem are formed by a piecewise application of the variational method. For better accuracy it will not be necessary to increase the order of the functions used, but it would be sufficient to use a finer mesh. In this way, the difficulties encountered in the direct application of the variational methods are overcome. The basic steps involved in the finite element method are as follows:

(i) *Discretization:* The given domain is divided into a number of finite elements. The points of intersection are called *nodes*. The nodes and the elements are both numbered.

(ii) *Derivation of element equations:* For the given differential equation, a variational formulation is constructed over a typical element. The element equations are obtained by substituting a typical dependent variable, say

$$u = \sum_{i=1}^{n} u_i \, \psi_i$$

into the variational formulation. After choosing ψ_i, the interpolation functions, the element matrices are computed.

(iii) *Assembly:* The next step is the assembly of the element equations so that the total solution is continuous. When this is done, the entire system takes the matrix form

$$K\mathbf{u}' = \mathbf{F}',$$

where K = assemblage property matrix, and $\mathbf{u}'$ and $\mathbf{F}'$ are column vectors containing unknowns and external forces.

(iv) *Boundary conditions:* The above system of equations is modified using the boundary conditions of the problem.

(v) *Solution of the equations:* After incorporating the boundary conditions, the system is solved by any standard technique, for example, the LU decomposition.

The preceding steps are quite general but they are common to most finite element approaches. In the following section, these steps are elaborated and explained with an example of one-dimensional problem. Since the two-dimensional problems are modelled by partial differential equations, their finite element analysis is more complicated and are therefore not considered here.

11.4.1 Finite Element Method for One-dimensional Problems

We consider the two-point boundary value problem defined by

$$\frac{d^2y}{dx^2} = -f(x), \quad 0 < x < 1 \tag{11.27}$$

with the boundary conditions

$$y(0) = 0, \qquad \left[\frac{dy}{dx}\right]_{x=1} = 0 \tag{11.28}$$

The basic steps involved in the finite element method are now elaborated and explained (see Reddy [1985]):

Step 1 (*Discretization of the region*)*:* In the present problem, the region of interest is the x-axis from $x = 0$ to $x = 1$. Suppose that this is divided into a set of subintervals, called *elements*, of unequal length, in general. The intersection points are called *nodes*. Let these be given by $x_0, x_1, x_2, \ldots, x_{n-1}$, x_n, where $x_0 = 0$ and $x_n = 1$. The elements are numbered as ①, ②, ③, …, ⓝ, a typical element being the eth element of length h_e from node $e - 1$ to node e. Let x_{e-1} and x_e be the values of x at the nodes $e - 1$ and e, and let $y^{(e-1)}$ and $y^{(e)}$ be the values of y at these nodes, respectively. In general, $y^{(e)}$ satisfies the condition that *outside e.*

$$y^{(e)}(x) = 0 \qquad \text{for all elements } e. \tag{11.29}$$

For example, in Fig. 11.2 $y^{(e)}(x_3)$ is nonzero whereas $y^{(e)}(x_4) = 0$.

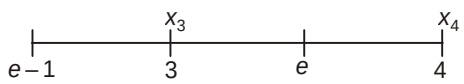


Figure 11.2 Typical eth element.

Using Eq. (11.29), it follows that the global approximate solution, $y(x)$, can be written as

$$y(x) = \sum_e y^{(e)}(x), \tag{11.30}$$

where the summation is taken over all the elements.

This completes the discretization process and in the next step, we choose a particular element e and formulate a variational principle for it.

Step 2 (*Variational formulation over the element e*)*:* From Eq. (11.27), we obtain

$$\int_{x_{e-1}}^{x_e} v\frac{d^2y}{dx^2}\,dx = -\int_{x_{e-1}}^{x_e} v\,f\,dx$$

which is written as

$$0 = \int_{x_{e-1}}^{x_e} \left[v\frac{d^2y}{dx^2} + vf \right] dx$$

$$= \left[v\frac{dy}{dx} \right]_{x_{e-1}}^{x_e} - \int_{x_{e-1}}^{x_e} v'\frac{dy}{dx}\,dx + \int_{x_{e-1}}^{x_e} vf\ dx$$

$$= -\int_{x_{e-1}}^{x_e} [v'y' - vf]\,dx + v(x_e)D_2^{(e)} + v(x_{e-1})D_1^{(e)}, \tag{11.31}$$

where

$$D_1^{(e)} = \left[-\frac{dy}{dx} \right]_{x_{e-1}} \quad \text{and} \quad D_2^{(e)} = \left[\frac{dy}{dx} \right]_{x_e}. \tag{11.32}$$

In the next step, we use a variational method to approximate Eq. (11.31). We demonstrate this by using the Rayleigh–Ritz method.

Step 3 (*Rayleigh–Ritz approximation over the element e*)*:* Let $y_e(x)$ be an approximation to $y(x)$ over the element e, so that

$$y_e(x) = \sum_{j=1}^{n} \alpha_j^{(e)} \phi_j(x), \tag{11.33}$$

where the α_j are parameters to be determined and $\phi_j(x)$ are approximation functions to be chosen. Substituting Eq. (11.33) in Eq. (11.31), we obtain

$$\sum_{j=1}^{n} \alpha_j^{(e)} \left[\int_{x_{e-1}}^{x_e} \phi_j'(x)\phi_i'(x)\,dx \right]$$

$$= \int_{x_{e-1}}^{x_e} f\phi_i(x)\,dx + \phi_i(x_e)D_2^{(e)} + \phi_i(x_{e-1})D_1^{(e)}, \qquad i = 1, 2, \ldots, n. \tag{11.34}$$

Equation (11.34) can be written in the matrix form

$$K_{ij}^{(e)} \alpha_j^{(e)} = F_i^{(e)}, \tag{11.35}$$

where K_{ij} and F_i are called the *stiffness matrix* and *force vector* respectively, and are given by

$$K_{ij}^{(e)} = \int_{x_{e-1}}^{x_e} \phi_i'(x)\phi_j'(x)\,dx \tag{11.36}$$

and

$$F_i^{(e)} = \int_{x_{e-1}}^{x_e} f\phi_i(x)\,dx + \phi_i(x_e)D_2^{(e)} + \phi_i(x_{e-1})D_1^{(e)}. \tag{11.37}$$

In the Rayleigh–Ritz and Galerkin methods, the system of equations is obtained in terms of the arbitrary parameters α_j. In the finite element method, on the other hand, the unknown values of the dependent variable y at the nodes are taken as parameters. This is done in the following way. Let

$$y(x) = \alpha_1 + \alpha_2 x \tag{11.38}$$

be an approximation in the element e. We have

$$\left.\begin{aligned} y(x_{e-1}) &= \alpha_1 + \alpha_2 x_{e-1} = y_1^{(e)} \\ y(x_e) &= \alpha_1 + \alpha_2 x_e = y_2^{(e)}. \end{aligned}\right\} \tag{11.39}$$

Solving the equations given in Eq. (11.39), we obtain

$$\alpha_1 = \frac{y_1^{(e)}x_e - y_2^{(e)}x_{e-1}}{x_e - x_{e-1}} \tag{11.40}$$

and

$$\alpha_2 = \frac{y_2^{(e)} - y_1^{(e)}}{x_e - x_{e-1}}. \tag{11.41}$$

Equation (11.38) now becomes

$$\begin{aligned} y(x) &= \frac{y_1^{(e)}x_e - y_2^{(e)}x_{e-1}}{x_e - x_{e-1}} + \frac{y_2^{(e)} - y_1^{(e)}}{x_e - x_{e-1}}x \\ &= \frac{x_e - x}{x_e - x_{e-1}}y_1^{(e)} + \frac{x - x_{e-1}}{x_e - x_{e-1}}y_2^{(e)} \\ &= \sum_{i=1}^{2} y_i^{(e)}\phi_i^{(e)}(x) \end{aligned} \tag{11.42}$$

where

$$\phi_1^{(e)}(x) = \frac{x_e - x}{x_e - x_{e-1}} \quad \text{and} \quad \phi_2^{(e)}(x) = \frac{x - x_{e-1}}{x_e - x_{e-1}}. \tag{11.43}$$

With $x_1 = x_{e-1}$ and $x_2 = x_e$, the functions $\phi_i^{(e)}$ have the property

$$\phi_i^{(e)}(x_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases} \tag{11.44}$$

Instead of Eq. (11.35), we now have

$$K_{ij}^{(e)} y_j^{(e)} = F_i^{(e)}, \tag{11.45}$$

where $K_{ij}^{(e)}$ and $F_i^{(e)}$ are given by Eq. (11.36) and (11.37).

With the choice of $\phi_i^{(e)}(x)$ as in Eq. (11.43), we now demonstrate the computation of $K^{(e)}$ and $F^{(e)}$. In particular, we choose $f = 2$. With $h_e = x_e - x_{e-1}$, we obtain

$$\frac{d\phi_1^{(e)}}{dx} = -\frac{1}{h_e} \quad \text{and} \quad \frac{d\phi_1^{(e)}}{dx} = -\frac{1}{h_e} \tag{11.46}$$

where

$$\left.\begin{aligned} K_{11} &= \int_{x_{e-1}}^{x_e} \left(-\frac{1}{h_e}\right)^2 dx = \frac{1}{h_e} \\ K_{12} &= \int_{x_{e-1}}^{x_e} -\frac{1}{h_e^2} dx = -\frac{1}{h_e} = K_{21} \\ K_{22} &= \int_{x_{e-1}}^{x_e} \frac{1}{h_e^2} dx = \frac{1}{h_e} \end{aligned}\right\} \tag{11.47}$$

and

$$\left.\begin{aligned} F_1^{(e)} &= 2 \int_{x_{e-1}}^{x_e} \frac{x_e - x}{h_e} dx + D_1^{(e)} = h_e + D_1^{(e)} \\ F_2^{(e)} &= 2 \int_{x_{e-1}}^{x_e} \frac{x - x_{e-1}}{h_e} dx + D_2^{(e)} = h_e + D_2^{(e)}. \end{aligned}\right\} \tag{11.48}$$

As a particular case, we consider the following example.

Example 11.6 We consider the following problem defined by

$$\frac{d^2 y}{dx^2} = -2, \quad 0 < x < 1, \qquad y(0) = 0, \; y'(1) = 0. \tag{i}$$

The exact solution of the above problem is given by

$$y(x) = 2x - x^2 \tag{ii}$$

Comparison with Eq. (11.27) shows that $f(x) = 2$.

(a) To demonstrate the steps involved in the finite element solution, we divide [0, 1] into two equal subintervals with $h_e = 1/2$. From Eqs. (11.43) and (11.44), we obtain the equations for both elements.

(i) $e = 1$: $x_{e-1} = 0,\ x_e = 1/2,$

$$K^{(1)} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}, \quad F^{(1)} = \begin{bmatrix} \frac{1}{2} + D_1^{(1)} \\ \frac{1}{2} + D_2^{(1)} \end{bmatrix}$$

(ii) $e = 2$: $x_{e-1} = 1/2,\ x_e = 1$

$$K^{(2)} = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}, \quad F^{(2)} = \begin{bmatrix} \frac{1}{2} + D_1^{(2)} \\ \frac{1}{2} + D_2^{(2)} \end{bmatrix}.$$

Having determined the equations for each element, these have to be assembled now to determine the global approximations. This will be the next step in the finite element solution.

Step 4 (*Assembly of element equations*)*:* We shall explain this step with reference to the two elements obtained in Example 11.6. In this case, the two elements are connected at the node 2. Since the function $y(x)$ is continuous, it follows that y_2 of element 1 should be the same as y_1 of element 2. For the two elements of Example 11.6, the correspondence can be expressed mathematically as follows:

$$y_1^{(1)} = Y_1, \qquad y_2^{(1)} = Y_2 = y_1^{(2)}, \qquad y_2^{(2)} = Y_3.$$

In the finite element analysis, such relations are usually called *interelement continuity conditions*.

Using the above relations, the global finite element model of the given boundary value problem is

$$\begin{bmatrix} 2 & -2 & 0 \\ -2 & 2+2 & -2 \\ 0 & -2 & 2 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1/2 + D_1^{(1)} \\ 1 + D_2^{(1)} + D_2^{(2)} \\ 1/2 + D_2^{(2)} \end{bmatrix}.$$

The next step is the imposition of boundary conditions.

Step 5 (*Imposition of boundary conditions*)*:* The homogeneous boundary condition gives $Y_1 = 0$. Then, we obtain the equations:

$$4Y_2 - 2Y_3 = 1, \qquad -2Y_2 + 2Y_3 = \frac{1}{2}$$

since $D_2^{(1)}$ and $D_2^{(2)}$ cancel each other and $D_2^{(2)} = 0$ is the natural boundary condition.

The solution of this system is given by

$$Y_2 = \frac{3}{4} \quad \text{and} \quad Y_3 = 1$$

Finally, the approximate solution throughout the interval [0, 1] can now be found using the formula

$$y(x) = \sum_{i=1}^{2} Y_i\, \phi_i^{(e)}(x) = \begin{cases} Y_1 \phi_1^{(1)}(x) + Y_2 \phi_2^{(1)}(x), & 0 \le x \le \frac{1}{2} \\ Y_2 \phi_1^{(2)}(x) + Y_3 \phi_2^{(2)}(x), & \frac{1}{2} \le x \le 1 \end{cases}$$

$$= \begin{cases} \frac{3}{2}x, & 0 \le x \le \frac{1}{2} \\ \frac{x+1}{2}, & \frac{1}{2} \le x \le 1 \end{cases}$$

on substitutions and simplification.

From the above, we obtain the approximate value of $y(1/4) \approx 3/8 = 0.375$, whereas its exact value $= 2(1/4) - 1/16 = 7/16$.

(b) To improve the accuracy, we now consider four elements of length 1/4. In this case, the element matrices become

(i) $e = 1$: $x_{e-1} = 0,\ x_e = 1/4$

$$K_{11}^{(1)} = 4, \qquad K_{12}^{(1)} = K_{21}^{(1)} = -4, \qquad K_{22}^{(1)} = 4$$

$$F_1^{(1)} = \frac{1}{4} + D_1^{(1)}, \qquad F_2^{(1)} = \frac{1}{4} + D_2^{(1)}$$

$$K^{(1)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \qquad F^{(1)} = \begin{bmatrix} 1/4 + D_1^{(1)} \\ 1/4 + D_2^{(1)} \end{bmatrix}$$

(ii) $e = 2$: $x_{e-1} = 1/4,\ x_e = 1/2$

$$K^{(2)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \qquad F^{(2)} = \begin{bmatrix} 1/4 + D_1^{(2)} \\ 1/4 + D_2^{(2)} \end{bmatrix}$$

(iii) $e = 3$: $x_{e-1} = 1/2,\ x_e = 3/4$

$$K^{(3)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \qquad F^{(3)} = \begin{bmatrix} 1/4 + D_1^{(3)} \\ 1/4 + D_2^{(3)} \end{bmatrix}$$

(iv) $e = 4$: $x_{e-1} = 3/4,\ x_e = 1$

$$K^{(4)} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix}, \qquad F^{(4)} = \begin{bmatrix} 1/4 + D_1^{(4)} \\ 1/4 + D_2^{(4)} \end{bmatrix}$$

To avoid confusion, we now write down the complete system for each element

$$e=1 \qquad \begin{bmatrix} 4 & -4 & 0 & 0 & 0 \\ -4 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1/4 + D_1^{(1)} \\ 1/4 + D_2^{(1)} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e=2 \qquad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & -4 & 0 & 0 \\ 0 & -4 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/4 + D_1^{(2)} \\ 1/4 + D_2^{(2)} \\ 0 \\ 0 \end{bmatrix}$$

$$e=3 \qquad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -4 & 0 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/4 + D_1^{(3)} \\ 1/4 + D_2^{(3)} \\ 0 \end{bmatrix}$$

$$e=4 \qquad \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -4 \\ 0 & 0 & 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ D_1^{(4)} \\ D_2^{(4)} \end{bmatrix}$$

Adding up the above, we obtain

$$\begin{bmatrix} 4 & -4 & 0 & 0 & 0 \\ -4 & 4+4 & -4 & 0 & 0 \\ 0 & -4 & 4+4 & -4 & 0 \\ 0 & 0 & -4 & 4+4 & -4 \\ 0 & 0 & 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1/4 + D_1^{(1)} \\ 1/2 + D_2^{(1)} + D_1^{(2)} \\ 1/2 + D_2^{(2)} + D_1^{(3)} \\ 1/2 + D_2^{(3)} + D_1^{(4)} \\ 1/4 + D_2^{(4)} \end{bmatrix}$$

By boundary condition, we have $Y_1 = 0$.

The system now becomes

$$8Y_2 - 4Y_3 = 1/2$$

$$-4Y_2 + 8Y_3 - 4Y_4 = 1/2$$

$$-4Y_3 + 8Y_4 - 4Y_5 = 1/2$$

$$-4Y_4 + 4Y_5 = 1/4 + D_2^{(4)}$$

But $D_2^{(4)} = y'(1) = 0.$

The solution of the system is

$$Y_2 = \frac{7}{16}, \quad Y_3 = \frac{3}{4}, \quad Y_4 = \frac{15}{16}, \quad Y_5 = 1.$$

Then the approximate solution valid for [0, 1] is

$$y(x) = \begin{cases} \dfrac{7x}{4}, & 0 \le x \le \dfrac{1}{4} \\ \dfrac{10x+1}{8} & \dfrac{1}{4} \le x \le \dfrac{1}{2} \\ \dfrac{6x+3}{8}, & \dfrac{1}{2} \le x \le \dfrac{3}{4} \\ \dfrac{x+3}{4}, & \dfrac{3}{4} \le x \le 1 \end{cases}.$$

From the above, we obtain $y(1/4) = 7/16$, which is the same as the exact solution.

11.5 CONCLUDING REMARKS

We have attempted, in this book, to supply to the reader some *basic* numerical methods which are indispensable for current scientific research. Many methods have been excluded since we do not intend to be exhaustive. There is indeed much more to include. Unfortunately, the limitations of space and our own experience have forced us to exclude many important topics such as eigenvalue problems in differential equations, linear and non-linear programming, convergence and stability criteria for partial differential equations and numerical solution of singular integral equations. Our motivation throughout has been to present the various methods in a very simple way so as to enable the reader to understand and apply them to solve the specific problems arising in his work. The book is, therefore, intended to be an introductory text only.*

*The reader who wishes to pursue the subject and understand the analysis of these methods is recommended to refer to: Isaacson and Keller, *Analysis of Numerical Methods*, and other references cited in the Bibliography.

EXERCISES

11.1 Explain the difference between a function and a functional. Prove that
(a) $\delta(dx) = d(\delta x)$.

(b) $\delta \int f(x)\ dx = \int \delta f(x)\ dx$.

11.2 Establish the Euler–Lagrange equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0,$$

for a functional defined by

$$I(y) = \int_a^b f(x, y, y')\, dx.$$

Obtain functionals for the following boundary value problems (Problems 11.3–11.8):

11.3 $\dfrac{d^2y}{dx^2} = g(x), \qquad y(0) = y(1) = 0.$

11.4 $\dfrac{d^2y}{dx^2} + ky = x^3, \qquad y(a) = 0, \ \left[\dfrac{dy}{dx}\right]_{x=b} = 1.$

11.5 $x^2 \dfrac{d^2y}{dx^2} + 2x\dfrac{dy}{dx} = g(x), \qquad y(0) = y(1) = 0.$

11.6 $\dfrac{d^2y}{dx^2} + p(x)y + q(x) = 0, \quad y(a) = y(b) = 0.$

11.7 $\dfrac{d^4y}{dx^4} + ky = f(x), \qquad 0 < x < 1,$

$y = \dfrac{d^2y}{dx^2} = 0$ at $x = 0$ and $x = 1$.

Solve the following boundary value problems by Rayleigh–Ritz method (Problems 11.8–11.10):

11.8 $\dfrac{d^2y}{dx^2} + y = x^2, \qquad y(0) = y(1) = 0.$

11.9 $\dfrac{d^2y}{dx^2} + 2x = 0, \qquad y(0) = y(1) = 0.$

11.10 $\dfrac{d^2y}{dx^2} - 64y + 10 = 0, \qquad y(0) = y(1) = 0.$

Apply Galerkin's method to solve the boundary value problems (Problems 11.11–11.12):

11.11 $\dfrac{d^2y}{dx^2} + y = x^2, \quad y(0) = y(1) = 0.$

11.12 $\dfrac{d^2y}{dx^2} - 64y + 10 = 0, \quad y(0) = y(1) = 0.$

Answers to Exercises

11.3 $I(v) = \displaystyle\int_0^1 \left[\frac{1}{2}\left(\frac{dv}{dx}\right)^2 + gv\right] dx$

11.4 $I(v) = \displaystyle\int_a^b \left[\left(\frac{dv}{dx}\right)^2 - kv^2 + 2vx^3\right] dx - v(b)$

11.5 $I(v) = \displaystyle\int_a^b v\left[2g - \frac{d}{dx}\left(x^2\frac{dv}{dx}\right)\right] dx$

11.6 $I(v) = \displaystyle\int_a^b \left[\left(\frac{dv}{dx}\right)^2 - pv^2 - 2qv\right] dx$

11.7 $I(v) = \displaystyle\int_0^1 \left[\left(\frac{d^2v}{dx^2}\right)^2 + kv^2 - 2fv\right] dx$

11.8 $I(v) = \displaystyle\int_0^1 \left[\left(\frac{dv}{dx}\right)^2 - v^2 + 2vx^2\right] dx;$

$$y(x) \approx v(x) = -\frac{10}{121}x(1-x) - \frac{7}{41}x^2(1-x)$$

11.9 $I(v) = \displaystyle\int_0^1 \left[\left(\frac{dv}{dx}\right)^2 - 4vx\right] dx; \quad y(x) \approx v(x) = \frac{1}{3}(x - x^3).$

11.10 $v(x) = \dfrac{25}{37}x(1-x)$

11.11 $v(x) = -\dfrac{10}{123}x(1-x) - \dfrac{7}{41}x^2(1-x)$

11.12 $v(x) = \dfrac{25}{37}x(1-x)$

Bibliography

BOOKS

Ahlberg, J.H., E.N. Nilson and J.L. Walsh, *The Theory of Splines and their Applications*, Academic Press, New York, 1967.

Allaire, P.E., *Basics of the Finite Element Method*, William C Brown, Dubuque, IA, 1985.

Atkinson, K.E., *An Introduction to Numerical Analysis*, John Wiley & Sons, New York, 1978.

Bathe, K.J. and E.L. Wilson, *Numerical Methods in Finite Element Analysis*, Prentice-Hall, New Jersey, 1976.

Berndt, R. (Ed.), *Ramanujan's Note Books*, Part I, Springer-Verlag, New York, 1985.

Booth, A.D., *Numerical Methods*, Academic Press, New York, 1958.

Brebbia, C.A. and J.J. Connor, *Fundamentals of Finite Element Techniques for Structural Engineers*, Butterworths, London, 1975.

Brigham, E.O., *The Fast Fourier Transform*, Prentice-Hall, New Jersey, 1974.

Carnahan, B., H.A. Luther and J.O. Wilkes, *Applied Numerical Methods*, Wiley, New York, 1969.

Conte, S.D., *Elementary Numerical Analysis*, McGraw Hill, New York, 1965.

Chapra, S.C. and Raymond P. Canale, *Numerical Methods for Engineers*, 3ed., Tata McGraw-Hill, New Delhi, 2000.

Davies, A.J., *The Finite Element Method: A First Approach*, Clarendon Press, Oxford, 1980.

Davis, P.J. and P. Rabinowitz, *Methods of Numerical Integration*, Academic Press, New York, 1984.

De Boor, C., *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

Fox, L. and I.B. Parker, *Chebyshev Polynomials in Numerical Analysis*, Oxford University Press, 1968.

Froberg, C.E., *Introduction to Numerical Analysis*, Addison–Wesley, Reading, Mass., 1965.

Gerald, C.F. and P.O. Wheatley, *Applied Numerical Analysis*, 3rd ed., Addison-Wesley, Reading, Mass., 1989.

Greville, T.N.E., *Introduction to Spline Functions*, In *Theory and Applications of Spline Functions*, Academic Press, New York, 1969.

Hartree, D.R., *Numerical Analysis*, Oxford University Press, London, 1952.

Henrici, P., *Elements of Numerical Analysis,* Wiley, New York, 1964.

Henrici, P., *Applied and Computational Complex Analysis*, John Wiley & Sons, New York, 1974.

Hildebrand, F.B., *Introduction to Numerical Analysis*, McGraw Hill, New York, 1956.

Ian Jacques and Colin Judd, *Numerical Analysis*, Chapman and Hall, New York, 1987.

Isaacson, E. and H.B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

Jain, M.K., *Numerical Analysis for Scientists and Engineers*, S.B.W. Publishers, Delhi, 1971.

Levy, H. and E.A. Baggott, *Numerical Solution of Differential Equations*, Dover, New York, 1950.

McCormick, J.M. and M.G. Salvadori, *Numerical Methods in FORTRAN*, Prentice-Hall of India, New Delhi, 1971.

———, *Modern Computing Methods*, HMSO, London, 1961.

Mitchell, A.R. and R. Wait, *The Finite Element Method in Partial Differential Equations*, John Wiley & Sons, London, 1977.

Nielsen, K.L., *Methods in Numerical Analysis*, Macmillan Co., New York, 1964.

Noble, B., *Numerical Methods*, Vol. 2, Oliver and Boyd, Edinburgh, 1964.

Phillips, G.M. and P.J. Taylor, *Theory and Applications of Numerical Analysis*, Academic Press, London, 1973.

Press, W.H., B.P. Flanamer, S.A. Tenkolsky and W.T. Vetterling, *Numerical Recipes, The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986, 1992.

Reddy, J.N., *An Introduction to the Finite Element Method*, McGraw Hill Book Co., Singapore, 1985.

Sastry, S.S., *Engineering Mathematics*, 3rd eds., Vols. 1 and 2, Prentice-Hall of India, New Delhi, 2004.

Scarborough, J.B., *Numerical Mathematical Analysis*, Johns Hopkins University Press, Baltimore, 1950.

Scheid, Francis, *Theory and Problems of Numerical Analysis*, Schaum Series, McGraw Hill, New York, 1968.

Schumaker, L.L., In *The Theory and Applications of Spline Functions*, T.N.E. Greville (Ed.), pp. 87–102, Academic Press, New York, 1969.

Smith, G.D., *Numerical Solution of Partial Differential Equations*, Oxford University Press, London, 1965.

Stanton, R.G., *Numerical Methods for Science and Engineering*, Prentice-Hall of India, New Delhi, 1967.

Wilkinson, J.H., *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

TABLES

Interpolation and Allied Tables, Nautical Almanac Office, HMSO, London, 1956.

Handbook of Mathematical Functions, by Milton Abramovitz and I.A. Stegun, US Department of Commerce, Washington, 1965.

Orthogonal Polynomials, by Milton Abramovitz and I.A Stegun, US Department of Commerce, Washington, 1965.

Tables of Integrals and Other Mathematical Data, by H.M. Dwight, Macmillan, & Co., London, 1934.

RESEARCH PAPERS

Allasiny, E.L. and W.D. Hoskins, Cubic spline solutions to two-point boundary value problems, *Computer Journal*, Vol. 12, p. 151, 1969.

Atkinson, K.E., The numerical solution of Fredholm integral equations of the second kind, *SIAM J. Num. Anal*., Vol. 4, p. 337, 1967.

Bauer, W.F., *J. SIAM*, Vol. 6, p. 438, 1958.

Bickley, W.G., Piecewise cubic interpolation and two-points boundary value problems, *Computer Journal*, Vol. 11, p. 206, 1968.

Clenshaw, C.W. and A.R. Curtis, A method for numerical integration in an automatic computer, *Numer. Math*., Vol. 2, p. 197, 1960.

Cox, M.G., The numerical evaluation of B-splines, *J. Inst. Maths. Applics*. Vol. 10, p. 134, 1972.

———, The numerical evaluation of a spline from its B-spline representation, *J. Inst. Maths. Applics*., Vol. 15, p. 95, 1975.

———, The numerical evaluation of a spline from its B-spline representation, *J. Inst. Maths. Applics.,* Vol. 21, p. 135, 1978.

Curtis, A.R. and M.J.D. Powell, Using cubic splines to approximate functions of one variable to prescribed accuracy, *AERE Harwell Report No. AERE-R5602*-(HMSO), 1967.

de Boor, C., On calculation with B-splines, *J. Approx. Theory*, Vol. 6, p. 50, 1972.

Delves, L.M., The numerical evaluation of principal value integrals, *Computer Journal*, Vol. 10, p. 389, 1968.

Einarsson, Bo, Numerical calculation of Fourier integrals with cubic splines, *BIT*, Vol. 8, p. 279, 1968.

Einarsson, Bo, On the calculation of Fourier integrals, *Information Processing*, Vol. 71, p. 1346, 1972.

El-Gendi, S.E., Chebyshev solution of differential, integrals and integro-differential equations, *Computer Journal*, Vol. 12, p. 282, 1969.

Elliott, D., A Chebyshev series for the numerical solution of Fredholm integral equations, *Computer Journal*, Vol. 6, p. 102, 1963.

Filon, L.N.G., On a quadrature formula for trigonometric integrals, *Proc. Roy. Soc. Edin*., Vol. 49, p. 38, 1928.

Fox, L. and E.T. Goodwin, The numerical solution of nonsingular linear integral equations, *Phil. Trans. Roy. Soc*., A, Vol. 245, p. 501, 1953.

Fyfe, D.J., The use of cubic splines in the solution of two point boundary value problems, *Computer Journal*, Vol. 12, p. 188, 1969.

———,The use of cubic splines in the solution of two point boundary value problems, *Computer Journal*, Vol. 13, p. 204, 1970.

Greville, T.N.E., Data fitting by spline functions, *M.R.C. Tech. Sum. Report*, 893, Maths. Res. Center, US Army, University of Wisconsin, Madison, Wisconsin, 1968.

Henrici, P., The quotient difference algorithm, *App. Math. Series*, U.S. Bureau of Standards, 49, p. 23, 1958.

Ichida and Kiyono, *Int. J. Comp. Maths*., Vol. 4, p. 111, 1974.

Kalaba, R.E. and E.H. Ruspini, Theory of invariant imbedding, *Int. J. Engg. Sc*., Vol. 7, p. 1091, 1969.

Kershaw, D., A note on the convergence of natural cubic splines, *SIAM J. Num. Anal*., Vol. 8, 67, 1971.

———, Two interpolatory cubic splines, *Tech. Rep*., Dept. of Maths, University of Lancaster, 1972.

———, A numerical solution of an integral equation satisfied by the velocity distribution around a body of revolution in axial flow, *ARC. Rep. No. 3308,* 1961.

Love, E.R., The electrostatic field of two equal circular coaxial conducting disks, *Quar. J. Mech. App. Math*., Vol. 2, p. 428, 1949.

Moore, E., Exponential fitting using integral equations, *Int. J. Num. Meth. in Engg*., Vol. 8, p. 271, 1974.

Muller, D.E., A method for solving algebraic equations using an automatic computer, *Math. Tables*, *Aids Comp*., Vol. 10, p. 208, 1956.

Patricio, F., Cubic spline functions and initial-value problems, *BIT*, Vol. 18, p. 342, 1978.

Phillips, J.L., The use of collocation as a projection method for solving linear operator equations, *SIAM J. Num. Anal*., Vol. 9, p. 14, 1972.

Poornachandra, et al., An efficient algorithm for noise elimination using B-splines in wavelet domain, Biosignal, pp. 94–96, 2004.

Rutishauser, H., *Z. Angew*, *Math. Phys*., Vol. 5, p. 233, 1954.

Sastry, S.S., A numerical solution of an integral equation of the second kind occurring in aerodynamics, *Ind. J. Pure and App. Math*., Vol. 4, p. 838, 1973.

Sastry, S.S., Numerical solution of nonsingular Fredholm integral equations of the second kind, *Ind. J. Pure and App. Math*., Vol. 6, p. 773, 1975.

———, Numerical solution of Fredholm integral equations with a logarithmic singularity, *Int. J. Num. Meth. in Engg*., Vol. 10, p. 1202, 1976.

Sastry, S.S., Finite difference approximations to one-dimensional parabolic equations using a cubic spline technique, *J. Comp. and App. Math*., Vol. 2, p. 23, 1976.

Schoenberg, I.J., Contributions to the problem of approximation of equidistant data by Analytical functions, *Quart. App. Maths*., Vol. 4, p. 45, 1946.

Srivastava, K.N. and R.M. Palaiya, The distribution of thermal stress in a semi-infinite elastic solid containing a pennyshaped crack, *Int. J. Engg. Sc*., Vol. 7, p. 641, 1969.

Vandrey, F., A direct iteration method for the calculation of velocity distribution of bodies of revolution and symmetrical profiles, *ARC R&M*, 1951, No. 3374.

Wilkinson, J.H., Householder's method for the solution of the Algebraic Eigenvalue Problem, *Comp. J.,* April 1960.

Wolfe, M.A., The numerical solution of nonsingular integral and integro-differential equations by iteration with Chebyshev series, *Computer Journal*, Vol. 12, p. 193, 1969.

Wynn, P., A sufficient condition for the instability of q-d algorithm, *Num. Math*., Vol. 1, p. 203, 1959.

Young, A., The application of approximate product integration to the numerical solution of integral equations, *Proc. Roy. Soc. A*, Vol. 224, p. 561, 1954.

Model Test Paper 1

B.E./B.Tech. Degree Examination (Numerical Methods)

Answer All Questions

Time: 3 Hours Max: 100 Marks

Section A (10 × 2 = 20 marks)

1. Evaluate $e^{-0.2}$ correct to 2 decimal places.

2. Find the sum of the numbers

$$143.3,\ 15.45,\ 0.1734.$$

3. State the formula for finding the pth root of a positive number N.

4. State Newton–Raphson formula for finding the roots of the equations $f(x, y) = 0$ and $g(x, y) = 0$.

5. State the errors in Newton's forward and backward difference formulae.

6. If $I_1 = 0.6785$, and $I_2 = 0.6920$, find I using Romberg's method.

7. Explain the solution of a system of linear equations by decomposition of the matrix into LU form.

8. Write down the fourth order Runge–Kutta formula for the solution of the problem

$$\frac{dy}{dx} = f(x, y),\ y(x_0) = y_0.$$

9. Define a two-point boundary value problem and state any two methods of solving it.

10. State Laplace's equation in two dimensions and give its finite difference analogue.

Section B (5 × 16 = 80 marks)

11. (a) A root of the equation $x^3 + 3x^2 - 3 = 0$ lies between –3 and –2. Find this root, correct to 3 decimal places, by bisection method. [8]

(b) Compute the value of $\sqrt{10}$ correct to 4 decimal places. [8]

12. (a) Design a computational algorithm to implement Lagrange's interpolation formula and use it to compute the value of $f(5)$ from the following data for x and $f(x)$: [8]

(2, 46), (7, 71), (10, 110)

(b) Fit a curve of the form

$$y = \frac{a}{x} + bx$$

by the method of least squares to the following data of x and $f(x)$:

(1, 5.43), (2, 6.28), (4, 10.32), (6, 14.86), (8, 19.51) [8]

13. Derive the key equations of Sande–Tukey algorithm for computing the DFT of the sequence $f_k = 0, 1, 2, \ldots, 7$. Apply this method to find the DFT of the sequence $f_k = \{1, 2, 3, 4, 4, 3, 2, 1\}$. Draw the flowgraph. [16]

14. (a) Define a cubic spline and derive its governing equations, viz.

$$s_i(x) = \frac{1}{h_i}\left[\frac{(x_i - x)^3}{6}M_{i-1} + \frac{(x - x_{i-1})^3}{6}M_i + \left(y_{i-1} - \frac{h_i^2}{6}M_{i-1}\right)(x_i - x) + \left(y_i - \frac{h_i^2}{6}M_i\right)(x - x_{i-1})\right]$$

and

$$\frac{h_i}{6}M_{i-1} + \frac{1}{3}(h_i + h_{i+1})M_i + \frac{h_{i+1}}{6}M_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \; h_i = x_i - x_{i-1}.$$ [10]

(b) The function $y = f(x)$ is satisfied by the points (1, 1), (2, 5), (3, 11), (4, 8). Fit a natural cubic spline approximation to this data and find an approximate value of $y(1.5)$. [6]

15. Evaluate

$$I = \int_0^1 \frac{1}{1 + x^2}\,dx$$

using trapezoidal rule with $h = 0.5$, 0.25 and 0.125. Then obtain a better estimate by Romberg's method. Give the errors in the solutions obtained. [16]

OR

Find $y(0.2)$ given that

$$\frac{dy}{dx} = 3x + \frac{y}{2}, \;\; y(0) = 1,$$

using the Euler, the modified Euler and the fourth order Runge–Kutta methods with $h = 0.05$. [16]

Model Test Paper 2

B.E./B.Tech. Degree Examination (Numerical Methods)

Answer All Questions

Time: 3 Hours Max: 100 Marks

Section A (10 × 2 = 20 marks)

1. List any three sources of errors which you encounter while solving a methematical problem.

2. State sufficient conditions for the convergence of the iteration method to find the roots of the system $x = F(x, y)$ and $y = G(x, y)$.

3. Evaluate $\Delta^4(x^4)$.

4. Form the divided difference table for the data:

x	2	5	10
y	5	29	139

5. Write the normal equations to fit a quadratic $y = a + bx + cx^2$ to the data (x_i, y_i), $i = 1, 2, \ldots, n$.

6. Define DFT and IDFT of a function $f(t)$ defined at points t_i, $i = 0, 1, 2, \ldots, N-1$.

7. What are cardinal splines? How are they related to cubic splines?

8. Write the formula for $\frac{dy}{dx}$ using Stirling's, interpolation formula.

9. Explain the difference between Jacobi and Gauss–Seidel methods for the solution of a system of equations.

10. Explain the difference between explicit and implicit methods for the solution of the equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$.

Section B (5 × 16 = 80 marks)

11. (a) Prove that Newton–Raphson method has quadratic convergence. [6]

(b) Use Bessel's formula to estimate the value of y when $x = 5.0$ from the following data:

(0, 14.27), (4, 15.81), (8, 17.72), (12, 19.96) [6]

(c) If the straight line $y = a_0 + a_1x$ is fitted to the data points (x_i, y_i), $i = 1, 2, \ldots, n$, show that

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0, \; i = 1, 2, \ldots, n. \qquad [4]$$

12. (a) Using the Cooley–Tukey algorithm, compute the DFT of the sequence $f_k = \{1, -1, 1, -1\}$. [10]

(b) Given the data

x	-2	-1	2	3
y	-12	-8	3	5

compute $y'(1.0)$ using cubic spline approximation. [6]

13. (a) Evaluate

$$\int_0^1 \frac{1}{1+x}\,dx$$

using the 4-point Gauss quadrature formula. [6]

OR

If $y = A + Bx + Cx^2$ and y_0, y_1, y_2 are the values of y corresponding to $x = a$, $a + h$ and $a + 2h$, respectively, prove that

$$\int_a^{a+2h} y\,dx = \frac{h}{3}(y_1 + 4y_1 + y_2). \qquad [6]$$

(b) Solve the following equations by triangularisation method:

$8x - 3y + 2z = 20$, $4x + 11y - z = 33$, $6x + 3y + 12z = 36$ [10]

14. Solve the following equations by Gauss–Seidel method correct to 3 decimal places:

$10x - 5y - 2z = 3$, $4x - 10y + 3z = -3$, $x + 6y + 10z = -3$ [16]

OR

Using Taylor's series, find the values of $y(0.1)$, $y(0.2)$, $y(0.3)$ for the initial value problem

$$\frac{dy}{dx} = xy + y^2, \; y(0) = 1.$$

Hence find the value of $y(0.4)$ using Milne's method. [16]

15. Solve the heat equation

$$\frac{\partial u}{\partial t} = 4\frac{\partial^2 u}{\partial x^2} \qquad [16]$$

with the conditions $u(0, t) = u(4, t) = 0$ and $u(x, 0) = 4x - x^2$, $0 \le x \le 4$ for $t = \frac{1}{8}k$, $k = 0, 1, 2, 3, 4, 5$. [16]

OR

Solve Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \ 0 \le x, y \le 1$$

with boundary conditions $u(0, y) = u(1, y) = 10$, $u(x, 0) = u(x, 1) = 20$. Take $h = 0.25$ and apply Gauss-Seidel method to find values correct to 3 decimal places. [16]

Model Test Paper 3

B.E./B.Tech. Degree Examination (Numerical Methods)

Answer All Questions

Time: 3 Hours Max: 100 Marks

Section A (10 × 2 = 20 marks)

1. State the formula for finding the absolute error in the function $u = f(x_1, x_2)$ if Δx_1 and Δx_2 are the errors in x_1 and x_2, respectively.
2. Explain how you can find the reciprocal of a number N by Newton–Raphson method.
3. Explain Graeffe's root-squaring method for finding the zeros of a polynomial $p_n(x)$ of degree n.
4. What is inverse interpolation? State any formula for it.
5. What are radix-2 algorithms?
6. Define a cubic B-spline and state Cox, de-Boor formula.
7. Explain how you will compute a double integral numerically.
8. Define eigenvalue and eigenvector of a matrix. Explain briefly the power method for finding the smallest eigenvalue of a matrix.
9. Write down the finite difference analogue of the equation
$$y'' + f(x)y' + g(x)y = h(x) \text{ at } x = x_i.$$
10. State the explicit and implicit formulae for the solution of the equation
$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

Section B (5 × 16 = 80 marks)

11. (a) Explain briefly the regula–falsi method to find an approximate root, correct to 3 decimal places, of the equation $x^3 - 3x - 5 = 0$ that lives between 2 and 3. [8]

    (b) Establish Newton's divided difference formula and give an estimate of the remainder term in terms of the appropriate derivative. Deduce Newton's forward and backward interpolation formulae as particular cases.

    If $f(x) = \dfrac{1}{x^2}$, find the divided differences $[a, b]$ and $[a, b, c]$ [8]

12. (a) Find the values of a, b, c such that the parabola $y = a + bx + cx^2$ fits the following data (x_i, y_i) in the least squares sense:

$$(0,\ 1),\ (1,\ 5),\ (2,\ 10),\ (3,\ 22),\ (4,\ 38).$$ [8]

(b) Given the points, viz.

$$(1,\ -8),\ (2,\ -1),\ \text{and}\ (3,\ 18)$$

satisfying the function $y = f(x)$, find the linear and quadratic splines approximating the function $f(x)$ and determine the value of $y(2.5)$ in each case. [8]

13. (a) Find the value of $\int_0^{\pi/2} \sin x\, dx$ with $h = \frac{\pi}{20}$ using both the trapezoidal rule and Simpson's $\frac{1}{3}$-rule. [8]

(b) Use Householder's method to reduce the matrix

$$A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{bmatrix}$$

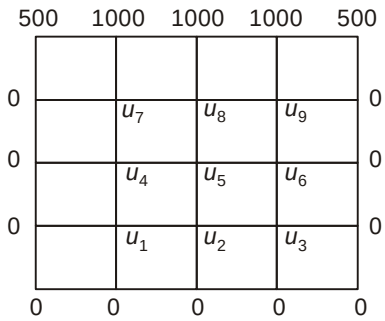
to a tridiagonal form. [8]

14. Using the fourth order Runge–Kutta formula, find $y(0.2)$ and $y(0.4)$ given that

$$\frac{dy}{dx} = \frac{y^2 - x^2}{y^2 + x^2},\ y(0) = 1.$$ [16]

OR

Solve Laplace's equation at the interior points of the square region given below. Use Gauss–Seidel method upto 7 iterations:



[16]

15. Derive the explicit scheme for the solution of the wave equation

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$$

Solve this equation with $a^2 = 4$ and the boundary conditions $u(0, t) = u(4, t) = 0$, $\frac{\partial u}{\partial t}(x, 0) = 0$ and $u(x, 0) = 4x - x^2$. With $h = 1$ and $k = 0.5$, find the values of $u(x, t)$ upto 3 time steps.

Model Test Paper 4

B.E./B.Tech. Degree Examination (Numerical Methods)

Answer All Questions

Time: 3 Hours Max: 100 Marks

Section A (10 × 2 = 20 marks)

1. Find the value of $\sqrt{3.02} - \sqrt{3}$, correct to 3 decimal places.

2. State the conditions to be satisfied by $\phi(x)$ if the equation $x = \phi(x)$ possesses a unique solution in $[a, b]$.

3. State the condition of convergence of Newton–Raphson formula for finding the root of $f(x) = 0$.

4. Show that $E = \dfrac{\Delta^2}{\delta^2}$.

5. Transform the equation $y = \dfrac{x}{a + bx}$ to a linear form.

6. Explain the terms 'Decimation in Time' and 'Decimation in Frequency'.

7. State the cubic spline formula for approximating the data (x_i, y_i), $i = 0, 1, \ldots, n$, in terms of the spline second derivatives.

8. Give the error in Simpson's $\dfrac{1}{3}$-rule.

9. What is meant by saying that the Runge–Kutta formula is of the fourth order?

10. State the local truncation error of the Crank–Nicolson formula.

Section B (5 × 16 = 80 marks)

11. (a) Explain the underlying principle of Muller's method and give the computational steps to compute a root of the equation $\cos x = xe^x$ by Muller's method. Find the root which lies between 0 and 1, correct to 4 decimal places. [10]

(b) Using Lagrange's formula, estimate the value of $y(10)$ from the following data (x, y):

(5, 12), (6, 13), (9, 14), (11, 16) [6]

12. (a) Obtain the first four orthogonal polynomials $f_n(x)$ on $[-1, 1]$ with respect to the weight function $W(x) = 1$. [6]

(b) using the Cooley–Tukey algorithm, compute the DFT of the sequence $\{1, -1, 1, -1\}$. Draw the flow-graph of the computations. [10]

13. (a) Fit natural and D_1 cubic splines for the following data satisfying the function $y = e^x$:

(0.10, 1.1052), (0.20, 1.2214), (0.30, 1.3499).

Approximate $e^{0.15}$ in each case and state which of these is the best fit. [10]

(b) Prove the minimization property of natural cubic splines. [6]

14. (a) Write down the formulae for computing the values of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ at any point, obtained from Newton's forward difference interpolation formula. Obtain the approximate values of $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$ for $x = 1.2$ from the following data:

(1.0, 2.7183), (1.2, 3.3201), (1.4, 4.0552), (1.6, 4.9530), (1.8, 6.0496), (2.0, 7.3891), (2.2, 9.0250). [8]

(b) Derive the trapezoidal rule

$$\int_a^b y\,dx = \frac{h}{2}\left[y_0 + 2(y_1 + y_2 + \cdots + y_{n-1}) + y_n\right]$$

and find an expression for the error in this formula. [8]

15. (a) Using Gauss–Seidel method upto 5 iterations, solve the system:

$30x - 2y + 3z = 75,\ 2x + 2y + 18z = 30,\ x + 17y - 2z = 48$

(b) Solve the boundary value problem

$$\frac{d^2y}{dx^2} - y = 0,\ y(0) = 0,\ y(2) = 3.62686$$

by finite difference method with $h = 0.5$. Find the value of $y(1.0)$ and state the absolute error in the solution. [8]

OR

Use Crank–Nicolson method to solve the equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ with the boundary conditions $u(x, 0) = 0$, $u(0, t) = 0$, $u(1, t) = 200t$. Find the values of $u(x, t)$ for two time steps taking $h = \frac{1}{4}$ and $k = \frac{1}{8}$. [16]

Index

FIFTH EDITION

Introductory Methods of Numerical Analysis

S.S. Sastry

This thoroughly revised and updated text, now in its fifth edition, continues to provide a rigorous introduction to the fundamentals of numerical methods required in scientific and technological applications, emphasizing on teaching students numerical methods and in helping them to develop problem-solving skills.

While the essential features of the previous editions such as References to MATLAB, IMSL, Numerical Recipes program libraries for implementing the numerical methods are retained, a chapter on Spline Functions has been added in this edition because of their increasing importance in applications.

This text is designed for undergraduate students of all branches of engineering.

NEW TO THIS EDITION

- Includes additional modified illustrative examples and problems in every chapter.
- Provides answers to all chapter-end exercises.
- Illustrates algorithms, computational steps or flow charts for many numerical methods.
- Contains four model question papers at the end of the text.

THE AUTHOR

S.S. SASTRY, Ph.D., is Formerly, Scientist/Engineer SF in the Applied Mathematics Division of Vikram Sarabhai Space Centre, Trivandrum. Earlier, he taught both undergraduate and postgraduate students of engineering at Birla Institute of Technology, Ranchi.

A renowned mathematical scientist and the author of the books on Engineering Mathematics (all published by PHI Learning), Dr. Sastry has a number of research publications in numerous journals of national and international repute.

Other books by the author

***Engineering Mathematics*, Vol. 1, 4th ed.**

***Engineering Mathematics*, Vol. 2, 4th ed.**

Advanced Engineering Mathematics

ISBN:978-81-203-4592-8

9 788120 345928

₹ 250.00

www.phindia.com